

Improving Dataset Quality through Metadata

Jo Cook | Astun Technology

Who are Astun Technology?

Who are Astun Technology?

We manage and host the SSDI for Scottish Government...

Who are Astun Technology?

...I sit on BSI IST36

Who are Astun Technology?

...chair the UK Gemini WG

Who are Astun Technology?

...and sit on the GeoNetwork PSC



Our Starting Point





Our Starting Point

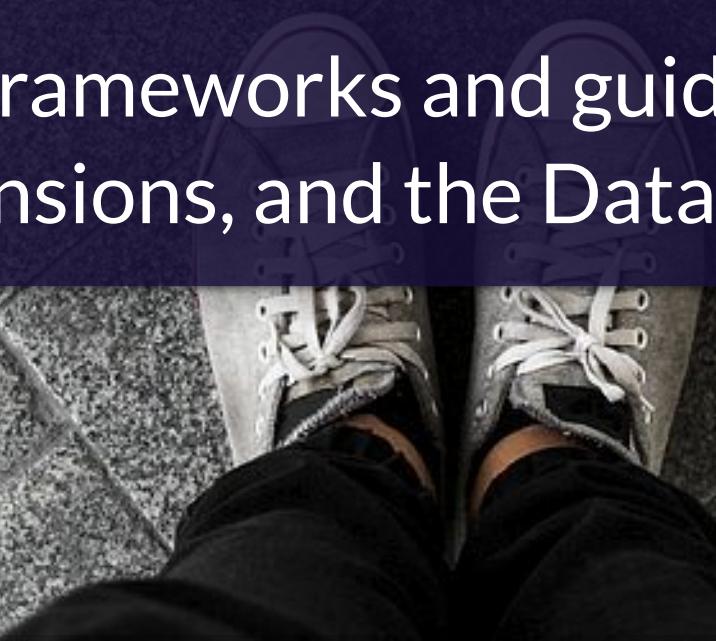
There's an increasing focus on Data Quality across the Public Sector...





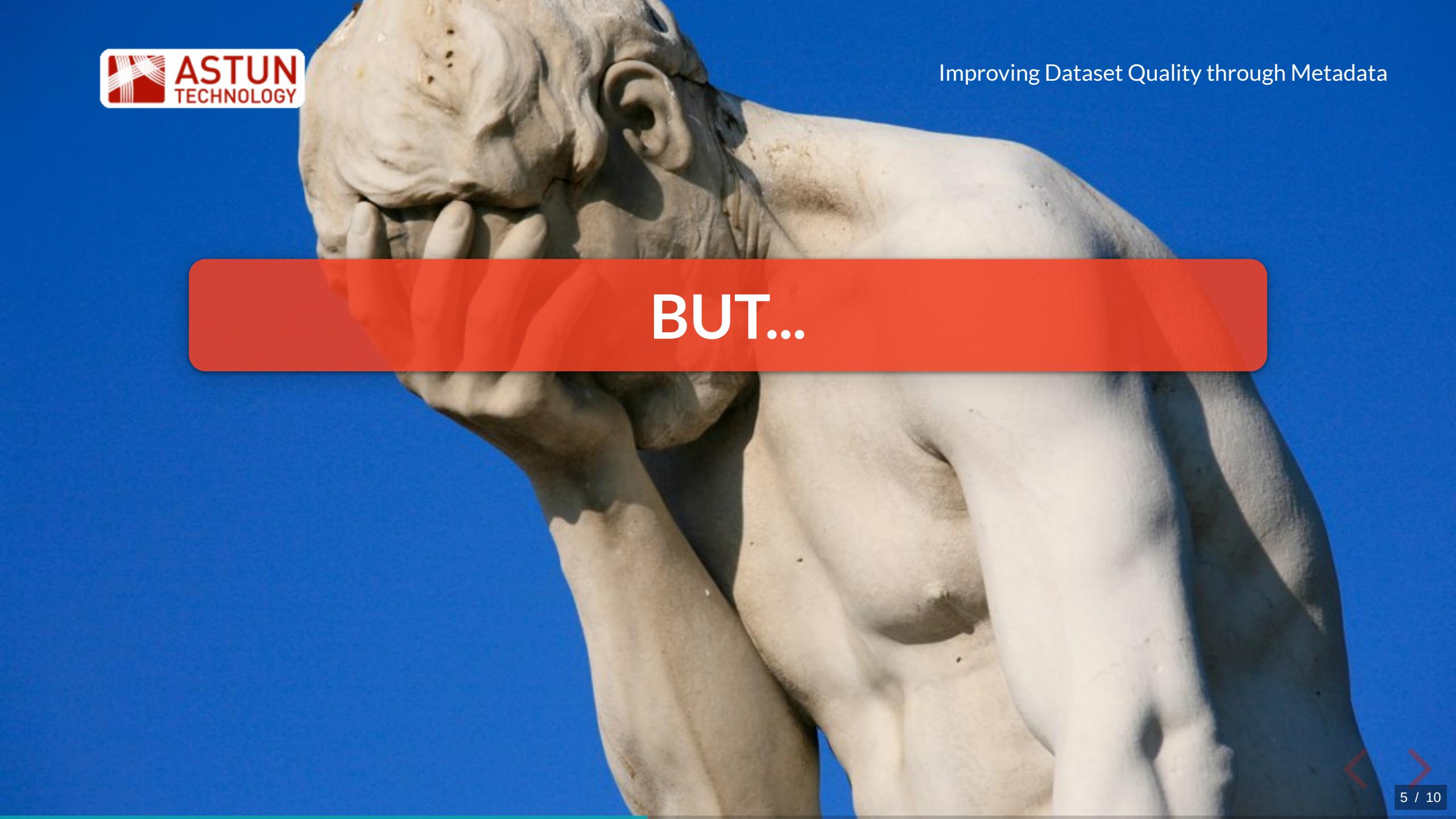
Our Starting Point

and many frameworks and guidance on Quality Dimensions, and the Data Lifecycle...

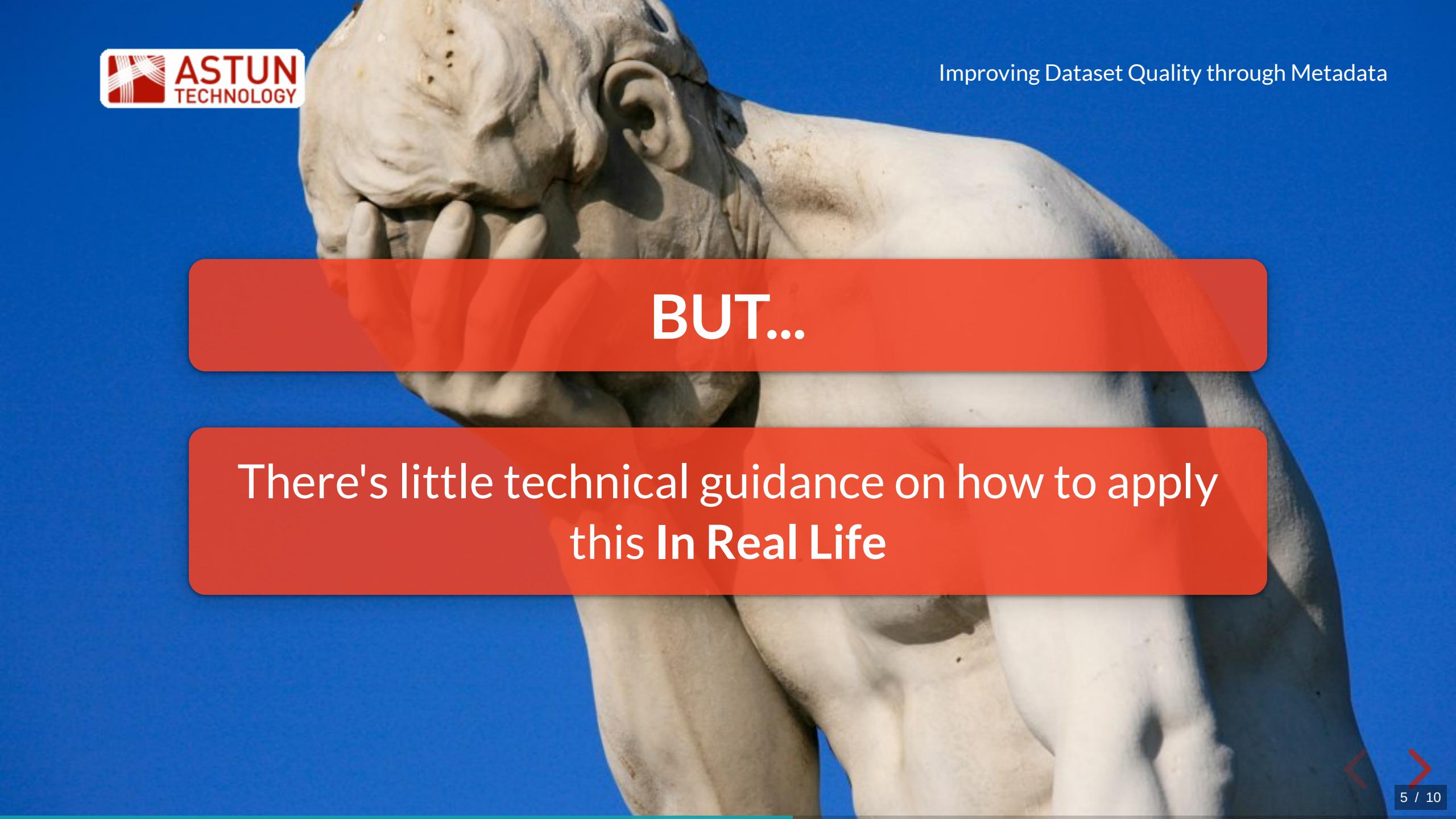


and everyone wants to make Data FAIR...





BUT...



BUT...

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

Some Practical Fixes...

Gemini 2.3

Gemini 2.3

Easier to use ➤

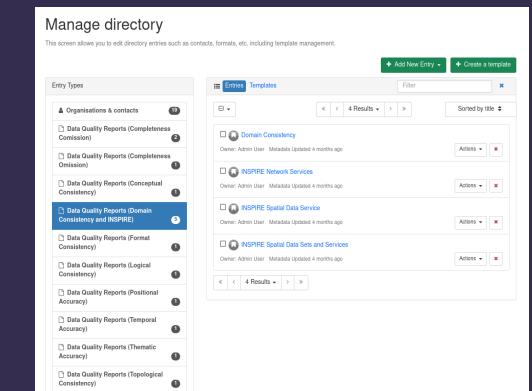
Gemini 2.3

Better machine-readable output ➡

Data Quality and Lineage

Data Quality and Lineage

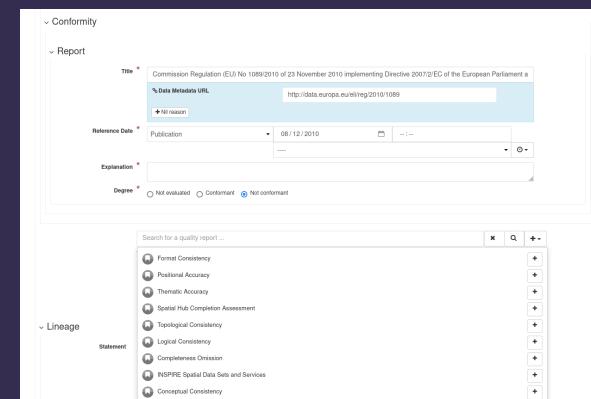
Snippets



The screenshot shows a 'Manage directory' interface. On the left, there's a sidebar titled 'Entry Types' with a tree view. Under 'Organisations & contacts', several items are listed, including 'Data Quality Reports (Completeness Consistency)', 'Data Quality Reports (Completeness Omission)', 'Data Quality Reports (Conceptual Consistency)', and 'Data Quality Reports (Domain Consistency and INSPIRE)'. Other categories like 'Data Quality Reports (Format Consistency)' and 'Data Quality Reports (Logical Consistency)' are also visible. On the right, there's a main panel titled 'Entries' with a table showing four results. The first result is 'Domain Consistency' (Owner: Admin User, Metadata Updated 4 months ago). The second is 'INSPIRE Network Services' (Owner: Admin User, Metadata Updated 4 months ago). The third is 'INSPIRE Spatial Data Service' (Owner: Admin User, Metadata Updated 4 months ago). The fourth is 'INSPIRE Spatial Data Sets and Services' (Owner: Admin User, Metadata Updated 4 months ago). Each entry has an 'Actions' button.

Data Quality and Lineage

Drop-down pick lists

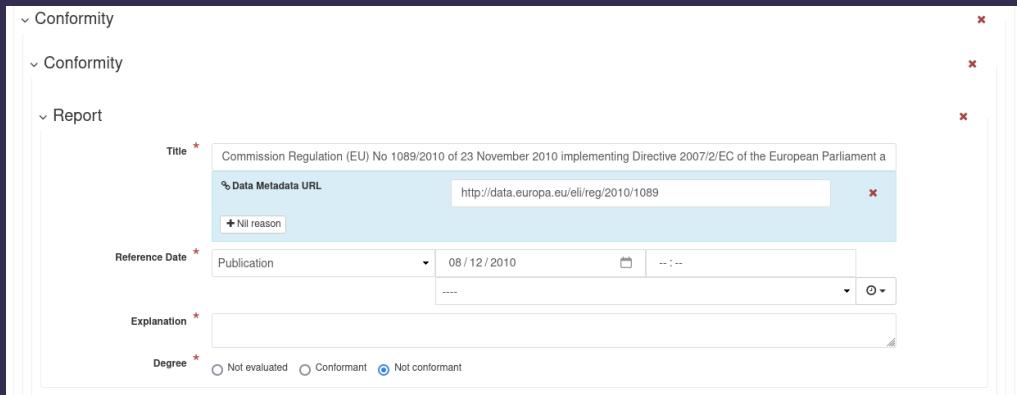


The screenshot shows a user interface for data quality reporting. At the top, there's a header with 'Conformity' and 'Report' sections. Under 'Report', there are fields for 'Title' (set to 'Commission Regulation (EU) No 1089/2010 of 23 November 2010 implementing Directive 2007/2/EC of the European Parliament and of the Council on the establishment of a Spatial Data Infrastructure in the European Union'), 'Data Metadata URL' (with a link to 'http://data.europa.eu/eu/reg/2010/1089'), and 'NI reason'. Below these are 'Reference Date' (set to 'Publication' on '08/12/2010') and 'Explanation' (a large text area). A 'Degree' section includes radio buttons for 'Not evaluated', 'Conform', and 'Not conformant' (which is selected). At the bottom, there's a search bar for 'Search for a quality report ...' and a list of quality statements under 'Statement'.

Statement	Action
Format Consistency	+/-
Positional Accuracy	+/-
Thematic Accuracy	+/-
Spatial Hub Completion Assessment	+/-
Topographical Consistency	+/-
Logical Consistency	+/-
Completeness/Omission	+/-
INSPIRE Spatial Data Sets and Services	+/-
Conceptual Consistency	+/-

Data Quality and Lineage

Visual changes for ambiguous elements

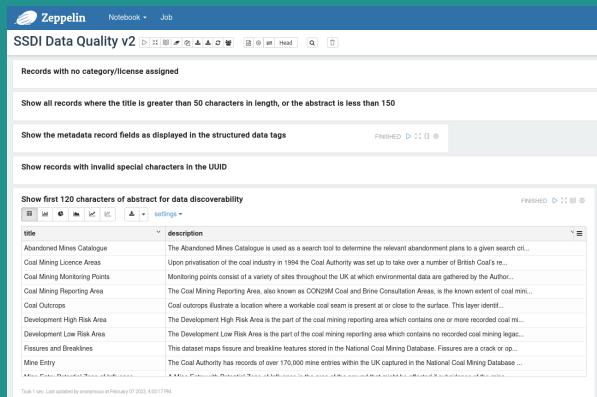


The screenshot shows a user interface for data entry, specifically for a 'Report' related to a 'Conformity' check. The interface includes the following fields:

- Title:** Commission Regulation (EU) No 1089/2010 of 23 November 2010 implementing Directive 2007/2/EC of the European Parliament a
- Data Metadata URL:** <http://data.europa.eu/eli/reg/2010/1089>
- Reference Date:** Publication date: 08 / 12 / 2010
- Explanation:** A large text area for notes.
- Degree:** Radio buttons for 'Not evaluated', 'Conformant', and 'Not conformant'. The 'Not conformant' option is selected.

Reporting

Dashboards for SEO and Data Quality metrics



The screenshot shows a Zeppelin notebook titled "SSDI Data Quality v2". The notebook interface includes a toolbar with various icons for file operations like Open, Save, and Print. Below the toolbar, there are several sections of text and tables:

- "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" (status: FINISHED)
- "Show records with invalid special characters in the UUID"
- "Show first 120 characters of abstract for data discoverability" (status: FINISHED)

Below these sections is a table with two columns: "title" and "description". The table lists various data quality issues:

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 criteria.
Coal Mining Licence Areas	Upon privatisation of the coal industry in 1994 the Coal Authority was set up to take over a number of British Coal's responsibilities.
Coal Mining Monitoring Points	Monitoring points consist of a variety of sites throughout the UK at which environmental data are gathered by the Authority.
Coal Mining Reporting Area	The Coal Mining Reporting Area, also known as CONCORM Coal and Brine Consultation Areas, is the known extent of coal mining in Great Britain.
Coal Outcrops	Coal outcrops illustrate a location where a workable coal seam is present at or close to the surface. This layer identifies locations of coal outcrops.
Development High Risk Area	The Development High Risk Area is the part of the coal mining reporting area which contains one or more recorded coal mining leases.
Development Low Risk Area	The Development Low Risk Area is the part of the coal mining reporting area which contains no recorded coal mining leases.
Fissures and Breaklines	This dataset maps fissure and breakline features stored in the National Coal Mining Database. Fissures are a crack or opening in rock.
Mine Entry	The Coal Authority has records of over 170,000 mine entries within the UK captured in the National Coal Mining Database.

At the bottom of the notebook, a footer note states: "Data last updated by anonymous at February 07 2023, 4:10:17 PM".



Any Questions?