Comparing element usage in NASA vs IDN metadata collections in the CMR

Overview

The CMR includes metadata that originate in three dialects: DIF, ECHO, and ISO. The largest portion of CMR collection records are in the SciOps collection, and originate in DIF format. These are referred to here as the SciOps group. The second largest portion of metadata for NASA collections come from the NASA DAACs and originate in ECHO format. These are referred to here as the NASA group. A third group of metadata in the CMR originate from agencies around the world that participate in the International Directory Network. These records originate in DIF and are referred to as the IDN group.

The DIF and ECHO dialects were originally developed to facilitate discovery of collections in the Global Change Master Directory or ECHO. The content of these “discovery” dialects is translated into the ISO dialects that are the eventual target for CMR. This translation is generally done without augmentation, so the content does not change very much.

Metadata providers have a choice about which metadata dialect(s) they use to submit metadata to the CMR. We compared the NASA and IDN collections to understand how this choice affects the metadata content.

Data Selection

This analysis compares item usage (elements and attributes) in the 18 NASA collection with the 8 non NASA (IDN) collections. This evaluation identifies items that exist in collections as well as items that are complete in collections. In order for an item to exist in a collection it must be present in at least 1 metadata record in the collection. In order for an item to be complete in a collection it must be present in all metadata records in the collection.

The table below shows the collections included in this evaluation and the record count for each collection.

Table 1. Collections and record counts

Collection	Organization	Count		Collection	Organization	Count		Collection	Organization	Count		Collection	Organization	Count
ASF	NASA	161		LARC	NASA	406		ORNL_DAAC	NASA	1216		ISRO	IDN	19
CDDIS	NASA	38		LARC_ASDC	NASA	606		PODAAC	NASA	603		JAXA	IDN	340
GES_DISC	NASA	1044		LPDAAC_ECS	NASA	285		SEDAC	NASA	202		LM_FIRMS	IDN	1
GHRC	NASA	361		NSIDC_ECS	NASA	223		USGS_EROS	NASA	11		NOAA_NCEI	IDN	5448
LAADS	NASA	130		NSIDC_V0	NASA	784		AU_AADC	IDN	2559		USGS_LTA	IDN	130
LANCEAMSR2	NASA	6		OB_DAAC	NASA	132		ESA	IDN	103				
LANCEMODIS	NASA	154		OMNIRT	NASA	5		EUMETSAT	IDN	23				

Not Provided Values

The CMR team are in the difficult position of trying to make a coherent and useful metadata repository by collecting metadata from many organizations and projects that have different goals and needs. This presents a challenge as the CMR evolves and new requirements emerge. Metadata managers need to account for content that is not provided by the metadata providers. At the current time, this is between three and five percent of the content.

The solution for both NASA provided collections and IDN provided collections is to add the string “Not provided” to expected fields that have no content. This clearly indicates that content is missing, except that tools that read the metadata or translate it must be aware of and consider this convention to get meaningful results. The tools we use for evaluating metadata completeness are agnostic to element and attribute values. Therefore the analysis presented below, which compares NASA Complete with IDN Complete does not include metadata fields with a value of 'Not Provided'. Below are the metadata fields with 'Not Provided' values that were excluded from the NASA Complete vs IDN Complete evaluation

NASA Collections - Fields with ‘Not Provided' values

Table 2 shows fields in the NASA Group with the 'Not provided' flag and the % of records from each data provider that include that value. In four cases these missing data flags make up over 50% of the content for fields. Elements with these missing values were not considered further in the analysis.

Table 2. Occurrences of missing data ('Not provided') in NASA collections

Number of Records		4	1044	50	154	305	2	11	19	783	16	
Paths - Data Provider	Count	CDDIS	GES_DISC	LAADS	LANCEMODIS	LARC	LARC_ASDC	LPDAAC_ECS	NSIDC_ECS	NSIDCV0	SEDAC	Average
/gmi:acquisitionInformation/gmi:instrument/gmi:type	10	100%	100%	78%	49%	83%	100%	91%	100%	75%	100%	86%
/gmd:contentInfo/gmd:processingLevelCode/gmd:code	8	0%	0%	90%	95%	2%	0%	91%	100%	100%	100%	64%
/gmd:identificationInfo/gmd:processingLevel/gmd:code	8	0%	0%	90%	95%	2%	0%	91%	100%	100%	100%	64%
/gmd:identificationInfo/gmd:abstract	6	0%	90%	90%	92%	100%	100%	91%	0%	0%	0%	63%
/gmi:acquisitionInformation/gmi:platform/gmi:description	6	0%	0%	66%	0%	30%	100%	9%	0%	100%	0%	34%
/gmd:contentInfo/gmd:dimension/gmd:otherProperty/gco:Record/eos:AdditionalAttributes/eos:AdditionalAttribute /eos:reference/eos:description	5	0%	87%	68%	1%	0%	0%	91%	0%	100%	0%	39%
/gmd:identificationInfo/gmd:descriptiveKeywords/gmd:keyword	4	0%	0%	66%	0%	30%	100%	0%	0%	53%	0%	28%
/gmd:identificationInfo/gmd:status/gmd:MD_ProgressCode	4	0%	0%	0%	1%	17%	0%	9%	0%	22%	0%	5%
/gmi:acquisitionInformation/gmi:platform/gmi:identifier/gmd:code	4	0%	0%	66%	0%	30%	100%	0%	0%	53%	0%	28%
/gmd:identificationInfo/gmd:aggregationInfo/gmd:aggregateDataSetName/gmd:citedResponsibleParty/gmd: contactInfo/gmd:onlineResource/gmd:linkage/gmd:URL	4	0%	100%	0%	100%	0%	0%	0%	0%	100%	100%	44%
/gmi:acquisitionInformation/gmi:instrument/eos:sensor/eos:type	3	0%	0%	0%	1%	4%	0%	0%	0%	0%	0%	0%
/gmd:identificationInfo/gmd:pointOfContact/gmd:individualName	2	0%	0%	4%	0%	0%	0%	9%	0%	0%	0%	1%
/gmd:identificationInfo/gmd:resourceConstraints/gmd:useLimitation	2	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	22%
/gmd:contact/gmd:organisationName	1	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	11%
/gmd:identificationInfo/gmd:pointOfContact/gmd:organisationName	1	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	11%
/gmd:distributionInfo/gmd:distributor/gmd:distributorTransferOptions/gmd:onLine/gmd:linkage/gmd:URL	1	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	11%

IDN Collections - Fields with ‘Not provided' values

Table 3 shows fields in the IDN Group with the 'Not provided' % of IDN records from each data provider that include that value. In seven cases these missing data flags make up over 50% of the content for fields and in two cases (processingLevelCodes) the values in all records are 'Not provided'. Elements with these missing values were not considered further in the analysis.

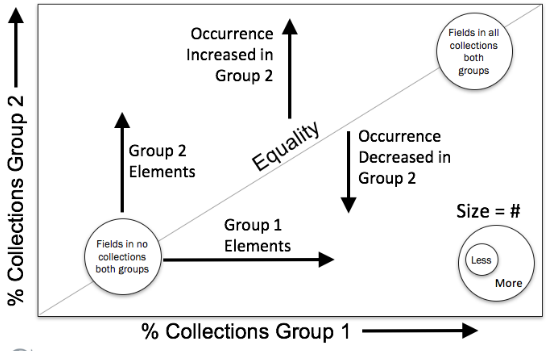
Table 3. Occurrences of missing data ('Not provided') in IDN collections

Number of Records		2559	1	58	130	5488	23	340	103	
Paths - Provider	Count	AU_AADC	LM_FIRMS	EUMETSAT	USGS_LTA	NOAA_NCEI	ISRO	JAXA	ESA	Average
/gmd:contentInfo/gmd:dimension/gmd:otherProperty/gco:Record/ eos:AdditionalAttributes/eos:AdditionalAttribute/eos:reference/eos:description	8	100%	100%	45%	100%	92%	91%	100%	100%	91%
/gmd:contentInfo/gmd:processingLevelCode/gmd:code	8	100%	100%	100%	100%	100%	100%	100%	100%	100%
/gmd:identificationInfo/gmd:processingLevel/gmd:code	8	100%	100%	100%	100%	100%	100%	100%	100%	100%
/gmi:acquisitionInformation/gmi:instrument/gmi:type	8	33%	100%	78%	82%	69%	100%	96%	97%	82%
/gmd:identificationInfo/gmd:citation/gmd:edition	7	100%	0%	40%	98%	86%	87%	28%	95%	67%
/gmd:identificationInfo/gmd:citation/gmd:identifier/gmd:version	7	100%	0%	40%	98%	86%	87%	28%	95%	67%
/gmi:acquisitionInformation/gmi:platform/gmi:description	7	100%	0%	100%	100%	100%	91%	100%	100%	86%
/gmd:identificationInfo/gmd:descriptiveKeywords/gmd:keyword	6	45%	0%	60%	76%	31%	0%	6%	30%	31%
/gmd:identificationInfo/gmd:status/@codeListValue	6	1%	0%	76%	13%	45%	0%	99%	43%	35%
/gmd:identificationInfo/gmd:status/gmd:MD_ProgressCode	6	1%	0%	76%	13%	45%	0%	99%	43%	35%
/gmi:acquisitionInformation/gmi:platform/gmi:identifier/gmd:code	6	45%	0%	60%	76%	31%	0%	6%	30%	31%
/gmd:identificationInfo/gmd:abstract	1	0%	100%	0%	0%	0%	0%	0%	0%	13

Presentation

Item usage for NASA collections and IDN collections is shown in Figure 1. The bubble chart interpretation graphic provides a schematic for interpreting the bubble plots.

Figure 1: *Bubble Chart Interpretation*

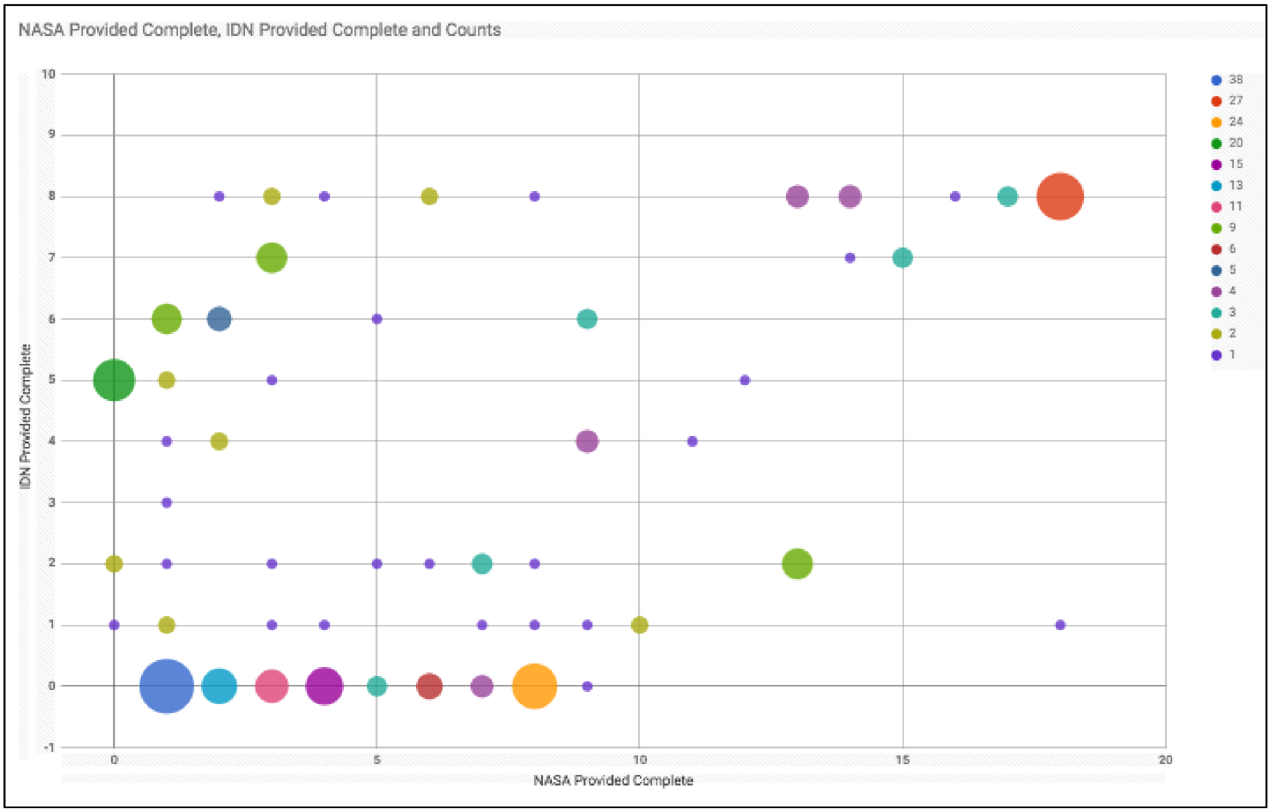

% Collections Group 2
Occurrence Increased in Group 2
Fields in all collections both groups
Group 2 Elements
Equality
Occurrence Decreased in Group 2
Size = #
Group 1 Elements
Fields in no collections both groups
Less
More
% Collections Group 1

NASA Complete vs IDN Complete

Figure 2 compares item completeness in NASA and IDN collections. The X axis shows the number of NASA collections that are complete with respect to an items. The Y axis shows number of IDN collections that are complete with respect to an items. The bubble size shows the number of items included in the 2 collections.

The large red bubble in the upper right corner of the plot shows represents items (elements and attributes) that are complete is all 18 NASA collections and in all 8 IDN collections. This bubble includes 27 items, as shown in the legend. The large blue bubble in the lower left corner of the plot represents the items that are complete in 1 NASA collection and in 0 IDN collections. This bubble includes 38 items, as shown in the legend. Click on the chart to view the data.

Figure 2: NASA Provided Complete vs IDN Provided Complete



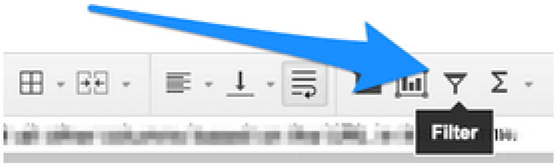
Accessing the Data

Click on Figure 2 (above) to view the data in Google Sheets. The Google Sheets display is interactive. It enables data identification for each bubble, and includes a look up table for identifying the ISO items associated with the bubble. To access and use the interactive version:

1. Click on the Figure 2 above to view the data in Google Sheets
2. Hover over the bubbles with your mouse to identify the number of complete NASA collections and the number of complete IDN collections associated with the bubble. The hover identification also shows the number of element (Counts) associated with each bubble.
3. To the left of the chart is a lookup table for identifying the xPaths associated with each of the bubbles. To identify the xPaths associated with a bubble, match the NASA Provided Complete and IDN Provided Complete values from the bubble with the values in the lookup table.

Elements complete in all NASA and IDN Collections

Twenty-seven (27) items (elements and attributes) are complete in all NASA and all IDN Collections: To identify these items, click here to access the spreadsheet and select the All NASA and All IDN filtered view from the toolbar.



Elements complete in all NASA Collections and some IDN Collections

Only one (1) element is complete in all NASA collections and some IDN collections: gmd:identificationInfo/gmd:citation/gmd:edition. This reflects the fact that this element is used in the NASA collections to provide version information for resources. To identify these items, click here to access the spreadsheet and select the All NASA and Some IDN filtered view from the toolbar.

Elements complete in all IDN Collections and some NASA Collections

Nineteen (19) items (elements and attributes) are complete in all IDN collections and complete in a smaller number of NASA collections. These are the bubbles along the top of the chart. To identify this items, click here to access the spreadsheet and select the All IDN and Some NASA filtered view from the toolbar.

Elements complete in no IDN Collections and some NASA Collections

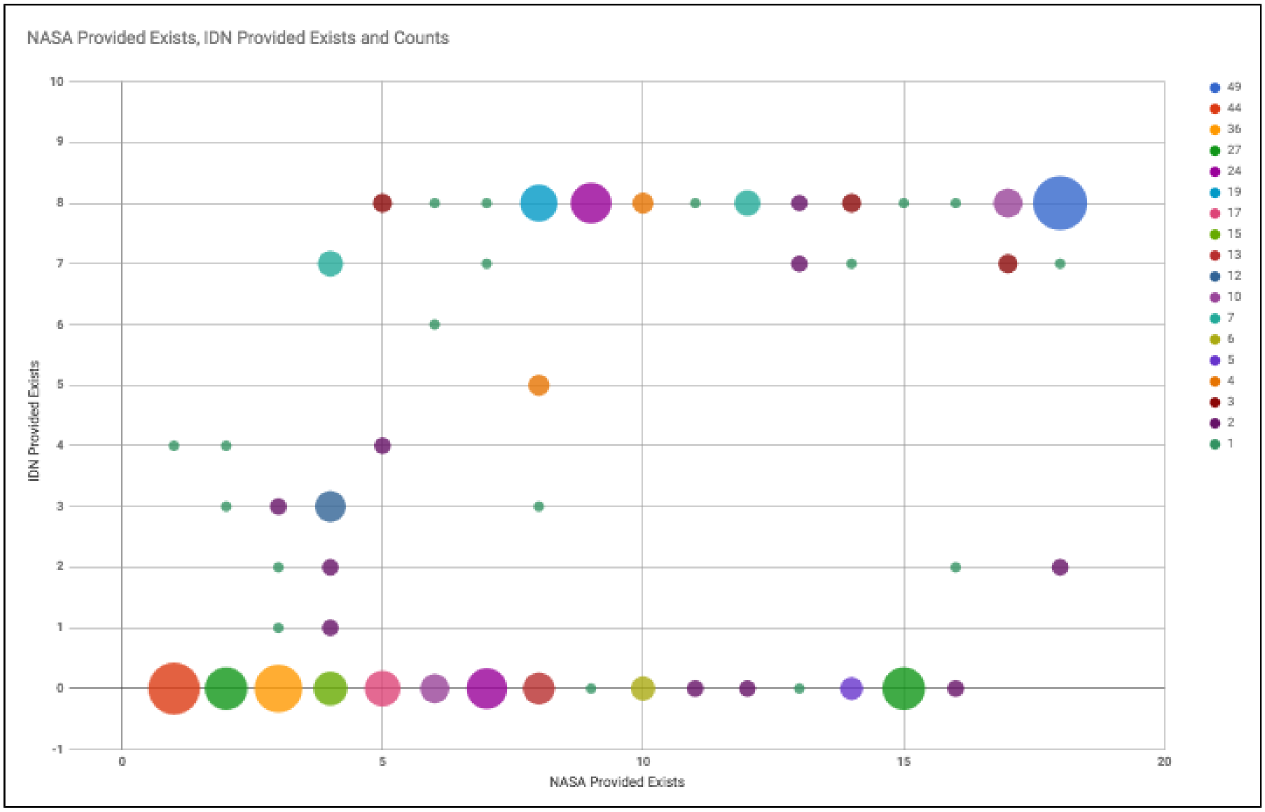
Sixty-three (63) elements are complete in no IDN Collections and some NASA Collections. These are the bubbles on the X axis in the chart. None are complete in more than nine NASA collections. Many of these elements are related to keyword thesaurus and additional attribute information. It is interesting to note that keyword thesaurus information is completely absent from the IDN collections, reflecting the fact that GCMD keywords are used in those collections, perhaps as a requirement of participation in the IDN. To identify this items, click here to access the spreadsheet and select the Some NASA and No IDN filtered view from the toolbar.

NASA Exists vs IDN Exists

Figure 3 compares item existence (elements and attributes) in NASA Collections with IDN collections. The X axis shows the number of NASA collections that include an item. The Y axis shows number of IDN collections that include an item. The bubble size shows the number of items included in the 2 collections.

The large blue bubble in the upper right corner of the plot shows the items (elements and attributes) that exist is all 18 NASA collections and in all 8 IDN collections. This bubble includes 49 items, as shown in the legend. The large red bubble in the lower left corner of the plot shows items that exist is one NASA collection and in zero IDN collections. This bubble includes 44 items, as shown in the legend. Click on the chart to view the data.

Figure 3: NASA Provided Exists vs IDN Provided Exists



Accessing the Data

Click on Figure 3 (above) to view the data in Google Sheets. The Google Sheets display is interactive. It enables data identification for each bubble, and includes a look up table for identifying the ISO items associated with the bubble. To access and use the interactive version:

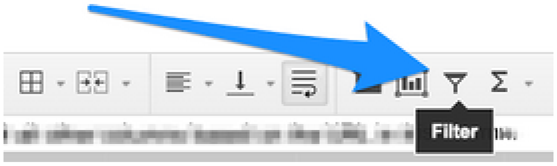
1. Click on the Figure 3 graphic above to view the data in Google Sheets
2. Hover over the bubbles with your mouse to identify the concept existence in NASA collections and concept existence in IDN collections associated with the bubble. The hover identification also shows the number of items (Counts) associated with each bubble.
3. To the left of the chart is a lookup table for identifying the xPaths associated with each of the bubbles. To identify the xPaths associated with a bubble, match the NASA Exists and IDN Exists values from the bubble hover with the values in the look up table.

Elements that exist in all IDN Collections or no IDN Collections

Figure 1 indicates that bubbles congregate near the equality line for similar collections. Bubbles congregate near the axis if collections differ significantly. Note that this comparison clearly suggests significant differences between the NASA and IDN collections. One hundred and twenty-six of four hundred and seven (126/407 = 31%) of the provided elements exist in all IDN collections (along top axis) and two hundred and thirty-two (232/407 = 57%) exist in none of the IDN collections (along the bottom axis). In total, 88% of the items are either complete in all or absent from all of the IDN collections.

Elements that exist in all NASA and IDN Collections

Forty-nine (49) items (elements and attributes) exist in all NASA and all IDN Collections (blue bubble in the upper right corner of the chart): To identify these items, click here to access the spreadsheet and select the All NASA and All IDN filtered view from the toolbar.



Elements that exist in some NASA Collections and all IDN Collections

Seventy-seven (77) items exist in all IDN Collections and some NASA collections. These are the bubbles along the top of the chart. To identify these items, click here to access the spreadsheet and select the All IDN and Some NASA filtered view from the toolbar.

Elements that exist in some NASA Collections and no IDN Collections

One hundred and thirty-five (135) items (elements and attributes) exist in some NASA collections and no IDN collections. These are the bubbles along the bottom of the chart. To identify these items, click here to access the spreadsheet and select the Some NASA and No IDN filtered view from the toolbar.

Other Filters

Elements or Attributes

To view just the elements, click here select the Elements or the Attributes filtered view from the toolbar.