

Improving Dataset Quality through Metadata

Jo Cook | Astun Technology

Our work focuses on GeoNetwork Open Source, but also metadata in general and the wider geospatial standards infrastructure

Who are Astun Technology?

People now understand the importance of data sharing and metadata, so we can now focus on improving what we share. Government Geospatial Commission, Data Quality Hub, EU Metadata Quality Assessment etc provide guidance

Our Starting Point

Data and Metadata Quality both contribute to Data Discovery and the FAIR principles. For example, if titles are too long or too short, abstracts/descriptions are truncated, duplicated records and so on, this can cause indexed records to be down-ranked and therefore make them less findable

Everyone wants to make Data FAIR



Encoding data quality and lineage in metadata is technically difficult. Standards documents cost money and are not the easiest to understand

But...

**There's little technical guidance on how to apply
this In Real Life**

Some Practical Fixes...

We've made improvements to the Gemini 2.3 plugin for Geonetwork, part funded by Scottish Government around both ease of use and machine-readability

Gemini 2.3

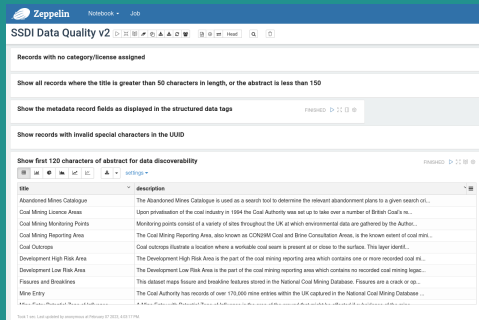
We've created snippets for the data quality elements so they can be picked from a list when editing records. We've done the same for lineage, to make it easier for people to include record more complex workflows and data lifecycles

Data Quality and Lineage

There's more to come! We're developing more dashboards and reporting based on what people ask us for.

Reporting

Dashboards for SEO and Data Quality metrics



Zeppelin Notebook - Job

SSDI Data Quality v2

Records with no category/license assigned

Show all records where the title is greater than 50 characters in length, or the abstract is less than 150

Show the metadata record fields as displayed in the structured data tags FINISHED 0 1 0 0

Show records with invalid special characters in the UUID

Show first 120 characters of abstract for data discoverability FINISHED 0 1 0 0

title	description
Abandoned Mines Catalogue	The Abandoned Mines Catalogue is used as a search tool to determine the relevant abandonment plans to a given search...
Coal Mining License Areas	Given proliferation of the coal industry in 1980 the Coal Authority was set up to take over a number of British Coal li...
Coal Mining Monitoring Points	Monitoring points consist of a variety of sites throughout the UK at which environmental data are gathered by the Author...
Coal Mining Reporting Areas	The Coal Mining Reporting Areas, also known as CONIRM Coal and Brown Consultation Areas, is the known extent of coal min...
Coal Outcrops	Coal outcrops (flintlike a location where a veinlike coal seam is present at or close to the surface. This layer ident...
Development High Risk Area	The Development High Risk Area is the part of the coal mining reporting areas which contains one or more recorded coal m...
Development Low Risk Area	The Development Low Risk Area is the part of the coal mining reporting areas which contains no recorded coal mining legac...
Features and Breaklines	This dataset maps feature and breakline features stored in the National Coal Mining Database. Features are a creek or sp...
Mine Entry	The Coal Authority has records of over 170,000 mine entries within the UK captured in the National Coal Mining Database...

View | Refresh | Download | Settings | Help

© 2015 Coal Authority. All rights reserved. | Privacy Policy | Terms of Use

