IMPROVING METADATA PRACTICES FOR NERC'S DATA CENTRES TO ENHANCE DATA FAIRNESS

A report by the NERC Digital Solutions Programme team.

Contents

Introduction

Metadata is used to describe and provide information about data, so that it can be correctly understood, interpreted and used. As we move towards an ever greater integration of data services and resources, it is essential that metadata is generated in alignment with the F.A.I.R. principles (https://www.go-fair.org/). These principles offer essential guidelines aimed at enhancing the Findability, Accessibility, Interoperability, and Reusability of data.

NERC's datasets are accompanied by geographical metadata, stored in XML files, which comply with the GEMINI UK standard. These metadata files provide a comprehensive range of information about the associated datasets, including details on the project, instruments, locations, and algorithms used in data production. By adhering to GEMINI UK, the metadata aims to enhance data FAIRness through the use of established methods for describing and publishing geographical data (GEMINI UK is based on ISO 19139 and the INSPIRE directive).

Despite NERC's significant efforts to comply with established standards, the initial exploratory phase led by the Digital Solution Programme (DSP) revealed that the state of NERC's data / metadata is less than ideal. This poses several challenges that prevent users from experiencing a seamless interaction with the data, with regard to various facets of the FAIR principles.

To effectively enhance the discovery and utilization of NERC's data, it is essential for the DSP to develop a comprehensive understanding of the metadata associated with these datasets.

State of data / metadata in NERC's data centres

The state of NERC's data / metadata, at the time of the exploratory phase led by the DSP, can be summarised under the following points:

- When users upload a dataset to NERC's data centres (DCs), they are required to input metadata using free text fields rather than selecting from a predefined set of values. This practice can lead to inconsistencies, redundancy in terminology, and challenges in filtering metadata entries, although this may not apply uniformly across all DCs.
- The datasets schemas are not documented in the corresponding metadata entry (although the ISO standards allow for it). The schemas are instead detailed in external files (Word, PDF, etc) documents that lack standardisation and coherence for machine data-consumability.
- Problematic lineage/versions documentation of datasets. This results in redundancy and the inability to identify in an automatic fashion the most recent version of a dataset.
- Incorrect URLs for datasets and incorrect/absent URLs for service's endpoints. This has obvious consequences in terms of findability, accessibility and reusability of data.
- Many datasets are not accessible through services (WMS, WFS, WCS). This means that data preview is not possible, impairing the assessment of the dataset relevance (licensing limitations should be reconsidered...?).
- Difficulty following the hierarchy of Projects, Dataset Collections and Datasets. This is not a problem of metadata as such, but of how data pertaining to the same project or the same collection is scattered in different datasets and presented to the user.

These issues were highlighted up by the DSP during the "NERC Data Centres Workshop" held in Manchester in March 2023. This event brought together experts from the various NERC's data centres (DCs) as well as members of the DSP and set the stage for continued discussions between NERC and the DSP concerning metadata.

Ideas on how to improve metadata of NERC's data holdings

During the "NERC Data Centres Workshop", the DSP suggested possible strategies to address some of the issues outlined in the previous section. These strategies primarily focused on enhancing the use of controlled vocabularies and implementing Natural Language Processing (NLP) models to standardize legacy metadata. Below, we present three examples of the approaches considered.

Example 1 - Inconsistency due to free text entries

Harmonizing metadata values is essential for addressing challenges related to interoperability and enhancing search and discovery capabilities. The NERC Vocabulary Server (NVS) was developed by NERC to tackle this issue, providing a comprehensive collection of controlled vocabularies specific to the environmental domain. While certain Data Centres (DCs), like BODC, have made notable progress in adopting these vocabularies, others seem to fall behind. This gap can be partially explained by the absence of enforcement in applying these vocabularies during the dataset upload process by users.

Such circumstances may lead to significant manual effort by the DC staff to ensure that existing as well as newly updated datasets align with the relevant vocabularies.

The DSP has recommended two complementary approaches to harmonize metadata:

1. Resorting to domain experts to map non-conformant metadata elements to the appropriate vocabulary terms of the NVS, which may involve expanding these vocabularies. This method reflects what has traditionally been undertaken by DC staff to harmonize legacy datasets.
2. Adopting Natural Language Processing (NLP) models to align metadata with the appropriate controlled vocabularies based on the abstracts of the datasets. These NLP models will analyse the text of the metadata abstracts and extract pertinent information, such as keywords and phrases, which can then be matched with the descriptions of terms in a controlled vocabulary.

Example 2 - Problematic lineage/versions documentation of datasets.

When multiple versions of the same dataset exist without a clear and machine-readable method for identifying or highlighting the one that supersedes the others, it becomes challenging to distinguish between them. This lack of clarity affects both the findability and reusability of the data, as users need assurance that they are utilizing the most up-to-date information available.

To effectively clean up and document the lineage of a dataset, the following steps could be considered:

- Identifying duplicates of a dataset. This could be effectively achieved using clustering techniques that focus on string similarity. Methods such as Jaccard similarity and Levenshtein distance are commonly employed for this purpose. In this approach, datasets are grouped based on the similarity of their titles and abstracts, under the assumption that different versions of the same dataset will typically have nearly identical or very similar titles.
- Determine the Superseding Version. Identify the most recent version of the dataset by comparing the *CI_Date* metadata element within the XML files.
- Create References to the Superseding Version. Establish links between the superseded versions and the current version using specific metadata elements from ISO 19115. For instance, set the "*CI_Citation/status*" element to "*obsolete*" (one of the accepted values) and utilize the "*DQ_DataQuality/lineage*" element as a reference to the superseding version.

Example 3 - The datasets schemas are not documented in the corresponding metadata entry

The datasets schemas are not documented in the corresponding metadata entry (although the ISO standards allow for it). The schemas are instead detailed in external files (Word, PDF, etc) documents that lack standardisation and coherence for machine data-consumability.

This situation presents an opportunity to leverage NLP models for improvement. The process could be structured as follows:

1. Extract the relevant text from the accompanying files of the dataset. This can be accomplished programmatically using Python libraries like PyWin32 or python-docx for Word documents, and PyPDF2 or pdfminer for PDF files.
2. Employ NLP techniques to map the pertinent information from the extracted text to GEMINI UK metadata elements.
3. Employ NLP techniques to map values of the identified metadata elements to terms in the NERC Vocabularies.

Since then, some of the data centres have begun to adopt the recommended solutions to resolve inconsistencies arising from free text entries (Example 1), or at least have implemented similar approaches.

Providing feedback to NERC

NERC advised the DSP to compile a list of metadata elements deemed essential to its operations. This initiative aimed to achieve two key objectives:

1. to provide the DCs with a clearer understanding of the specific information required by the DSP to enhance the characterization, querying, and display of NERC's datasets through the Digital Solutions Hub (DSH), which the DSP is expected to develop.
2. to identify opportunities for improving and standardising metadata practices across the NERC's data centres.

What follows is the list of metadata elements / descriptors as originally compiled by the DSP team in May 2023 to be discussed with the DCs.

Metadata for the DSP

Dataset Metadata

This section contains metadata elements of interest for the DSH that pertain to the dataset as a whole. The values of these metadata elements are provided quite consistently in the XML files exposed through the cataloguing services. Potential issues here are mainly related to how the dataset version and its relation to other datasets are documented.

License - It establishes the legal framework through which different datasets can be used. This is relevant not only in relation to the end users, but also in relation to the DSH in its role of facilitator for the discovery and consumption of NERC's environmental data. What are the datasets that can be used by the DSH (e.g. datasets for public use without any license restriction)? What are the datasets that must not be used by the DSH? Is there any grey area we should be aware of (e.g. some dataset may need to be password protected and/or available only to a specific kind of users – such as academics – and not available to other users – e.g. commercial partners).

Title - Name of dataset.

Abstract- Necessary to query the datasets semantically using NLP models, as well as to provide user with a succinct description of the dataset.

Author(s) - Who generated the dataset. Needed for making sure that the responsibility on the quality of the data is correctly attributed.

Discipline - The field(s) or domain(s) the dataset pertains to. Needed for quick and coarse filtering of the dataset based on the general domain of interest for the user.

Geographical extent - Bounding box [[lat1, lon1], [lat2, lon2]]. This will help us users to filter down datasets depending on the geographical relevance to the question at hand.

Temporal extent - From date(time) To date(time). This will help the user to assess the temporal relevance to the question at hand.

Format - File format(s) of data (netCDF, Shapefile etc.). Needed to understand if the data can be used by user, they must be standardised.

Dataset version - Is the dataset superseded? Does it supersede another dataset?

Related datasets - Links to other parent dataset, datasets that are part of same collection, or datasets that are part of this dataset. This will help us understand the dataset hierarchy (standalone, part of collection, etc.) describes hierarchy of datasets.

URL - Link to location of data on Data Centre website from where it can be downloaded.

Metadata about the inner schema of a dataset

The metadata elements listed in this section describe the structure / schema of a dataset. The value of many of these metadata elements is often not provided in the XML files or is provided inconsistently across different datasets.

Fields - List of fields contained in files, e.g. observables that have been measured (level of CO, NOx, PM2.5 etc.) or variables from model data.

Units of measurement - Necessary (also) in case one wants to compare two maps where the same observable is measured, and the comparison is made based on the absolute magnitude of the corresponding values.

Feature of interest - The thing whose property is being measured or calculated in the course of an observation or a model simulation (e.g. when measuring the atmospheric concentration of NOx, the atmosphere is the feature of interest).

Temporal resolution - (where applicable) e.g. in cases the readings are part of a time series. Although there is no metadata element in the ISO19115 standard that capture specifically the time resolution of the dataset or the observable, some useful information may be contained in other elements...?

Spatial resolution - (where applicable) horizontal and/or vertical resolution for gridded datasets.

Projection - (where applicable) projection used for map-based products.

Data Acquisition Process - The process through which the data was acquired. On a very coarse level the user my want to know whether the data was measured or calculated. At more fine level one may want to include things such as sampled, averaged.

Source - Description of tool used to generate data. For example, in case of observations, this would be type and name of the instrument used to measure the variables (This would make data more meaningfully comparable across datasets). For datasets derived from model simulations, this would be name and version of model used to generate data.

Data Quality - The metadata elements in the ISO19115 standard that can be used to store information about data quality (all part of the Data Quality section) are intended to describe the accuracy, reliability, and completeness of the data. These elements have free text domains, so we may want to come up with a metric that describes with a numerical code the overall assessment of data quality (not an easy task!!)

Service metadata

The association between a dataset/layer and the corresponding service relies on internal metadata of the viewing service (Service Capabilities Document) that lacks standardisation. This makes it difficult to discover the dataset/layer of interest in the map viewing service among the other layers provided through the same service (map composition).

It is necessary to devise and adhere to agreed best practices in order to enable layer-level discoverability rather than map composition-level discoverability (which leaves to the end user the task of selecting the layer of interest among other provided layers).

Other Desirable Metadata

Directory structure - Machine-readable details of the directory structure and filenames contained within the dataset as it appears in the Data Centre, to allow access to specific data and files within the dataset.

Funding body - Funding body(ies) and grant number(s) that funded the research that generated the dataset.

Cumulative vs non-cumulative - This specifies whether an observable property is cumulative or non-cumulative. The count of pedestrians along a street is an example of cumulative property as it has to be added over time. The speed of vehicles along a street is non-cumulative, as it has to be averaged over time. Discriminating between these two kinds of properties is necessary when the user wants to manipulate the time resolution at which the data is shown. This is very likely something not included in the original metadata and needs to be added by the DSH.

Improving metadata for internal DSP usage

After the DSP submitted to NERC the document containing the list of the desirable metadata, a couple of online meetings were held between members of the data centres and the DSP. During these discussions, the document underwent a partial review, where the data centres offered valuable feedback on a number of metadata descriptors. The insights gained from these meetings were utilized by the DSP to enhance the metadata for its internal use. After downloading the complete set of metadata records from NERC's data centres, the DSP applied various enhancement techniques to refine and augment the metadata locally. The following three sections provide a detailed account of these improvement efforts.

Refactoring Datasets subject into curated list

In the ISO19115 standard, there are two distinct methods for describing the subject of a resource:

1. gmd:topicCategory: This element provides a high-level categorization of the resource's topic, helping to organize and search for resources based on general topic areas. Examples of gmd:topicCategory values can be “farming”, “health”, “economy”, etc.

```
<gmd:topicCategory>
  <gmd:MD_TopicCategoryCode>climatologyMeteorologyAtmosphere</gmd:MD_To
picCategoryCode>
</gmd:topicCategory>
```

2. MD_KeywordTypeCode=theme: This element is a crucial component of the MD_Keywords structure defined in ISO 19115. This element is designed to identify a keyword that distinctly represents the theme related to the resource. Unlike the broader topics indicated by gmd:topicCategory, these theme keywords provide more detailed insights into the resource's content. For instance, examples of such themes include "agricultural practices," "weather patterns," "economic indicators," and "atmospheric conditions."

```
<gmd:MD_Keywords>
  <gmd:keyword>
    <gco:CharacterString>atmospheric conditions</gco:CharacterString>
  </gmd:keyword>
  <gmd:type>
    <gmd:MD_KeywordTypeCode codeList="http://standards..."
codeListValue="theme">theme</gmd:MD_KeywordTypeCode>
  </gmd:type>
  <gmd:thesaurusName>
    <gmd:CI_Citation>
      <gmd:title>
        <gco:CharacterString>GEMET - INSPIRE themes, version
1.0</gco:CharacterString>
      </gmd:title>
      <gmd:date>
        <gmd:CI_Date>
          <gmd:date>
            <gco:Date>2008-06-01</gco:Date>
          </gmd:date>
          <gmd:dateType>
            <gmd:CI_DateTypeCode
codeList="http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO
_19139_Schemas/resources/codelist/gmxCodelists.xml#CI_DateTypeCode"
codeListValue="publication">publication</gmd:CI_DateTypeCode>
          </gmd:dateType>
        </gmd:CI_Date>
      </gmd:date>
    </gmd:CI_Citation>
  </gmd:thesaurusName>
</gmd:MD_Keywords>
```

At the time of this writing, there are a total of 15,704 entries (which include geographical datasets, non-geographical datasets, and series) available through the https://data-search.nerc.ac.uk/geonetwork/srv/eng/catalog.search#/home . Among these entries, 5,351 lack an associated theme, meaning they do not have a keyword formally designated to represent the resource's theme using the MD_KeywordTypeCode=theme code value. This absence hinders our ability to effectively filter the dataset by subject. To address this issue, we have implemented a strategy aimed at filling these gaps by associating one or more themes with the resources that currently lack a theme code-valued keyword.

The themes referenced in NERC's metadata entries are derived from the NVS controlled vocabulary "GEMET - INSPIRE themes, version 1.0". The only exceptions are "Agricultural and aquaculture facilities" and "Utility and governmental services," which are associated with only 14 resources.

This controlled vocabulary was downloaded and utilized as the source of terms from which to select the themes that best represent the subject matter of the 5,351 resources without a theme code-valued keyword.

Specifically, for each of these 5,351 resources, an NLP model (GPT-3.5-turbo with temperature=0) was employed to identify the theme from the controlled vocabulary that most accurately describes the subject matter of the resource based on its title and abstract.

The model was permitted to assign a maximum of three themes to the same resource if multiple themes were deemed representative, in accordance with the GEMINI UK standard.

Accessibility of datasets

The datasets maintained by NERC's data centres come with varying access restrictions and licensing agreements. The level of accessibility for each dataset significantly influences how the DSP can utilize it, and in some cases, whether it can be used at all. Initially, the DSP should focus on datasets that offer open access, as these will be fully accessible through the Digital Solution Hub without any licensing requirements for users.

The way accessibility constraints are encoded in the UK-GEMINI 2.3 standard (the profile of ISO 19139 used by NERC's Data Centres) is the following:

```
<gmd:MD_LegalConstraints>
  <gmd:accessConstraints>
    <gmd:MD_RestrictionCode
codeList="http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/resources/codelist/gmxCodelists.xml#MD_RestrictionCode
[standards.iso.org]"
codeListValue="otherRestrictions">otherRestrictions</gmd:MD_RestrictionCode
>
  </gmd:accessConstraints>
  <gmd:otherConstraints>
    <gmx:Anchor xlink:href="link to the relevant entry of the INSPIRE list
code">
      Free text description of the accessibility constraints.
    </gmx:Anchor>
  </gmd:otherConstraints>
</gmd:MD_LegalConstraints>
```

In general, the <MD_LegalConstraints> element contains information regarding restrictions and legal prerequisites for (1) *accessing* and (2) *using* the underlying resource. Accessibility restrictions and use restrictions on a resource are encoded in separate <MD_LegalConstraints> elements.

The information about the *accessibility* restrictions is organised in the child elements <accessConstraints>/<MD_RestrictionCode> and <gmd:otherConstraints>/<gmx:Anchor> as depicted in the snippet above. In particular:

- the `codeListValue` property of the child element `<accessConstraints>/<MD_RestrictionCode>` must be set to "otherRestrictions".
- the `<gmd:otherConstraints>/<gmx:Anchor>` child element specifies the kind of accessibility constraint in the following way:
  - the free text should provide a short description of the accessibility constraints.
  - the `xlink:href` property should point to the relevant kind of limitation from the INSPIRE Metadata registry.

Although NERC follow the UK-GEMINI standard, the info about the accessibility constraints is sometimes encoded in ways that differ from the standard recommendations. In particular, the <gmd:otherConstraints> / <gmx:Anchor> element is sometimes replaced by the <gmd:otherConstraints>

/ <gco:CharacterString> element, which allow for the free text but not for the xlink:href attribute. This is intended to accommodate accessibility constraints that are not included in the INSPIRE code list (or any other controlled vocabulary).

Another important aspect to consider is the fact that *no limitations* does not necessarily mean direct downloadability of the underlying resource. Many NERC resources with no accessibility limitations encoded in their metadata entries still require users to log into the data centres where the resources are stored. Although such resources do not have *legal* constraints in terms of accessibility, there are *practical* considerations such as this one that may impede a user to directly accessing them.

To identify those resources that have no legal limitations in terms of accessibility AND are not password protected, the DCs advised the DSH to look for the following criteria in the metadata XML files:

1. the `xlink:href` property of the `<gmd:otherConstraints> / <gmx:Anchor>` metadata element should be equal to http://inspire.ec.europa.eu/metadata-codelist/LimitationsOnPublicAccess/noLimitations (term corresponding to "no limitations" in the INSPIRE controlled vocabulary used to denote limitations on public access).
2. The free text of the `<gmd:otherConstraints> / <gmx:Anchor>` element denotes that there are no access limitations. This free text differs across the data centres but are consistent within each one of them. In particular, the concept of no access limitations is rendered by the following strings, depending on the datacentre:

   a) "no limitations" (CEH),
   b) "licenceOGL" (BGS)
   c) "No limitations apply" (BODC)
   d) "Public data: access to these data is available to both registered and non-registered users." (CEDA)

3. there exists at least one element `<gmd:CI_OnlineResource>` that contains (i) a child element `<gmd:linkage> / <gmd:URL>` with a non empty URL pointing to an online resource and (ii) a child element `<gmd:function> / <gmd:CI_OnLineFunctionCode>` whose `codeListValue` property is equal to "download". The snipped below show an example of such occurrence.

```
<gmd:CI_OnlineResource>
  <gmd:linkage>

<gmd:URL>http://www.sciencedirect.com/science/article/pii/01968904950005
7K</gmd:URL>
  </gmd:linkage>
  ......
  <gmd:function>
    <gmd:CI_OnLineFunctionCode
codeList="https://schemas.isotc211.org/schemas/19139/resources/codelist/
```

```
    gmxCodelists.xml#CI_OnLineFunctionCode"
    codeListValue="download">download</gmd:CI_OnLineFunctionCode>
      </gmd:function>
    </gmd:CI_OnlineResource>
```

A Jupyter Notebook was utilized to extract information regarding the accessibility constraints of NERC resources from their corresponding metadata XML files. The focus was specifically on identifying resources that do not have *legal* accessibility limitations or *practical* barriers to their downloadability, such as password protection.

The diagram below illustrates the results of the analysis conducted on the metadata records using the Jupyter Notebook.

15701 metadata records *
|
|_____ **769** (~5%) without properly encoded info about accessibility constraints
| |
| |_____ **1** with no `<MD_RestrictionCode>` element with `codeListValue="otherRestrictions"`
| |
| |_____ **768** with neither `<gco:CharacterString>` nor `<gmx:Anchor>` in `<gmd:otherConstraints>`
|
|_____ **14932** (~95%) with properly encoded info about accessibility constraints **
   |
   |_____ **5303** satisfying the three criteria for no legal nor practical accessibility limitations
   |
   |_____ **9629** not satisfying at least one of the of three criteria mentioned above.

* These corresponding to geographical and non-geographical datasets, series and models.

** These include cases where the element <gmx:Anchor> with an INSPIRE code value is not present, and instead the <gmd:otherConstraints> element is found, containing some descriptive free text.

Harmonizing file format from UK-GEMINI XML files

When searching for and exploring datasets of interest, a user may want to know the format in which the dataset is provided (and in fact using it as a filtering criterion in the search). This piece of information is particularly relevant to assess the usability of the dataset in the context in which the user operates. For example, the user by be interested in image files specifically, or in tabular data, etc.

The specifics for encoding this piece of information in UK-GEMINI is provided at this address https://www.agi.org.uk/gemini/40-gemini/1062-gemini-datasets-and-data-series/#21 . Here we only mention that the metadata element used to provide information about the data format:

- is mandatory.
- can have multiple occurrences (i.e. can be present more than once in the same metadata record to denote different data formats).
- should contain a human readable term from a controlled vocabulary as per best practice.

An initial analysis of the XML metadata files provided by NERC, resulted in the following statistics:

- only 6.8% of the metadata records (XML files) use terms from a controlled vocabulary to denote the format of the underlying data resource.
- virtually the totality (99.7%) of the metadata records that use a controlled vocabulary to denote the data format is provided by the British Oceanographic Data Centre (BODC is denoted in the XML file as the responsible party for the dataset).
- all the terms used to denote the data format are from the same vocabulary, the MEDIN data format categories

The statistics listed above confirm that BODC is ahead of the other Data Centres in the adoption of controlled vocabularies whenever their use is encouraged and compatible with the UK-GEMINI standard. BODC seems to (or should) be leading the way towards standardising the domain values of the UK-GEMINI metadata elements.

Another statistic worth mentioning is the ratio of metadata records that exclusively mention a single data format, accounting for 54% of the total count (8144 of 15000 total records).

A total of 15,000 XML metadata files associated with NERC's datasets (as of June 21, 2023) underwent parsing to extract information regarding the file format(s) of the underlying data resource. This extracted information was then used as input for a Large Language Model (GPT-3.5-turbo). The model's task was to find the best match among the terms within the MEDIN data format categories (since these are already extensively employed by BODC).

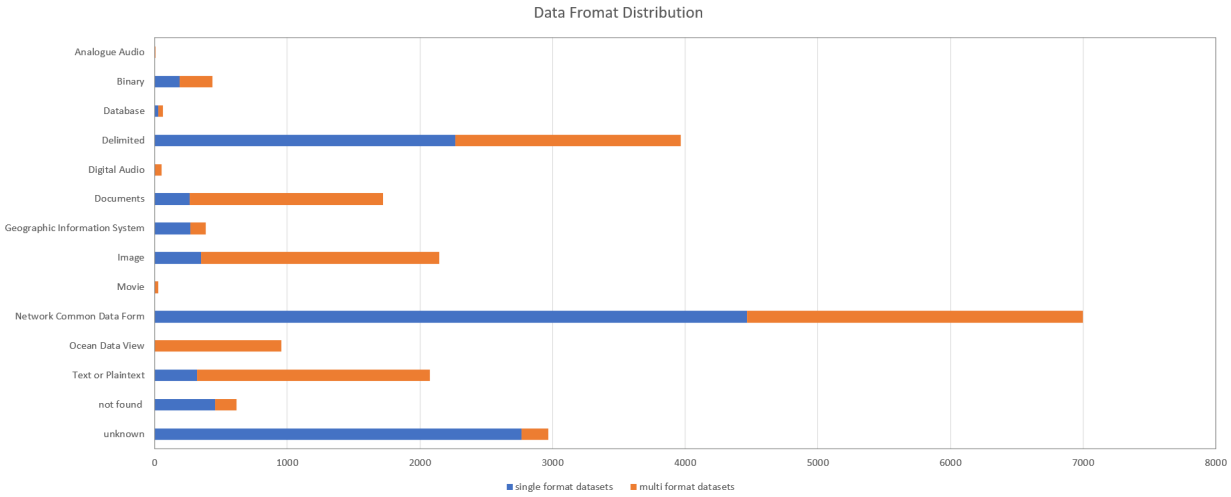
Specifically, for each individual XML file, the extracted data format description was fed into the model alongside a set of instructions that can be summarized as follows:

1. If the format description provided as input referred to only one data format, then only one term from the MEDIN vocabulary was to be selected (assuming a suitable match was found).

2. If the format description provided as input referred to more than one data format, each one of them was to be mapped to one term from the MEDIN vocabulary (assuming a suitable match was found)
3. If no match was found for a data format described in the input, then "not found" was to be returned.
4. If the format description provided was empty or "unknown", then "unknown" was to be returned.
5. The model was also explicitly instructed to map the NASA Ames file format to "Delimited".

Following the utilization of the LLM on all 15,000 inputs, the outcomes underwent refinement using ad-hoc Python code. This code was employed to verify the implementation of rules 3, 4, and 5, and to rectify any exceptions encountered, ensuring their proper enforcement.

The bar chart below shows the counts of the different terms of the MEDI vocabulary, as inferred by the LLM. The count is done separately for XML files referring to single file formats, and XML files referring to multiple file formats.



Full-sized version in Annex I.

Current state of work and discussion around NERC metadata

NERC and the data centres

During the DOG meeting held on November 20, 2024, representatives from various data centres provided updates on their progress towards achieving a higher degree of data FAIRness. The discussion focused on three main domains: the *data catalogue* (primarily addressing data-centre infrastructure), *metadata guidelines* (aimed at improving content quality and interoperability), and *metadata reviews* based on user feedback.

Feedback on the metadata from the DSP was provided to the datacentres quite timely, as it coincided with the launch of the UKRI DRI 1b project and was used to inform the new Metadata Guidance document (version 2.0), published in September 2024. Community feedback, including input from the DSP, highlighted several key areas for improvement in metadata content. These included: overcoming limitations caused by inconsistent free-text elements,

addressing issues with keywords (due to a lack of agreed-upon vocabularies), and resolving challenges in data access (e.g., data not accessible via the URLs provided in metadata).

A major focus for NERC is improving machine readability and interoperability. This is being done by semantically enhancing metadata through greater use of controlled vocabularies, revising certain GEMINI-UK specifications to make requirements more stringent, and fixing inconsistencies or missing data access links.

Additionally, NERC is working to enhance metadata quality across its data centres by identifying one or more sets of metadata descriptors that represent the most essential pieces of information to accompany all NERC data holdings. These descriptors include elements such as titles, abstracts, lineage, legal constraints, and resource constraints. NERC aims to expand the number of metadata records that incorporate these descriptors, ensuring they are properly formatted, make appropriate use of controlled vocabularies when needed, and comply fully with GEMINI-UK standards.

DSH

The discussion between the DSP and NERC regarding metadata was temporarily paused. Members of the DSP recommended that the group shift its focus to the architecture of the DSH and finalize the data schema for the geodatabase. This step is crucial to ensure the meaningful use of the data. Only after this can the DSP effectively define how the DCs can assist us with metadata.

Currently, the DSP's priority is to consolidate metadata from all data centers into a single database. Once the DSP has gathered all the necessary information, it can proceed with the following tasks:

1. Assess whether any additional information or resources are needed from each data center.
2. Identify the services that can be offered to each data center.

Meanwhile, the DSH team will concentrate on developing a Metadata Catalogue for all datasets within the database.

Annex I

Data format distribution