

# Hierarchical data (NetCDF and HDF5)

## [Hierarchical data \(NetCDF and HDF5\)](#)

[Current State:](#)

[Proposal:](#)

[Summary:](#)

[Changes:](#)

[Benefits:](#)

[Limitations](#)

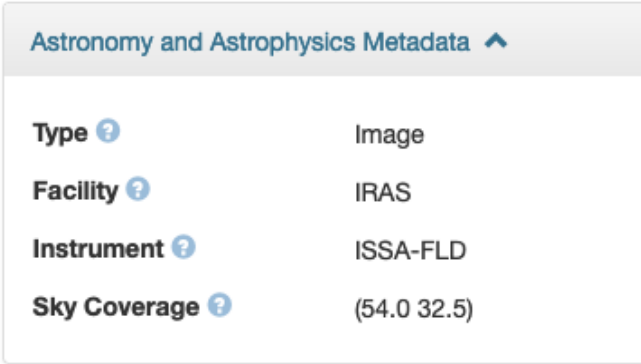
[Earlier Discussion:](#)

[Use cases:](#)

## Current State:

Harvard Dataverse, among other Dataverse installations, aims to be a general-purpose data repository, which means that it hosts a variety of research data from a range of scientific disciplines. Given that different data file formats are used in different disciplines, Dataverse needs to provide support for these data through adequate file detection and domain-specific metadata capture (see already supported schemas on [GitHub](#) or [Google Sheet](#)).

A good example of domain-specific data support in Dataverse is enabled for astronomy and astrophysics. FITS is a data format designed for astronomical data, which also stores metadata information about the origin of the data. It is the most commonly used file format in astronomy and is supported in Dataverse with automatic file detection, domain-specific metadata block and automatically populating metadata. In particular, when the “Astronomy and Astrophysics” [metadata block is enabled](#), Dataverse automatically extracts information from the FITS file and populates the block as shown in Fig. 1.



Astronomy and Astrophysics Metadata ^	
Type ?	Image
Facility ?	IRAS
Instrument ?	ISSA-FLD
Sky Coverage ?	(54.0 32.5)

Fig. 1: Astronomy and Astrophysics Metadata block.

The Network Common Data Form ([NetCDF](#)) and Hierarchical Data Format version 5 ([HDF5](#)) are specialized file formats commonly used to store data in earth and environment sciences, including climate, weather, air quality and oceanography data. Similar to the FITS file format, NetCDF and HDF5 often store metadata about the measurement, including (but not limited to) variable names, units, geographic coordinates, creators' names, emails and affiliations. However, as this metadata is embedded in the file, it is only visible when the file is downloaded and opened.

Dataverse already offers some support for data from earth and environmental sciences. In particular, it offers a geospatial metadata block (Fig. 2), which a user [can enable](#) before depositing data. However, the metadata block needs to be manually filled out by the user.

*Fig. 2: Geospatial metadata block in Dataverse*

Currently, the Dataverse software does not have specific support for NetCDF and HDF5 files. Inspired by the existing support for the FITS file format, this technical design document outlines new developments to better support these file formats.

The proposed work is funded by an NIH grant (see [short description](#)) and is primarily carried out by [Ana Trisovic](#) and [Phil Durbin](#).

## Proposal:

### Summary:

We propose the following developments to the Dataverse software to better support NetCDF/HDF5 data and geospatial data in general:

1. File detection and automatic extraction of the embedded metadata at file upload.
2. Facilitating visibility and display of extracted metadata in Dataverse.
3. Better dataset documentation through domain-specific metadata.

## Changes:

**1. File detection for NetCDF and HDF5 files and automatic extraction of the embedded metadata and file hierarchy (if available) at file upload.** When the files are detected in Dataverse, an appropriate MIME type is assigned to them. Next, we use the Unidata [netCDF-Java](#) library to extract embedded metadata from the file.

The embedded metadata can be presented as free-form key-value pairs (known as attributes) with optional conventions that are followed within disciplines that do not necessarily match the standard and adopted metadata schemas used by Dataverse. For this reason, is it not possible to blindly “map” the extracted metadata into the geospatial metadata block (or any other metadata block). As a result, the extracted metadata is stored in the dataset as an *auxiliary file* of the original file.

The extracted metadata is formatted in XML format (other formats like JSON, GeoJSON, or text can also be supported) as an auxiliary file and automatically added to the dataset at upload.

The change is documented in the following GitHub issues:

- <https://github.com/IQSS/dataverse/issues/9117>
- <https://github.com/IQSS/dataverse/pull/9152>
- <https://github.com/IQSS/dataverse/issues/9053>
- <https://github.com/IQSS/dataverse/issues/7947>
- <https://github.com/IQSS/dataverse/pull/9239>
- <https://github.com/IQSS/dataverse/issues/9153>
- <https://github.com/IQSS/dataverse/issues/9442>
- <https://github.com/IQSS/dataverse/issues/9480>

**2. Facilitating visibility and display of extracted metadata in Dataverse.** The extracted metadata is saved as an XML auxiliary file, which can be viewed directly from the Dataverse interface, as shown in Fig. 3.

In cases where metadata extraction fails, the auxiliary file won't be saved and the “eyeball” preview button won't be available, but the file can still be downloaded as normal.



The screenshot shows the Dataverse interface for a dataset. At the top, there are tabs for 'Files', 'Metadata', 'Terms', and 'Versions'. The 'Files' tab is active, showing a list of files. One file is listed: 'HadEX3-0-2\_cwd\_ann\_1901-2018.nc'. Below the file name, it says 'Network Common Data Form - 12.5 MB', 'Published Feb 6, 2023', '0 Downloads', and 'MD5: c7d...d50'. To the right of the file list, there is a preview button (an eye icon) and a download button (a download icon). The preview button is active, showing a preview of the extracted metadata in XML format.

```
<?xml version="1.0" encoding="UTF-8"?>
<netcdf xmlns="http://www.unidata.ucar.edu/namespaces/netcdf/ncm"
  <dimension name="time" length="118" isUnlimited="true" />
  <dimension name="longitude" length="192" />
  <dimension name="bnds" length="2" />
  <dimension name="latitude" length="144" />
  <variable name="time" shape="time" type="double">
  <attribute name="standard_name" value="time" />
  <attribute name="units" value="days since 1850-6-30 12:00:00" />
  <attribute name="calendar" value="proleptic_gregorian" />
```

*Fig. 3: Previewing the extracted metadata by clicking on the preview “eyeball” next to the NetCDF file. The previewer shows the metadata from the XML auxiliary file.*

The change is documented in the following GitHub issues:

- <https://github.com/gdcc/dataverse-previewers/pull/18>
- <https://github.com/gdcc/dataverse-previewers/issues/17>

---

Optional developments include:

**3. External tool support with h5web and Binder.** We'd like to visualize data with h5web, and allow data exploration with Binder. We'd like to - **ADD MORE INFO AT THE END OF THE WEEK**

**4. Better dataset documentation through domain-specific metadata.** Dataverse already supports a number of geospatial metadata fields as an independent block (see Fig. 2). As part of meeting this goal, we consider two changes: first, automatically populating the geospatial metadata block ([#9331](#)), inspired by the similar support for the FITS data format), and second, adding new standard metadata fields to improve the documentation of the dataset ([#6713](#), [#7455](#), [Release 5.13](#) & [PR #8239](#)).

The metadata fields such as *geographicBoundingBox* could be automatically populated from the extracted metadata below. To enable this, we consider using GDAL software or the netCDF-Java library mentioned above.

```
NC_GLOBAL#geospatial_lat_max=90
NC_GLOBAL#geospatial_lat_min=-90
NC_GLOBAL#geospatial_lat_resolution=1.25
NC_GLOBAL#geospatial_lat_units=degrees
NC_GLOBAL#geospatial_lon_max=360
NC_GLOBAL#geospatial_lon_min=0
NC_GLOBAL#geospatial_lon_resolution=1.875
NC_GLOBAL#geospatial_lon_units=degrees
NC_GLOBAL#Metadata_Conventions=Unidata Dataset Discovery v1.0,CF
Discrete Sampling Geometries Conventions
```

Additionally, we consider [adding the following geospatial metadata fields](#):

**Proposed new fields and changes to Dataverse's Geospatial Metadata Block:**  
(corresponding to OpenGeoMetadata & ISO 19139/19115 & FGDC/CSDGM)

\*Indicates fields that exist in Dataverse currently, changes are proposed

Metadata field	Subfields	OGM	ISO 19115/19139	FGDC/CSDGM
Spatial Reference System Information	Coordinate system and projection name (CV)  <a href="#">EPSG Registry</a> Identifier code (CV)	N/A	MD_ReferenceSystem RS_Identifier  Ex. label: "NAD 83" Ex. identifier '4269' (EPSG code)	Map_Projection Grid_Coordinate_System
Geographic Extent (optional within; repeatable set)	*Geographic Coverage  *Geographic Unit  Geographic Coordinates /Centre/Centroid point  *Geographic Bounding Box  Geographic Coordinate Units (measure used for the latitude and longitude values. Ex. "Decimal degrees")	<a href="#">Spatial Coverage   OpenGeoMetadata</a>  <a href="#">Centroid   OpenGeoMetadata</a> (similar to geo_point since DV v5.13.)  <a href="#">Bounding Box   OpenGeoMetadata</a>  N/A	MD_Keywords (type: 'place')  Geographic coordinate unit > encoding Ex: <gmd:westBoundLongitude> <gco:Decimal>-76</gco:Decimal> </gmd:westBoundLongitude>	Need 4.1.1.3 <a href="#">Geographic Coordinate Units</a>
Spatial Representation Information	Geometric Object Type (Ex. Point, Line, Polygon)  Spatial Representation Type (Ex. Raster, Vector)	Relates to: <a href="#">Geometry   OpenGeoMetadata</a>	Relates to: <a href="#">Geometric Object Type</a> (Ex. Point, Composite, Complex)  Spatial Representation Type (Ex. 'Grid', 'Vector')	Need

Source Information (repeatable)	Georeferenced Source	<a href="#">Georeferenced   OpenGeoMetadata</a> <a href="#">Source   OpenGeoMetadata</a>	Source Lineage	Need
Spatial Resolution (repeatable)	Value (ex. '8') Units (ex. 'cm') Scale (ex. 1:250,000)	N/A	MD_Resolution, spatialResolution	Review: Longitude_Resolution Latitude_Resolution
Web map service (repeatable)	URI (Ex. ArcGIS rest service path)  Type (Ex. WFS, WMS, etc.)	<a href="#">WxS Identifier   OpenGeoMetadata</a>	Linkage Online Resource	Review: Geospatial_Data_Presentation_Form

The change is documented in the following GitHub issues:

- <https://github.com/IQSS/dataverse/issues/9331>
- <https://github.com/IQSS/dataverse/issues/9069>
- <https://github.com/IQSS/dataverse/issues/6713>
- <https://github.com/IQSS/dataverse/issues/7091>
- <https://github.com/IQSS/dataverse/issues/7455>
- <https://github.com/cf-convention/cf-conventions/issues/435>

## 5. Automatically enabling a metadata block when FITS/NetCDF/HDF5 files are detected.

(#9410) While domain-specific metadata is valuable for better documentation of research data and adherence to [the FAIR principles](#), from the examples below, we see that it is rarely used in practice.

1. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/9ZSHYB>
2. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/L5WNU4>
3. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/E0HLON>
4. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/WH9OQR>
5. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BKWJAI>

This is because the domain-specific metadata blocks need to be manually enabled (as described [here](#)). As part of this change, we are considering automatically enabling the

domain-specific metadata block (if permissions allow) when FITS/NetCDF/HDF5 files are detected, which would be automatically populated with the embedded metadata, if it can be extracted. In the case of not enough permission, the user could be prompted to reach out to an admin who can help.

The change is documented in the following GitHub issues:

- <https://github.com/IQSS/dataverse/issues/9410>

## Benefits:

General benefits include:

- Meeting the research community's needs by supporting free and open source file formats (NetCDF and HDF5), which are popular in many scientific fields.
- Improved transparency, findability and reuse of the data by displaying extracted metadata, thus avoiding the need for a user to retrieve and open the file locally (with specialized software) just to understand what it stores.
- The new feature of automatically creating auxiliary files can be used for future developments and integrations in Dataverse (i.e., provenance, runtime environment).

Automatically populating metadata and new metadata fields would benefit:

- Automatically extracting and populating metadata would alleviate the burden of the data depositor and minimize human error in file documentation.
- The *geographicBoundingBox* metadata field, which can be automatically populated, will be used by Dataverse geospatial search, which looks for datasets based on a geographic point and radius (see [documentation](#), [technical design doc](#), and [the GitHub thread](#)).

## Limitations

Based on our ongoing discussions, we recognize the following limitations:

- Automatically extracting metadata will not be possible in all cases, as not all files will have metadata.
- We experimented with extracting or generating a figure from NetCDF/HDF5 files and setting it as a thumbnail, however, it might be inverted (flipped), misleading or in other ways, unusable.
- We considered generating new metadata from the variables available in the file, but it is common practice to use numerical “flags” in the geospatial data to capture information such as “recording device malfunction”, which would make generated information such as mean and standard deviation inaccurate. We plan instead to promote the use of specific tools more tailored to analyzing the files in a containerized environment, such as

Binder, launched from Dataverse. We may provide some documentation and guidance in this area.

- Integration with a THREDDS or OPeNDAP server is currently out of scope.

## Earlier Discussion:

As part of this effort, we organized additional weekly open-to-all Dataverse community meetings on Mondays at 10 AM ET. All meeting notes are [publicly available](#). Particularly interesting ideas are recorded in [the Ideas Google Document](#).

The project was introduced in the Dataverse Community News [post](#).

## Use cases:

This work could be useful in the following use cases:

- Geospatial data in Dataverse and the integration with GeoBlackLight (see [presentations](#) by Marc McGee, Maura Carbone, Kristial Allen, Jamie Jamison and Paul Dante)

Use cases evident by using a variety of NetCDF/HDF5 data from these examples:

1. Surface PM2.5: <https://sites.wustl.edu/acag/datasets/surface-pm2-5/#V5.GL.03>
2. GridMET data: <https://www.northwestknowledge.net/metdata/data/>
3. Global Workshop on Earth Observation  
<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OYBLGK>
4. NetCDF data from Harvard Dataverse:  
[https://dataverse.harvard.edu/dataverse/harvard?q=\\*.nc](https://dataverse.harvard.edu/dataverse/harvard?q=*.nc)