

This one cool data trick you won't believe!

How applying clickbait and SEO
techniques to your data can make it
more discoverable

Jo Cook | Data Discovery Lead | Astun Technology

Who is this "Astun Technology" of which you speak?

Speaker notes

The sorts of services we supply are mapping platforms, spatial databases and metadata portals to clients such as Scottish Government, DEFRA, EA, Ordnance Survey and Local Authorities. We also contribute to software development, documentation and guidance, and standards implementation and development.

Who is this "Astun Technology" of which you speak?

Astun Technology was founded in 2006. We're based in Epsom, but our 25-ish staff are spread across Europe. We do spatial and data "stuff", based on an open source technology stack.

Speaker notes

The sorts of services we supply are mapping platforms, spatial databases and metadata portals to clients such as Scottish Government, DEFRA, EA, Ordnance Survey and Local Authorities. We also contribute to software development, documentation and guidance, and standards implementation and development.

...and who are you?



...and who are you?

Hi I'm Jo 

I've worked at Astun for over 10 years and lead on data discovery (metadata etc). I'm mad keen on problem solving and acting as an enabler. I also do "stuff" with OSGeo and the AGI.



Speaker notes

The finding data and metadata pieces of the puzzle work just as much for internal data as external



Today's Problem

Speaker notes

The finding data and metadata pieces of the puzzle work just as much for internal data as external



Today's Problem

People need to find data

Speaker notes

The finding data and metadata pieces of the puzzle work just as much for internal data as external

A close-up photograph of a large stack of silver-colored metal cans, likely food cans, stacked in several rows. The cans are shiny and reflective, with some labels visible on the top ones.

Today's Problem

People need to find data

Metadata helps, but it's hard!

Speaker notes

The finding data and metadata pieces of the puzzle work just as much for internal data as external



Today's Problem

People need to find data

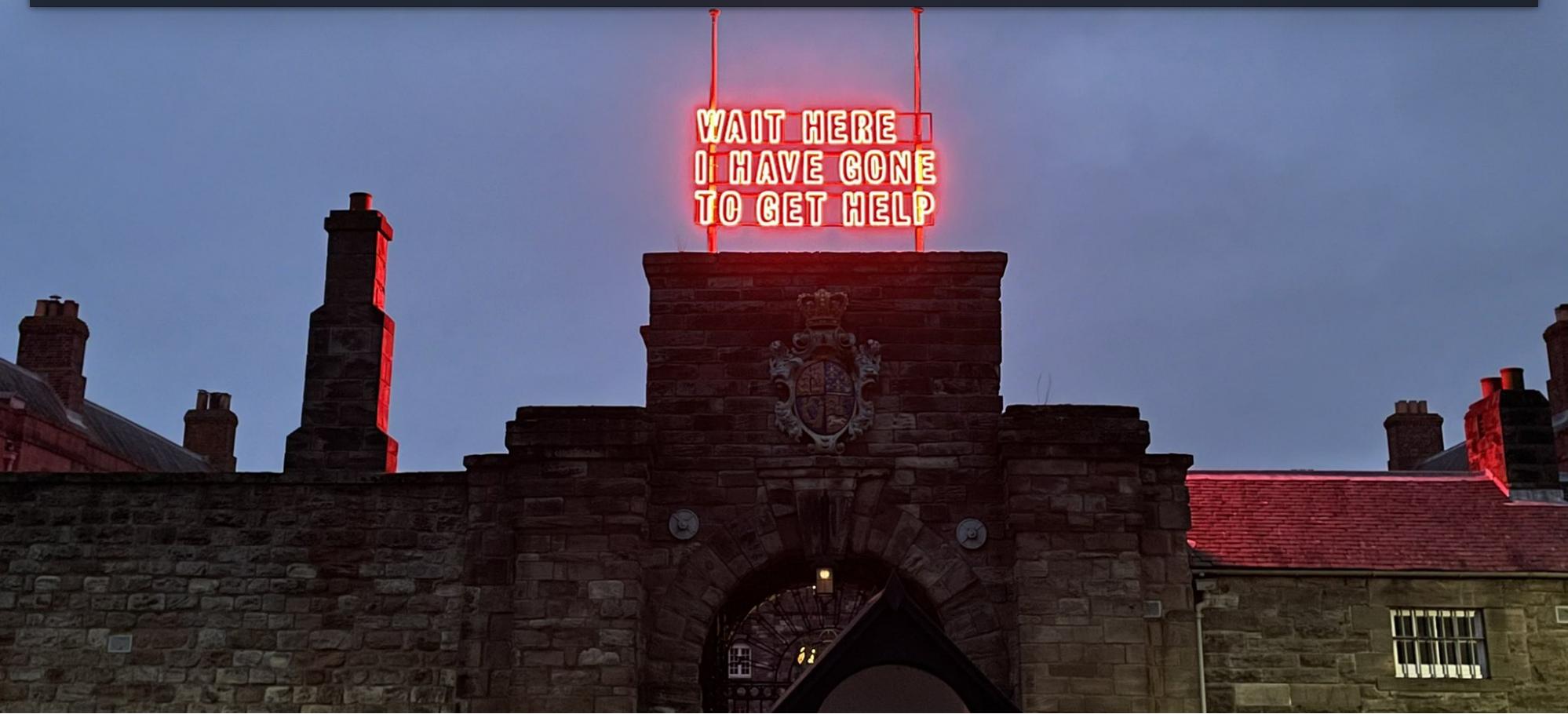
Metadata helps, but it's hard!

It's all about search engines

Speaker notes

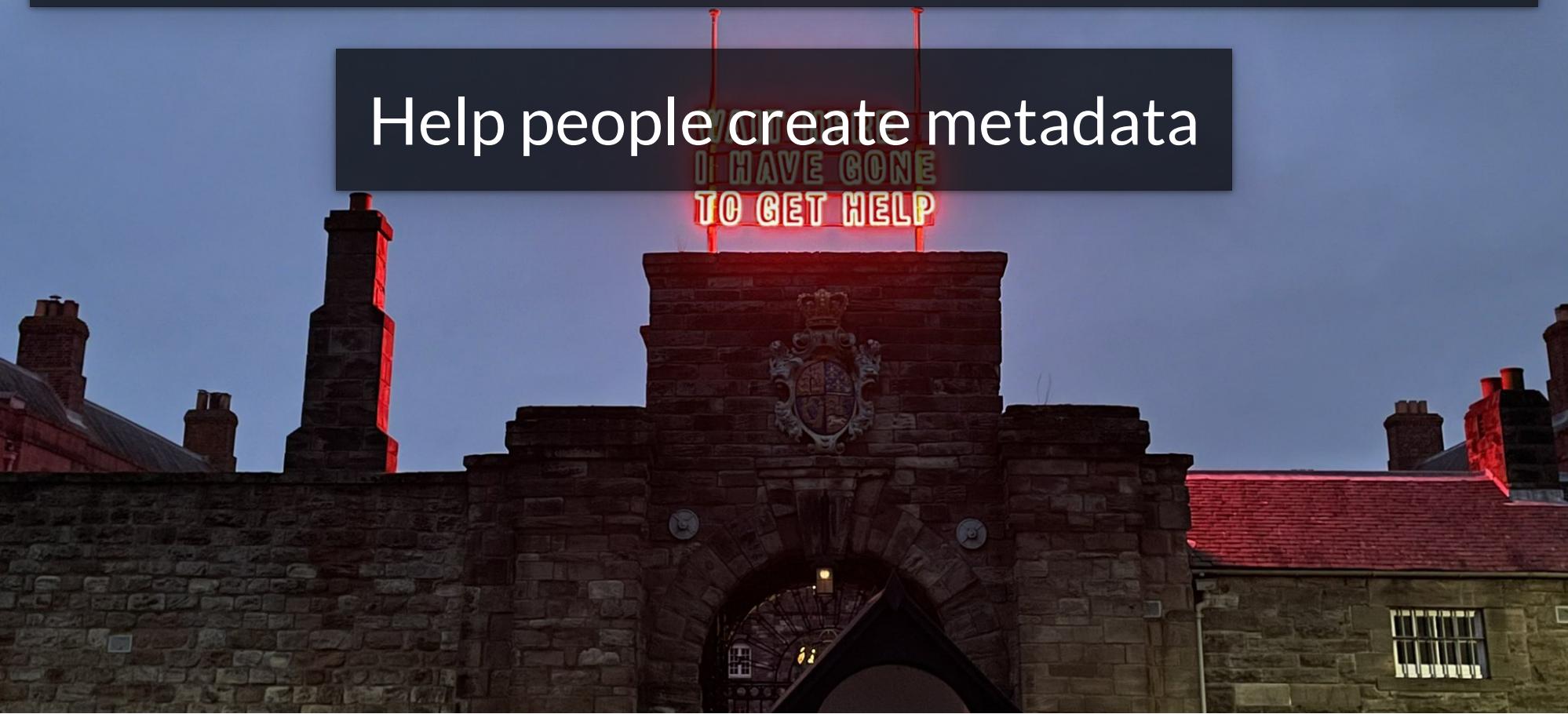
The finding data and metadata pieces of the puzzle work just as much for internal data as external

The Solution



The Solution

Help people create metadata



The Solution

Help people create metadata

I HAVE GONE
TO GET HELP

Help search engines understand it

The Solution

Help people create metadata

I HAVE GONE
TO GET HELP

Help search engines understand it

Get the dataset to appear high up in search results

Helping people create metadata

*"Quality metadata is key to making
data discoverable"*

*Geospatial Commission "Recommendations For SEO of UK Government's
Geospatial Metadata" 2019*

Speaker notes

Data producers need to create high-quality metadata, but also to employ SEO techniques to make the data more search-engine and user-friendly. Modern search engines also use the data to display richer results. Any solution must work at scale, not just for single datasets, and must be controllable by the data producer if it is to inspire confidence. Geospatial Commission Research from 2019 found that 70% of users searching for data will use search engines rather than portals. So, datasets are competing with lots of other results, and people need to be able to find the right dataset amongst the noise. They may not have the specialist skills to use a portal, or knowledge of the correct terms to search for.



x

Layer Properties - high_water_polyline | Information

x

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

Some metadata elements can be easily derived from the dataset

Speaker notes

You can get at this sort of metadata in your file browser- think titles, locations, change dates, or tools like QGIS, which will show you spatial extents. What's more difficult is automatically doing this for lots of datasets, so we've got an ETL "Crawler" which automatically discovers data (flat files, database tables etc) and automatically derives this simple metadata.



We need help to derive some elements such as human-readable titles and descriptions

Speaker notes

Human-readable titles, descriptions, keywords and so on are much harder to do programatically, and also hard for users to do at scale. What works manually for a few records doesn't scale to thousands.

Increasing river temperatures are a threat to many of Scotland's freshwater species which are often adapted to live in cool environments. This includes ecologically and economically important freshwater fish species such as Atlantic salmon and brown trout. Management of riparian woodland is proven to protect cold water habitats. However, Scotland has ca. 108,000 km of rivers, of which only ca. 35% are protected by any substantial tree cover. Furthermore, the creation of new riparian woodland can be costly and logistically challenging compared to other forms of large scale woodland creation. It is therefore important that riparian tree planting is prioritised to areas where it can have greatest benefits for river temperature, specifically, where rivers are (1) hottest (2) most sensitive to climate change (see SRTMN Predictions: <http://marine.gov.scot/information/scotland-river-temperature-monitoring-network-srtmn-predictions-river-temperature-and>) and (3) can be effectively cooled by riparian woodland (see tree planting prioritisation layer). These three individual criteria can be combined with an equal weight to provide a single riparian woodland prioritisation score that looks to maximise the benefits of riparian tree planting for protecting Scotland's rivers from the adverse effects of climate change.

Details of the modelling work that produced the river temperature and climate sensitivity predictions can be found in the peer reviewed manuscript: Jackson et al (2018) 'A spatio-temporal statistical model of maximum daily river temperatures to inform the management of Scotland's Atlantic salmon rivers under climate change.' (see link under 'Additional Information' Tab).

IRL Example!

Details of the modelling work that identifies where riparian trees can have the greatest effect in reducing summer maximum river temperatures can be found in: Jackson, F.L., Hannah, D.M., Ouellet, V. and Malcolm, I.A. (2021) A deterministic river temperature model to prioritise management of riparian woodlands to reduce summer maximum river temperatures (see link under 'Additional Information' Tab).

Given the variety of potential tree planting options (southerly banks, northerly banks, both banks) and the need to scale results both nationally and locally, the outputs are illustrated as six layers on Marine Scotland Maps NMPI:

- 1. Nationally scaled tree planting prioritisation score where trees are planted on both banks*
- 2. Nationally scaled tree planting prioritisation score where trees are planted on only the most southerly bank*
- 3. Nationally scaled tree planting prioritisation score where trees are planted on only the most northerly bank*
- 4. Locally scaled tree planting prioritisation score where trees are planted on both banks*
- 5. Locally scaled tree planting prioritisation score where trees are planted on only the southerly bank*
- 6. Locally scaled tree planting prioritisation score where trees are planted on only the northerly bank*

Speaker notes

This is a real example of putting a complex block of text through our prototype ML workflow. We can generate keywords and geographic keywords, and auto-summarise to create a short description. Metadata keywords should be used inside the abstract. This needs to be done intelligently, eg as part of sentences rather than lists. You should include keywords that work for all levels of expertise, and take into consideration variations in spelling, synonyms and so on. Geographic keywords help to provide context for the user.

Increasing river temperatures are a threat to many of Scotland's freshwater species which are often adapted to live in cool environments. This includes ecologically and economically important freshwater fish species such as Atlantic salmon and brown trout. Management of riparian woodland is proven to protect cold water habitats. However, Scotland has ca. 108,000 km of rivers, of which only ca. 35% are protected by any substantial tree cover. Furthermore, the creation of new riparian woodland can be costly and logistically challenging compared to other forms of large scale woodland creation. It is therefore important that riparian tree planting is prioritised to areas where it can have greatest benefits for river temperature, specifically, where rivers are (1) hottest (2) most sensitive to climate change (see SRTMN Predictions: <http://marine.gov.scot/information/scotland-river-temperature-monitoring-network-srtmn-predictions-river-temperature-and>) and (3) can be effectively cooled by riparian woodland (see tree planting prioritisation layer). These three individual criteria can be combined with an equal weight to provide a single riparian woodland prioritisation score that looks to maximise the benefits of riparian tree planting for protecting Scotland's rivers from the adverse effects of climate change.

Details of the modelling work that produced the river temperature and climate sensitivity predictions can be found in the peer reviewed manuscript: Jackson et al (2018) 'A spatio-temporal statistical model of maximum daily river temperatures to inform the management of Scotland's Atlantic salmon rivers under climate change.' (see link under 'Additional Information' Tab).

IRL Example!

Details of the modelling work that identifies where riparian trees can have the greatest effect in reducing summer maximum river temperatures can be found in: Jackson, F.L., Hannah, D.M., Ouellet, V. and Malcolm, I.A. (2021) A deterministic river temperature model to prioritise management of riparian woodlands to reduce summer maximum river temperatures (see link under 'Additional Information' Tab).

Extract Keywords lochs, riparian, rivers, woodland

Given the variety of potential tree planting options (southerly banks, northerly banks, both banks) and the need to scale results both nationally and locally, the outputs are illustrated as six layers on Marine Scotland Maps NMPI:

- 1. Nationally scaled tree planting prioritisation score where trees are planted on both banks*
- 2. Nationally scaled tree planting prioritisation score where trees are planted on only the most southerly bank*
- 3. Nationally scaled tree planting prioritisation score where trees are planted on only the most northerly bank*
- 4. Locally scaled tree planting prioritisation score where trees are planted on both banks*
- 5. Locally scaled tree planting prioritisation score where trees are planted on only the southerly bank*
- 6. Locally scaled tree planting prioritisation score where trees are planted on only the northerly bank*

Speaker notes

This is a real example of putting a complex block of text through our prototype ML workflow. We can generate keywords and geographic keywords, and auto-summarise to create a short description. Metadata keywords should be used inside the abstract. This needs to be done intelligently, eg as part of sentences rather than lists. You should include keywords that work for all levels of expertise, and take into consideration variations in spelling, synonyms and so on. Geographic keywords help to provide context for the user.

Increasing river temperatures are a threat to many of Scotland's freshwater species which are often adapted to live in cool environments. This includes ecologically and economically important freshwater fish species such as Atlantic salmon and brown trout. Management of riparian woodland is proven to protect cold water habitats. However, Scotland has ca. 108,000 km of rivers, of which only ca. 35% are protected by any substantial tree cover. Furthermore, the creation of new riparian woodland can be costly and logistically challenging compared to other forms of large scale woodland creation. It is therefore important that riparian tree planting is prioritised to areas where it can have greatest benefits for river temperature, specifically, where rivers are (1) hottest (2) most sensitive to climate change (see SRTMN Predictions: <http://marine.gov.scot/information/scotland-river-temperature-monitoring-network-srtmn-predictions-river-temperature-and>) and (3) can be effectively cooled by riparian woodland (see tree planting prioritisation layer). These three individual criteria can be combined with an equal weight to provide a single riparian woodland prioritisation score that looks to maximise the benefits of riparian tree planting for protecting Scotland's rivers from the adverse effects of climate change.

Details of the modelling work that produced the river temperature and climate sensitivity predictions can be found in the peer reviewed manuscript: Jackson et al (2018) 'A spatio-temporal statistical model of maximum daily river temperatures to inform the management of Scotland's Atlantic salmon rivers under climate change.' (see link under 'Additional Information' Tab).

IRL Example!

Details of the modelling work that identifies where riparian trees can have the greatest effect in reducing summer maximum river temperatures can be found in: Jackson, F.L., Hannah, D.M., Ouellet, V. and Malcolm, I.A. (2021) A deterministic river temperature model to prioritise management of riparian woodlands to reduce summer maximum river temperatures (see link under 'Additional Information' Tab).

Geoparse scotland, atlantic

Given the variety of potential tree planting options (southerly banks, northerly banks, both banks) and the need to scale results both nationally and locally, the outputs are illustrated as six layers on Marine Scotland Maps NMPI:

- 1. Nationally scaled tree planting prioritisation score where trees are planted on both banks*
- 2. Nationally scaled tree planting prioritisation score where trees are planted on only the most southerly bank*
- 3. Nationally scaled tree planting prioritisation score where trees are planted on only the most northerly bank*
- 4. Locally scaled tree planting prioritisation score where trees are planted on both banks*
- 5. Locally scaled tree planting prioritisation score where trees are planted on only the southerly bank*
- 6. Locally scaled tree planting prioritisation score where trees are planted on only the northerly bank*

Speaker notes

This is a real example of putting a complex block of text through our prototype ML workflow. We can generate keywords and geographic keywords, and auto-summarise to create a short description. Metadata keywords should be used inside the abstract. This needs to be done intelligently, eg as part of sentences rather than lists. You should include keywords that work for all levels of expertise, and take into consideration variations in spelling, synonyms and so on. Geographic keywords help to provide context for the user.

Increasing river temperatures are a threat to many of Scotland's freshwater species which are often adapted to live in cool environments. This includes ecologically and economically important freshwater fish species such as Atlantic salmon and brown trout. Management of riparian woodland is proven to protect cold water habitats. However, Scotland has ca. 108,000 km of rivers, of which only ca. 35% are protected by any substantial tree cover. Furthermore, the creation of new riparian woodland can be costly and logistically challenging compared to other forms of large scale woodland creation. It is therefore important that riparian tree planting is prioritised to areas where it can have greatest benefits for river temperature, specifically, where rivers are (1) hottest (2) most sensitive to climate change (see SRTMN Predictions: <http://marine.gov.scot/information/scotland-river-temperature-monitoring-network-srtmn-predictions-river-temperature-and>) and (3) can be effectively cooled by riparian woodland (see tree planting prioritisation layer). These three individual criteria can be combined with an equal weight to provide a single riparian woodland prioritisation score that looks to maximise the benefits of riparian tree planting for protecting Scotland's rivers from the adverse effects of climate change.

Details of the modelling work that produced the river temperature and climate sensitivity predictions can be found in the peer reviewed manuscript: Jackson et al (2018) 'A spatio-temporal statistical model of maximum daily river temperatures to inform the management of Scotland's Atlantic salmon rivers under climate change.' (see link under 'Additional Information' Tab).

IRL Example!

Details of the modelling work that identifies where riparian trees can have the greatest effect in reducing summer maximum river temperatures can be found in: Jackson, F.L., Hannah, D.M., Ouellet, V. and Malcolm, I.A. (2021) A deterministic river temperature model to prioritise management of riparian woodlands to reduce summer maximum river temperatures (see link under 'Additional Information' Tab).

Auto-Summarise

Given the variety of potential tree planting options (southerly banks, northerly banks, both banks) and the need to prioritise management of riparian woodlands to reduce summer maximum river temperatures and a riparian woodland prioritisation score that looks to maximise the benefits of Riparian tree planting prioritisation score where trees are planted on both banks prioritisation score where trees are planted on only the most southerly bank prioritisation score where trees are planted on only the most northerly bank prioritisation score where trees are planted on both banks prioritisation score where trees are planted on only the southerly bank

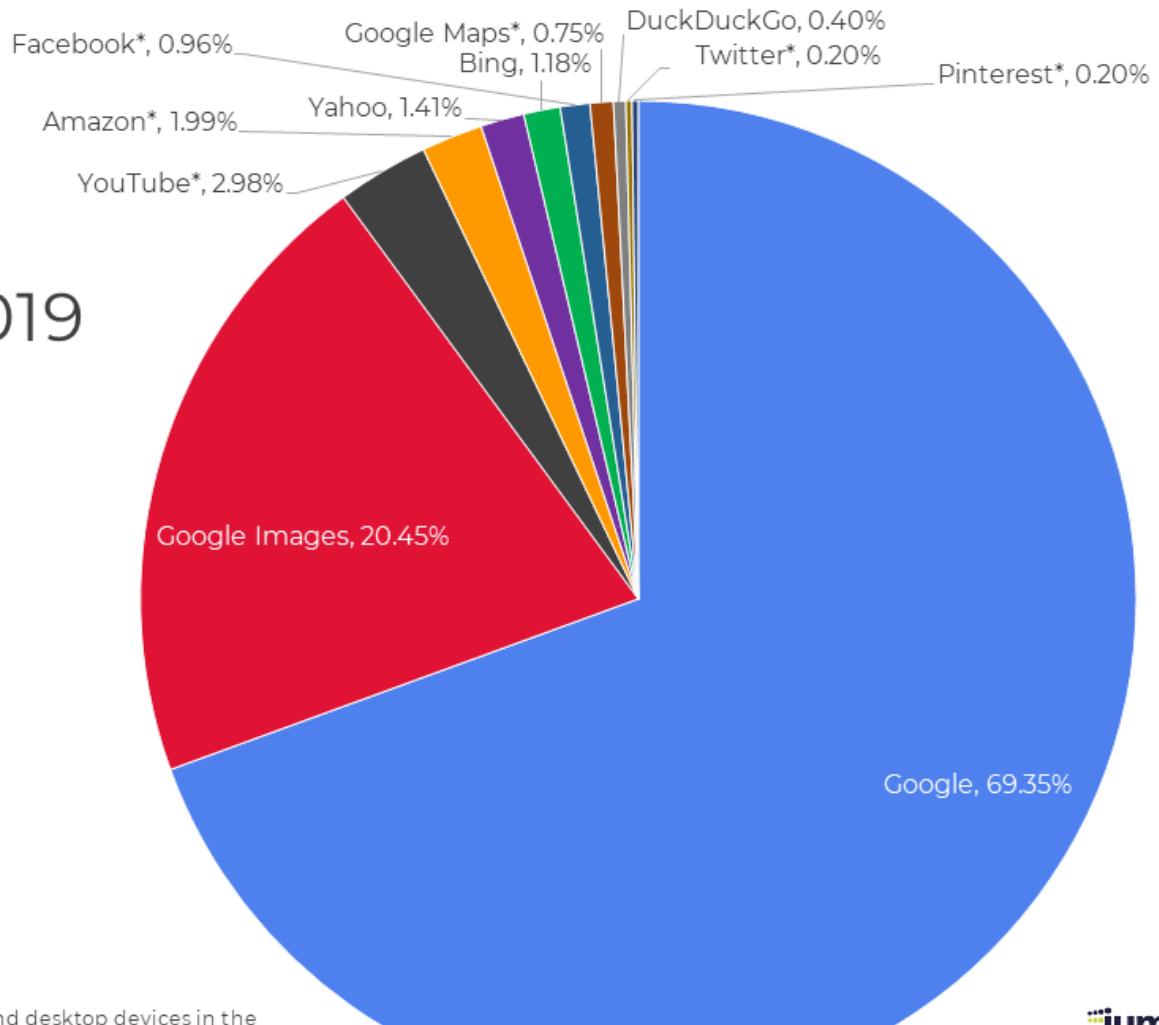
Speaker notes

This is a real example of putting a complex block of text through our prototype ML workflow. We can generate keywords and geographic keywords, and auto-summarise to create a short description. Metadata keywords should be used inside the abstract. This needs to be done intelligently, eg as part of sentences rather than lists. You should include keywords that work for all levels of expertise, and take into consideration variations in spelling, synonyms and so on. Geographic keywords help to provide context for the user.

Help search engines understand your
data

Search Engine Market Share Q2 2019

94%
of all searches happen
on a **Google** property



* Data from 230B+ browser-based searches on millions of mobile and desktop devices in the

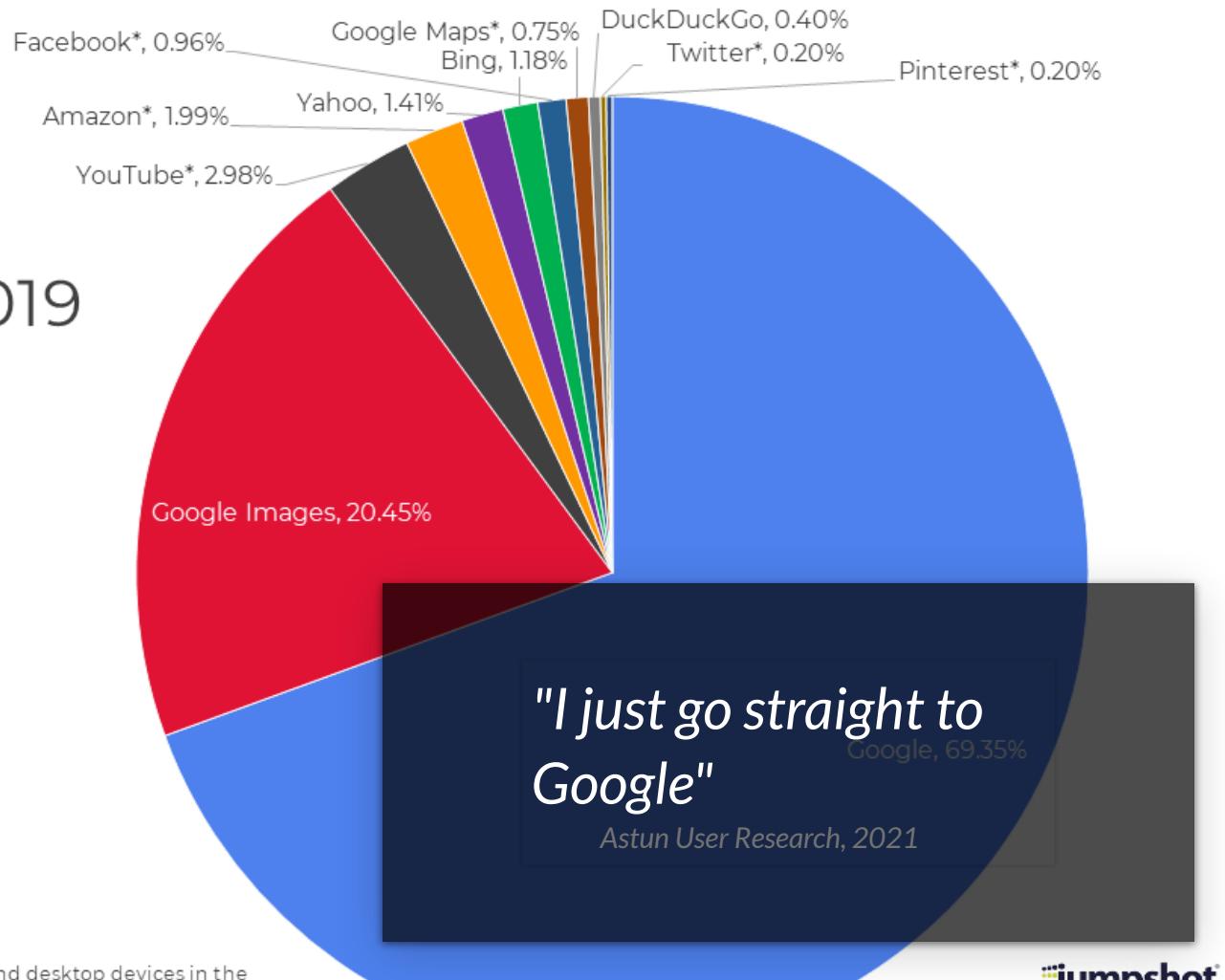
iumshot

Speaker notes

There's no point in adding extra complexity to people's workflows- they are going to use search engines anyway so let's just make it as easy as possible to do that. Google has a 94% market share across all of its products

Search Engine Market Share Q2 2019

94%
of all searches happen
on a **Google** property



* Data from 230B+ browser-based searches on millions of mobile and desktop devices in the



Speaker notes

There's no point in adding extra complexity to people's workflows- they are going to use search engines anyway so let's just make it as easy as possible to do that. Google has a 94% market share across all of its products

[Back to search](#)< PreviousNext >[Download](#) ▾[Display mode](#) ▾

Dedicated Woodlands

Description:

The Dedication Scheme (Basis I & II) was introduced in 1947 in order to encourage landowners to retain their land in forestry and to introduce good forestry practice. Basis III was introduced in 1974, providing grants for new planting and additional supplements for broadleaves.

The Dedication Scheme was closed to new applications in 1981. Land still under Dedication could continue to be within the scheme but Dedication would terminate on a change of ownership.

Dedication schemes without a Plan of Operations and therefore receiving no grant, are deemed to be under Negative Covenant.

Dataset Attributes:

Descriptor Dataset name
Case_Name Dedication Scheme name
Case_No Dedication Scheme reference number
Basis Basis number of scheme (I, II or III)
Date_Appr Date woodland became Dedicated
PlanOfOps Period of current (or last) Plan of Operations
Covenant Scheme in 'Positive' or 'Negative' Covenant
Grid_Ref Grid Reference of property
Total_Area Total area of scheme (* OL not always included)
Man_Area Total area of scheme under management

Download and links

[Scottish Forestry Open Data](#)[Open link](#)

Spatial extent



Temporal extent

Creation date

2002-09-01

Revision date

2019-11-01

Speaker notes

This is done automatically in GeoNetwork at least, it doesn't require user intervention

[Back to search](#)[Previous](#)[Next >](#)[Download](#)[Display mode](#)

Dedicated Woodlands

Description:

The Dedication Scheme (Basis I & II) was introduced in 1947 in order to encourage landowners to retain their land in forestry and to introduce good forestry practice. Basis III was introduced in 1974, providing grants for new planting and additional supplements for broadleaves.

The Dedication Scheme was closed to new applications in 1981. Land still under Dedication could continue to be within the scheme but Dedication would terminate on a change of ownership.

Dedication schemes without a Plan of Operations and therefore receiving no grant, are deemed to be under Negative Covenant.

Dataset Attributes:

Descriptor Dataset name
Case_Name Dedication Scheme name
Case_No Dedication Scheme reference number
Basis Basis number of scheme (I, II or III)
Date_Appr Date woodland became Dedicated
PlanOfOps Period of current (or last) Plan of Operations
Covenant Scheme in 'Positive' or 'Negative' Covenant
Grid_Ref Grid Reference of property
Total_Area Total area of scheme (* OL not always included)
Man_Area Total area of scheme under management

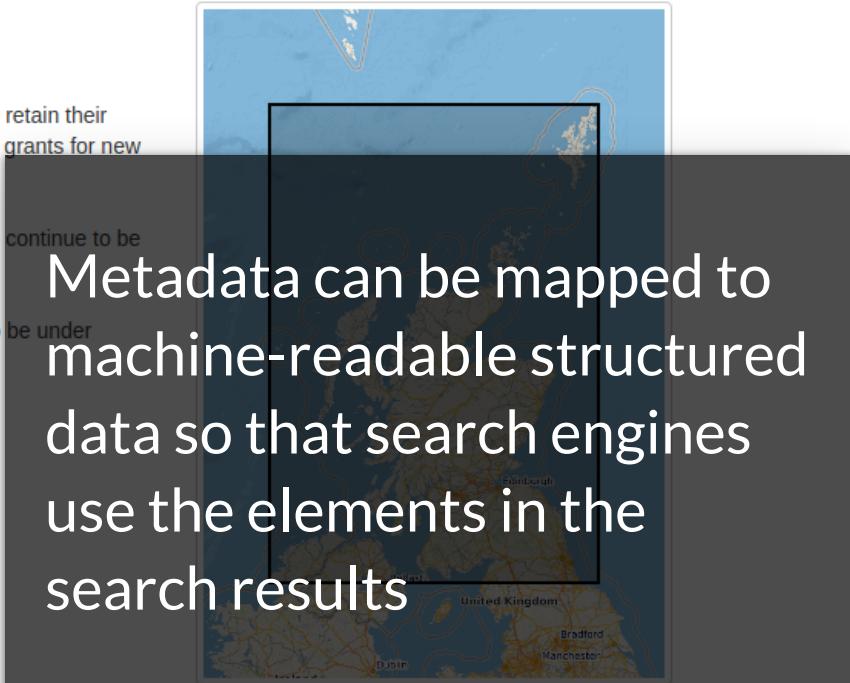
Download and links



Scottish Forestry Open Data

[Open link](#)**Revision date**
2019-11-01

Spatial extent

<https://data.gov.uk> > dataset > dedicated-woodlands**Dedicated Woodlands - data.gov.uk**

23 Jul 2020 — Description: The Dedication Scheme (Basis I & II) was introduced in 1947 in order to encourage landowners to retain their land in forestry ...

Speaker notes

This is done automatically in GeoNetwork at least, it doesn't require user intervention



Getting your data to appear high up in search results

A photograph of Leonardo DiCaprio as Jordan Belfort from the movie "The Wolf of Wall Street". He is wearing a black tuxedo and a white shirt with a bow tie. He is looking directly at the camera with a slight smile and is holding a clear wine glass in his right hand, which is raised slightly. The background is dark with some blurred lights and other people.

Quick wins (aka clickbait)

Speaker notes

Ensuring that the title and description/abstract are present and not too short or long, and ensuring you don't have duplicate titles or broken links are simple fixes that ensure the web pages are not seen as low-quality and hence downgraded in search results.

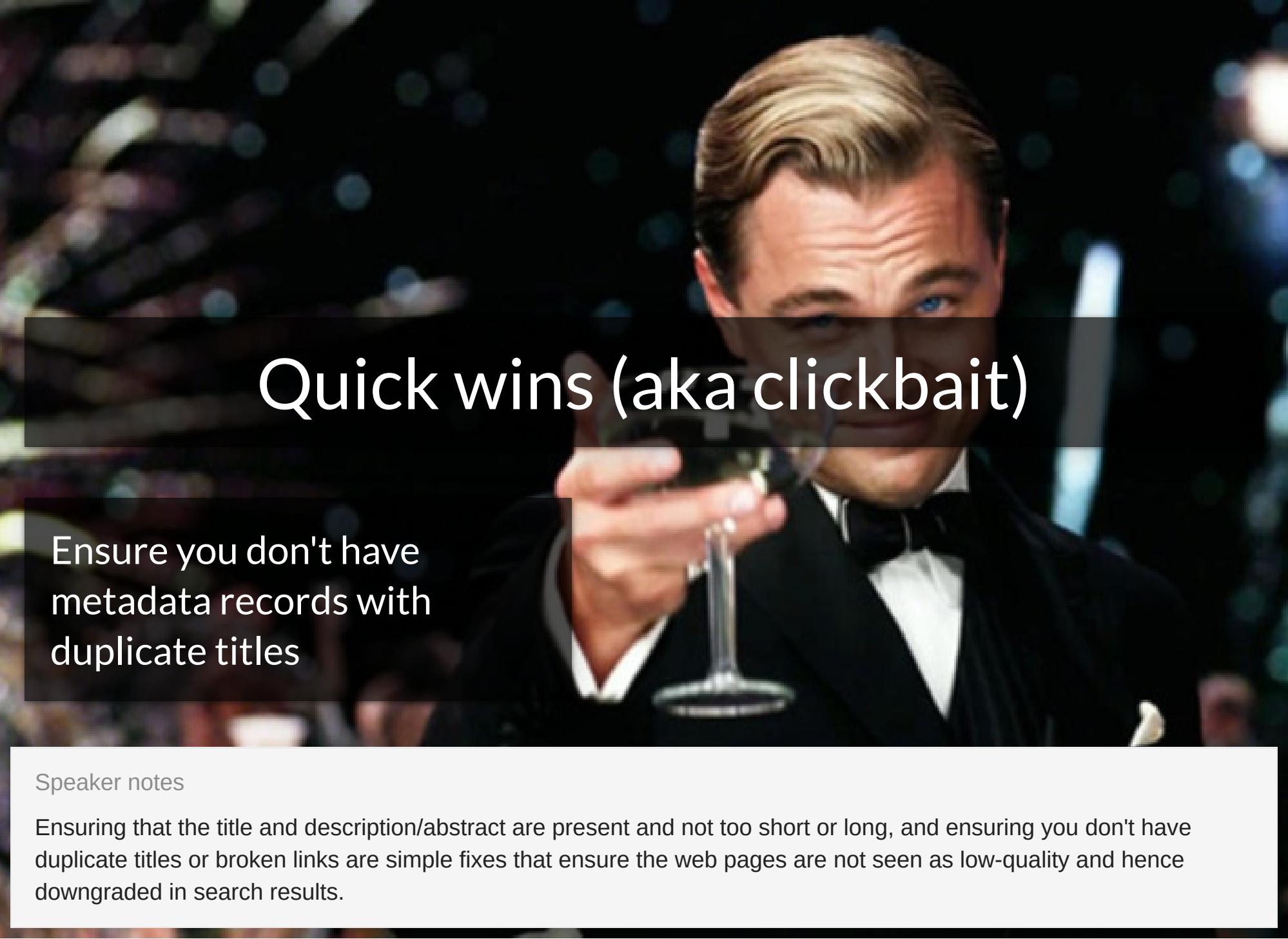
A photograph of Leonardo DiCaprio as Jordan Belfort from the movie 'The Wolf of Wall Street'. He is wearing a dark tuxedo and white shirt, looking slightly upwards and to the side with a serious expression. He is holding a clear cocktail glass in his right hand, which is raised towards the camera.

Quick wins (aka clickbait)

Ensure that the title and abstract are present and follow basic SEO guidelines for length and language

Speaker notes

Ensuring that the title and description/abstract are present and not too short or long, and ensuring you don't have duplicate titles or broken links are simple fixes that ensure the web pages are not seen as low-quality and hence downgraded in search results.

A dramatic close-up of Leonardo DiCaprio's face as his character Jordan Belfort. He has blonde hair and is looking intensely at the camera with a serious expression. His right hand is raised, holding a clear wine glass with a dark liquid inside. The background is dark and out of focus, showing some blurred lights and other people.

Quick wins (aka clickbait)

Ensure you don't have
metadata records with
duplicate titles

Speaker notes

Ensuring that the title and description/abstract are present and not too short or long, and ensuring you don't have duplicate titles or broken links are simple fixes that ensure the web pages are not seen as low-quality and hence downgraded in search results.

A dramatic close-up of Leonardo DiCaprio's face as Jordan Belfort. He has blonde hair and is looking intensely at the camera with a serious expression. His right hand is raised, holding a clear cocktail glass with a dark liquid inside. The background is dark and out of focus, showing some blurred lights and other people.

Quick wins (aka clickbait)

Fix broken links

Speaker notes

Ensuring that the title and description/abstract are present and not too short or long, and ensuring you don't have duplicate titles or broken links are simple fixes that ensure the web pages are not seen as low-quality and hence downgraded in search results.



Speaker notes

Do people stay on the site, and go on to download a dataset, or do they leave very quickly? Are particular datasets more "popular"? Are people commonly mis-spelling a term, or using gaelic or welsh terms? Get direct qualitative feedback from users too



**YOUR
FEEDBACK
Feedback
MATTERS!**

Web analytics, Google trends, search statistics can provide useful insights into who's using your data

Speaker notes

Do people stay on the site, and go on to download a dataset, or do they leave very quickly? Are particular datasets more "popular"? Are people commonly mis-spelling a term, or using gaelic or welsh terms? Get direct qualitative feedback from users too

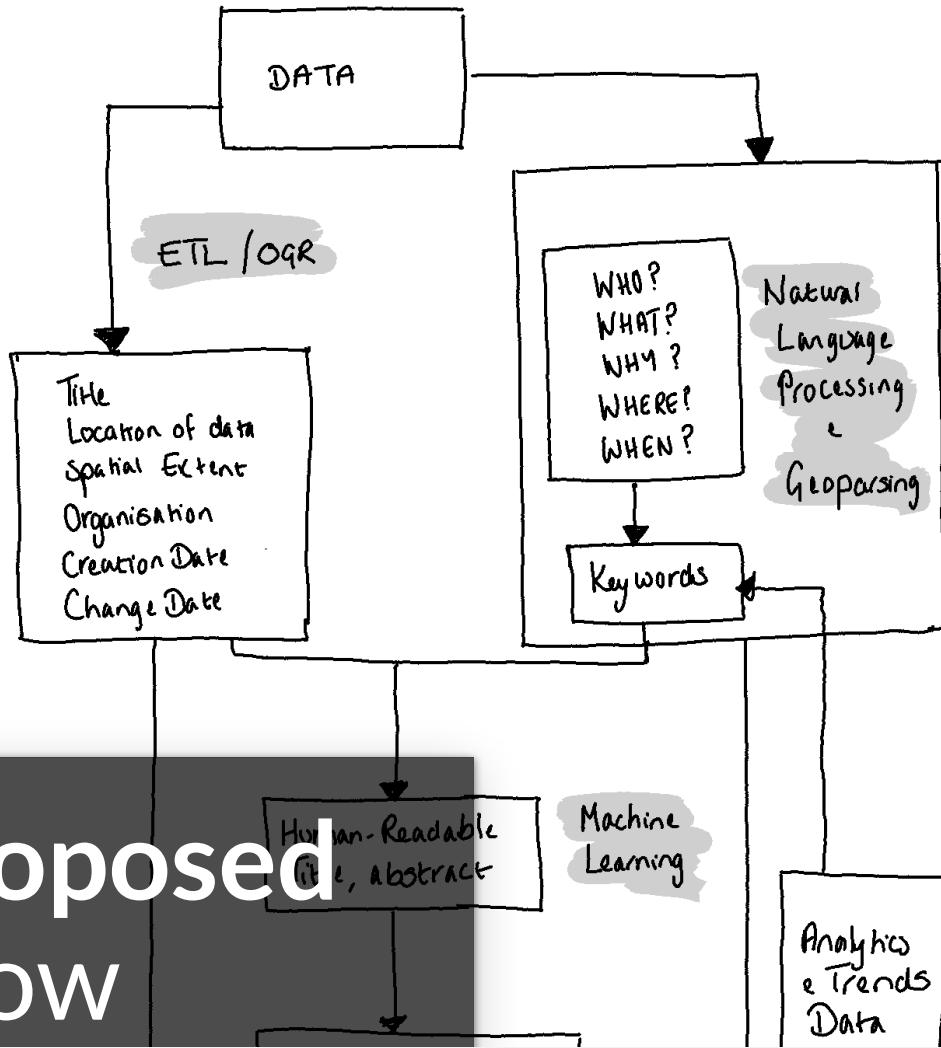


**YOUR
FEEDBACK
Feedback
MATTERS!**

Give people the chance to submit feedback (good or bad)

Speaker notes

Do people stay on the site, and go on to download a dataset, or do they leave very quickly? Are particular datasets more "popular"? Are people commonly mis-spelling a term, or using gaelic or welsh terms? Get direct qualitative feedback from users too



Our proposed workflow

Speaker notes

Our proposed workflow starts with finding the data and doing both automatic extraction of simple metadata, and using ML to derive the other elements. Then we pull them together to create the full metadata record. People can review this, and see it internally, or publish it so it's visible to google. The search analytics and usage statistics can feed back into the keywords and hence improve the metadata in a virtuous circle.



Where we are at the moment

Speaker notes

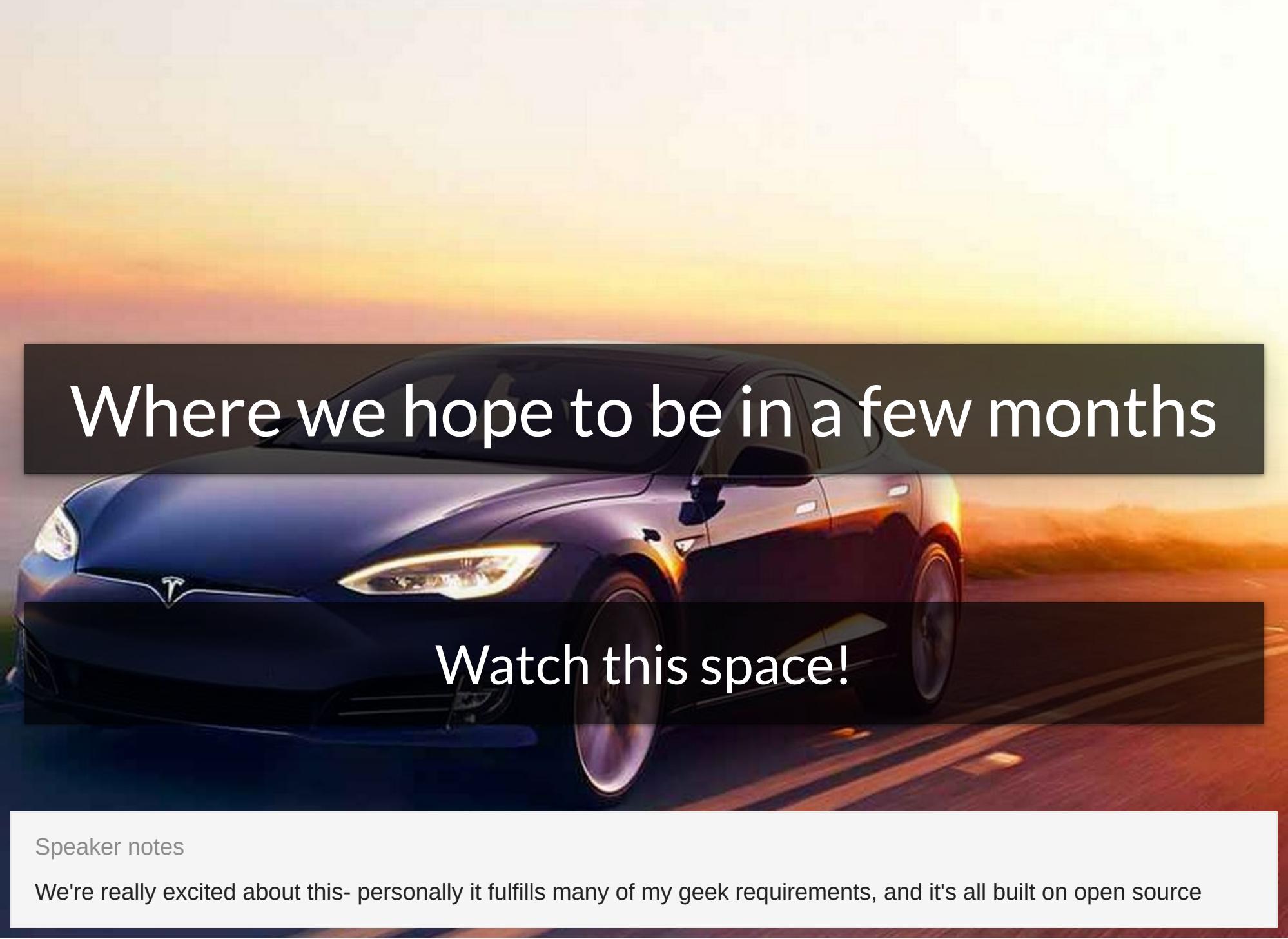
We have all the pieces but need to glue them together into a usable product. It's a bit heath robinson right now. There are lots of moving parts, and areas where we're going to get expert assistance such as on SEO (to test our ML results against what they would do) and to ensure we're using the right ML models and so on.

A blue Tesla Model S is shown from a front three-quarter perspective, driving on a road. The background features a warm, orange and yellow sunset sky. The car's headlights are on, and its sleek lines are highlighted by the low light.

Where we hope to be in a few months

Speaker notes

We're really excited about this- personally it fulfills many of my geek requirements, and it's all built on open source

A blue Tesla Model S is shown from a front-three-quarter angle, driving on a road. The background is a blurred sunset or sunrise, suggesting speed. The car's headlights are on.

Where we hope to be in a few months

Watch this space!

Speaker notes

We're really excited about this- personally it fulfills many of my geek requirements, and it's all built on open source

A close-up photograph of a white cat's face, focusing on its large green eyes. The cat is looking directly at the camera. It is positioned above a stack of several thick, greyish-blue books. The background is blurred, showing warm, out-of-focus colors.

Thank you for having me!

jocook@astuntechnology.com | @archaeogeek