

Directory Interchange Format (DIF) Usability Survey

NASA's Earth Science Data Systems Standards Process Group (SPG) is considering the DIF (Directory Interchange Format) specification, developed by the GCMD (Global Change Master Directory), for adoption as a community standard. Your responses to this survey on the usability of the DIF and the suitability of this specification for Earth science data will be helpful.

Please answer as many of the questions below as you can..

1. Please provide your name, organization and contact information (including email address).

2. Are you answering for your entire organization, for a smaller group, or individually?

Entire organization

3. Are you a data producer, data consumer, or both?

Both

4. How long have you been using the DIF?

The AADC started with DIF 7 in 1999 as a separate standalone database within the AADC, which was periodically copied and sent to the GCMD for hosting there on the Antarctic Master Directory. Currently we no longer run a publicly accessible standalone database, but instead a master copy of all our metadata records are directly hosted on the GCMD site.

5. Please describe how the DIF is used in your organization.

We use DIF to catalogue our scientific datasets. Our metadata catalogue is hosted on the GCMD website and has an IDN Node of AMD/AU. We also contribute to the Antarctic Master Directory (also hosted on the GCMD website) as part of our obligations to the international Antarctic community, and specifically the Committee for Antarctic Data Management.

We currently manage 2000+ records at the GCMD. Each datapoint/observation/collection/database held within the AADC also references a DIF

record via a connection with the DIF Entry_ID. A simple script allows end-users to click on a link and go directly to the GCMD page for that metadata record.

We have our own copy of the MD9.4 database schema, from an earlier instance when we hosted our own DIF records. We take the DIF XML and update our database at the same time that we update or create DIF records at the GCMD. We can then use parts of a DIF, such as data quality or access constraints, in links to the download files we show on our website. Spatial coverages are also stored at our site and these are used to map DIF record extents on some of our local mapping tools. It also gives us a chance to correct spatial coverages by mapping many at one time.

The URLs stored in the 'related_url' table can be tested for accuracy/existence at any point in the life of a DIF record, and we can also use this table to check if revisions across many metadata records need to be changed. This is in addition to the feature of the GCMD DocBuilder tool which checks the validity of URLs at the point of creation or update of a DIF record.

We have a secondary table that links the MD people records to our corporate database. From this we can report on outcomes of individuals which comprise ownership of metadata records, investigators on project or authors on publications. These metrics are used extensively to monitor and encourage science productivity. Individuals with poor ratings or overdue metadata records may not obtain approval for their science projects.

For the following, you can answer either about the DIF alone, or relative to other, comparable specifications.

6. What are the strengths of the DIF? How has the use of the DIF helped your organization?

DIF has helped us effectively organise and catalogue our datasets, such that it is now easier for us to locate and access data. DIF has an advantage over other metadata formats (in my opinion) in that it is a little more accessible to the average user. It doesn't have the sheer, daunting bulk of ISO19115 or FGDC, but is also more detailed than Dublin Core.

DIF also benefits from a good tool created by the GCMD. The DIF search engine is quite powerful, logical and relatively easy to use. The DIF entry and edit tool, DocBuilder is perhaps also the best metadata entry and edit tool currently available. I have personally reviewed a large number of metadata entry and edit tools (eg Geonetwork, M3Cat, Medi), and DocBuilder is by far the most straightforward and simple. Having said that, it is still a complex tool that requires a degree of skill and understanding to use, but of all the tools I have seen it is the one that presents the average user with the greatest chance of "successfully" completing a metadata record. Having said that, the average user still

finds metadata to be a complex and time-consuming beast. The DocBuilder tool also benefits enormously from the practice of ensuring that when a user creates/edits a metadata record it is not immediately uploaded to the database, but is sent to a science coordinator for checking first (who can then actually complete/polish the record).

7. What are the weaknesses of the DIF? What would you like to change about the DIF or what would make the DIF a better specification?

Currently the DIF format, or rather the GCMD tool, does not support the ISO19115 format, and is not capable of producing metadata compliant with the 19139 specification. Once the DIF tool is capable of doing this then its use and power will increase enormously. Some aspects of DIF are also not compatible with ISO equivalents. For example, the ISO version of “dataset release place” also requires a time to be entered, whereas DIF has no such equivalent. DIF’s “data resolution ranges”, are also practically impossible to convert into ISO 19115 format using an XSLT. Other aspects of the DIF (eg “DIF Distribution Media”) would also benefit from using a controlled list rather than allowing free-text. This would enable greater (and easier) compatibility with ISO 19115.

There are common elements in all metadata schemas such as people, address’s etc. OGC is working on harmonising common elements in their many schemas, and it may be beneficial to identify common elements in DIF that can be readily digested by other communities.

8. How well does the DIF solve your metadata storage, discovery, and/or interchange needs? Are there specific areas it is applicable to vs. areas where it is not applicable or not used?

DIF does an excellent job of solving our metadata storage and discovery needs. As mentioned above in point 7 though, it is not yet fully capable of satisfying our interchange needs.

There will always be domain-specific metadata schemas such as EML. The AADC hosts the widest variety of scientific data and DIF suits this well.

9. How suitable is the DIF for representing your data holdings?

Very suitable. We have been active in contributing new science keywords, especially in the area of sea-ice.

10. Do you use the DIF to track your own data holdings (i.e. do you use DIF in your own data management activities)?

Yes all the time. We have added a database table to hold the following information (one record per DIF) about our control and reporting on DIF records.

Name	Purpose
Metadata has data	A yes /no flag to indicate that the metadata record has links to actual data via either a URL to a file download, or a URL to an external data location. It is not possible to view all the URL's in a record and to ascertain that it really points to some form of data. We use this to pick out metadata records that we do not yet have data for and we put effort into those to get the data for archiving and/or insertion into a database.
Data held by AAD	If Yes to the first flag, then this flag is either Yes or NO if the AAD holds the data. Some of the AADC metadata records describe data held in external organisations.
Data is a download file	Set to Yes if the data linked to a record is simply a zipped file of the original data. Other data often gets inserted into a database.
Metadata Status	Has four values C = complete – record is complete and adequate N = New - record is initially new and has not been vetted I = Incomplete – record requires updating and might have missing descriptions on data quality, etc. Used by AADC to chase investigators. O = Overdue – data custodian has not updated the record in a timely manner.

11. What are the limitations of the DIF? Does the DIF prevent you from doing things you would like to do? Does its use make other things more difficult?

As well as the items laid out in Point 7, we feel that it would be beneficial if the GCMD tool had some web based administration reporting mechanisms built into it. Currently at the AADC we have to build our own tools if we want to produce reports (eg show me all the broken URLs in every metadata record, or show me all records which are missing the field “Access Constraints”). This would increase the versatility of the tool for science coordinators like myself, who are not based at the GCMD.

12. Do you think ESDS-RFC-012 (and thus the DIF) should be endorsed as a NASA Earth Science Data Systems Standard? Why or why not?

Yes it should become a standard. In the next few years, there will be tools to help transform DIF to other formats like ISO, EML etc. it is important that the community of use has a stable controlled standard.