DataCite Recommendation Analysis for the National Center for Atmospheric Research

Table of Contents

Executive Summary

This report presents the results of a completeness evaluation of the National Center for Atmospheric Research (NCAR) metadata with respect to the DataCite recommendation. DataCite is an organization formed to help improve consistent identification of data and other resources with the goal of making data more accessible and useful. DataCite provides a metadata recommendation that includes mandatory, recommended, and optional elements. This recommendation places a high priority on making data, people and organizations discoverable through the use of unique identifiers.

NCAR has many ways of sharing the data they produce and archive. The Data Stewardship Engineering Team (DSET) is responsible for helping NCAR labs share their data efficiently and in a unified manner. Currently there are 9 labs, each with varying types of information collections. There are also a variety of metadata practices employed; some groups use XML standards from external sources, some use XML standards formed at NCAR, and some have their own structured documentation in the form of a database or ASCII headers. This report focuses on the ISO, MODS, DataCite, netCDF, and RDA-CISL holdings at NCAR. We acquired samples of each of these collections in order to explore completeness with respect to the DataCite recommendation and as an introduction to the metadata evaluation and improvement tools we are developing. Applying the DataCite recommendation to a data center's metadata, regardless of the dialect they utilize, can help prepare organizations that are hoping to improve the identification of their information collections through the use of Digital Object Identifiers (DOIs).

One important observation is that only the DataCite dialect contains all of the concepts required by the DataCite recommendations. Given the focus of DataCite, several of the missing concepts are related to identifiers for resources and people/organizations. RDA-CISL is missing three mandatory concepts, two recommended concepts, and two optional concepts. ISO is missing three mandatory concepts, and two recommended concepts. MODS is missing two mandatory concepts, two recommended concepts, and two optional concepts. netCDF is missing three mandatory concepts, four recommended concepts, and four optional concepts. EOL is missing four mandatory concepts, four recommended concepts, and one optional concept. If it is essential to meet the DataCite recommendation, choices must be made regarding expanding or changing the current dialects in use at NCAR. It is also interesting to note that all of the DataCite shared metadata records from NCAR labs do not use any of the concepts in the DataCite recommendation that dialects like ISO do not contain.

The metadata samples included 2505 records from 7 NCAR labs. Of those, no records included all of the metadata concepts in the DataCite recommendation that are represented in the dialect the record is written in. The first two signature groups of the ISO dialect are the most complete records at NCAR labs with respect to the DataCite Recommendation. CGD did not provide assets with sharable metadata. Mapping was done with CGD to determine whether the CESM experiments database contained the concepts in the DataCite Recommendation and with EOL to provide comparisons of unstructured (unshared) metadata to the concepts in the DataCite Recommendation. ACOM did not have metadata, or machine readable structured documentation. As such, research was done to find shared vocabularies and ontologies in chemistry to help them

document their data. Decisions have to made about how they want to store and create metadata before the assets can be assessed.

NCAR Labs and the Dialects Analyzed

	CGD	DataCite	EOL	ISO	MODS	netCDF	RDA-CISL
Atmospheric Chemistry Observations and Modeling (ACOM)							
Climate & Global Dynamics (CGD)	X						
Computational and Informational Systems Lab (CISL)							X
Earth Observing Lab (EOL)			X	X			
High Altitude Observatory (HAO)		X					
Integrated Information Services (IIS)					X		
Mesoscale and Microscale Meteorology (MMM)		X					
Research Applications Lab (RAL)						X	
Unidata (UCP)		X		X			

DataCite – What is it?

DataCite is an organization founded to help make data more accessible and usable. DataCite is the originating organization of the DOI, and manages the distribution of these identifiers by their member organizations. Their purpose is to develop and support methods to locate, identify and cite data and other research objects. Specifically, they develop and support the standards behind persistent identifiers for data.

In the context of the terminology we use (see Glossary), DataCite is an organization that created a set of recommendations at three levels, mandatory, recommended, and optional (described in the DataCite Metadata Schema) and an XML schema (a dialect) for implementing those recommendations. Concepts included in all three levels are listed with definitions and XPaths in several dialects on the DataCite Recommendation Page. The dialect is currently being used in the DataCite search portal and in creating DOI landing pages. The recommendations are useful for communities looking for expert guidance about metadata concepts that are intended to enhance data discovery. Applying the DataCite recommendation to a data center's metadata, regardless of the dialect they utilize, can help prepare organizations that are currently hoping to improve the identification of their dataset through DOIs.

The DataCite Metadata Schema is a list of metadata elements defined by DataCite for the accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions. The schema is intended to help DOI users document resources that have been assigned DOIs. The resource that is being identified with a DOI can be of any type, but it is typically a dataset (used in its broadest sense). It may include not only numerical data, but also any other research outputs.

This assessment of a sampling of 9 collections from 7 NCAR labs is based on the DataCite 3.1 recommendation. The NCAR collections are in 5 XML dialects, ISO, MODS, DataCite, netCDF, RDA-CISL. The mapping of the EOL Zith database as well as CGD's CESM experiments database are also added to the comparison of dialect maxima with the DataCite recommendation to highlight the opportunities labs at NCAR have to develop structured documentation, or metadata into shared metadata.

Recommendation Dialect Comparison – How Does My Dialect Fit?

Recommendations are created in order to address metadata requirements perceived by the organizations that create them, e.g. data discovery, use, understanding. It is important to understand the fit, and the misfit, between the recommendation and the dialects.

This section provides information about similarities and differences between the DataCite recommendations and the NCAR dialect implementations. We describe the recommendation-dialect fit in the following ways:

- A recommendation comparison report
- A chart comparing the concepts in the DataCite recommendation and the NCAR dialects.
- A graph comparing the maximum number of concepts in each dialect compared to the recommendation.
- A Recommendation/Dialect comparison that lists all concepts in the DataCite recommendation and NCAR dialects.
- Tables that describe the concepts in the DataCite recommendation that are missing in the NCAR dialects.

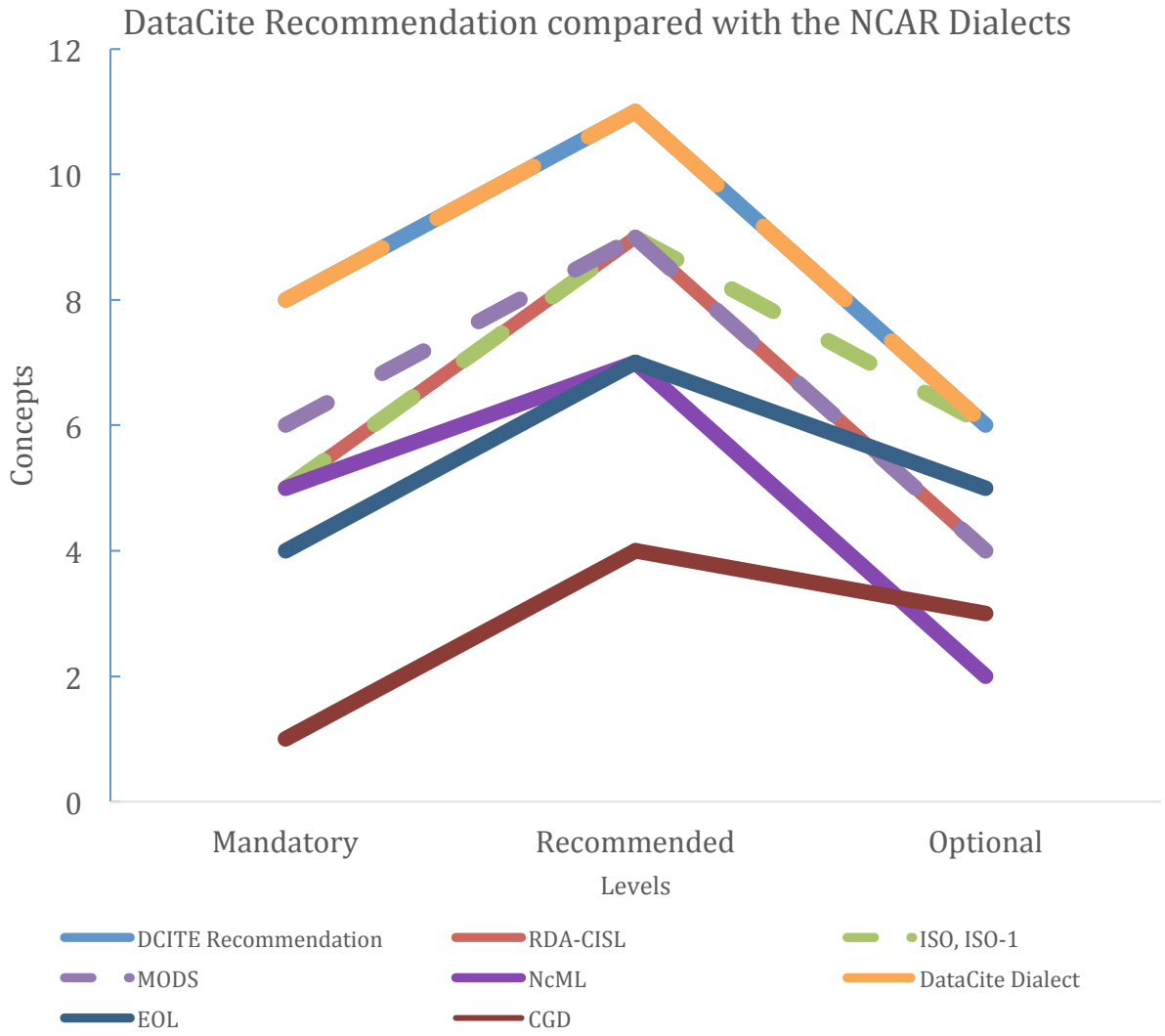
Recommendation Level Comparison Report

The purpose of the Recommendation Level Comparison report is to show the concepts that are included in each of the recommendation levels (see Glossary) being compared. A concept is a generalized term for a documentation entity, and a recommendation level is a list of concepts that an organization identifies as necessary for achieving a documentation goal. A concept may appear in multiple levels because the concept is general and may have multiple specific elements that appear in different recommendation levels. For example, there may be multiple resource identifiers that identify different resources. The recommendation levels included in this study are: mandatory, recommended, optional (see table below).

Concept	Score	Description	Mandatory	Recommended	Optional
Abstract	1	A paragraph describing the resource.		X	
Author / Originator	1	The principal author of the resource	X		
Author / Originator Identifier	1	A unique identifier for a resource author or originator	X		
Author / Originator Identifier Type	1	The type of unique identifier for a resource author or originator	X		
Contributor Name	1	Contributor to the resource		X	
Contributor Role	1	The role of any individuals or institutions that contributed to the creation of the data.		X	
Keyword Vocabulary	1	If you are following a guideline or using a shared vocabulary for the words/phrases in your 'keywords' attribute, put the name of that guideline here.		X	
Publisher	1	Publisher of the cited resource	X		
Related Resource Identifier	1	Identifier for a resource related to the resource being described.		X	
Resource Creation/Revision Date	2	The date the resource was created or revised	X	X	
Resource Format	1	The physical or digital manifestation of the resource			X
Resource Identifier	2	Identifier for the resource described by the metadata	X		X
Resource Identifier Type	1	The type of identifier used to uniquely identify the resource.	X		
Resource Language	1	The language of the resource.			X
Resource Title	1	A short description of the resource. The title should be descriptive enough so that when a user is presented with a list of titles the general content of the data set can be determined.	X		
Resource Type	1	A resource code identifying the type of resource; e.g. dataset, a collection, an application (See MD_ScopeCode) for which the metadata describes.		X	
Resource Version	1	Version of the cited resource			X
Responsible Party Identifier	1	A unique identifier for a person or an organization		X	
Responsible Party Identifier Type	1	The type of a unique identifier for a person or an organization		X	
Rights	1	Information about rights held in and over the resource			X
Spatial Extent	1	The spatial extent of the resource.		X	
Theme Keyword	1	A word or phrase that describes some aspect of a resource. Can be one of several types.		X	
Transfer Size	1	The size of the digital resource			X

Recommendation/Dialect Maximum Graph

This graph compares the number of concepts included in the DataCite recommendations (recommendation maximum) to the maximum number of these concepts supported by the NCAR dialects (dialect maximum). The three levels of the DataCite recommendation (mandatory, recommended and optional) include 8, 11, and 6 concepts respectively as indicated by the upper line in the Figure below. This Recommendation Maximum defines the highest completeness scores with respect to these recommendations for any metadata dialect. The difference between the Recommendation Maximum (8 11 6) and the Dialect Maximum e.g. RDA-CISL (5 9 4) indicates that there are three mandatory DataCite concepts that are missing from the RDA-CISL dialect, as well as two recommended concepts and two optional concepts. The ISO dialect is missing three mandatory concepts, and two recommended concepts. MODS is missing two mandatory concepts, two recommended concepts, and two optional concepts. The netCDF dialect is missing three mandatory concepts, four recommended concepts, and four optional concepts. The EOL dialect is missing four mandatory concepts, four recommended concepts, and one optional concept. The preliminary mapping of CGD is missing seven mandatory concepts, seven recommended and three recommended concepts.



Recommendation/Dialect Comparison Report

These tables identify all of the concepts included in the DataCite recommendation, and verify their existence in the NCAR dialects with an "X".

Mandatory Level

Concept	Score	Description	DCITE	RDA-CISL	ISO	MODS	netCDF	CGD	EOL
Author / Originator	6	The principal author of the resource	X	X	X	X	X		X
Author / Originator Identifier	1	A unique identifier for a resource author or originator	X						
Author / Originator Identifier Type	1	The type of unique identifier for a resource author or originator	X						
Publisher	5	Publisher of the cited resource	X	X	X	X	X		
Resource Creation/Revision Date	6	The date the resource was created	X	X	X	X	X		X
Resource Identifier	6	Identifier for the resource described by the metadata	X	X	X	X	X		X
Resource Identifier Type	2	The type of identifier used to uniquely identify the resource.	X			X			
Resource Title	7	A short description of the resource. The title should be descriptive enough so that when a user is presented with a list of titles the general content of the data set can be determined.	X	X	X	X	X	X	X

Recommended Level

Concept	Score	Description	DCITE	RDA-CISL	ISO	MODS	netCDF	CGD	EOL
Abstract	7	A paragraph describing the resource.	X	X	X	X	X	X	X
Contributor Name	7	Contributor to the resource	X	X	X	X	X	X	X
Contributor Role	6	The role of any individuals or institutions that contributed to the creation of the data.	X	X	X	X	X		X
Keyword Vocabulary	5	If you are following a guideline or using a shared vocabulary for the words/phrases in your 'keywords' attribute, put the name of that guideline here.	X	X	X	X	X		
Related Resource Identifier	6	Identifier for a resource related to the resource being described.	X	X	X	X		X	X
Resource Creation/Revision Date	6	The date the resource was created	X	X	X	X	X		X
Resource Type	5	A resource code identifying	X	X	X	X		X	

		the type of resource; e.g. dataset, a collection, an application (See MD_ScopeCode) for which the metadata describes.							
Responsible Party Identifier	1	A unique identifier for a person or an organization	X						
Responsible Party Identifier Type	1	The type of a unique identifier for a person or an organization	X						
Spatial Extent	6	The spatial extent of the resource.	X	X	X	X	X		X
Theme Keyword	7	A word or phrase that describes some aspect of a resource. Can be one of several types.	X	X	X	X	X	X	X

Optional Level

Concept	Score	Description	DCITE	RDA-CISL	ISO	MODS	netCDF	CGD	EOL
Resource Format	6	The physical or digital manifestation of the resource	X	X	X	X		X	X
Resource Identifier	6	Identifier for the resource described by the metadata	X	X	X	X	X		X
Resource Language	4	The language of the resource.	X		X	X			X
Resource Version	3	Version of the cited resource	X		X	X			
Rights	5	Information about rights held in and over the resource	X	X	X		X		X
Transfer Size	4	The size of the digital resource	X	X	X				X

DataCite Concepts missing from NCAR Dialects

The Tables below provide lists of the DataCite recommendation concepts that are missing from the NCAR dialects for each of the three levels. The DataCite dialect contains all concepts. If a level is missing from the dialect's section it is complete; there are no missing concepts.

RDA-CISL

Missing Mandatory Concepts

Concept	Description
Author / Originator Identifier	A unique identifier for a resource author or originator
Author / Originator Identifier Type	The type of unique identifier for a resource author or originator
Resource Identifier Type	The type of identifier used to uniquely identify the resource.

Missing Recommended Concepts

Concept	Description
Responsible Party Identifier	A unique identifier for a person or an organization
Responsible Party Identifier Type	The type of a unique identifier for a person or an organization

Missing Optional Concepts

Concept	Description
Resource Language	The language of the resource.
Resource Version	Version of the cited resource

ISO

Missing Mandatory Concepts

Concept	Description
Author / Originator Identifier	A unique identifier for a resource author or originator
Author / Originator Identifier Type	The type of unique identifier for a resource author or originator
Resource Identifier Type	The type of identifier used to uniquely identify the resource.

Missing Recommended Concepts

Concept	Description
Responsible Party Identifier	A unique identifier for a person or an organization
Responsible Party Identifier Type	The type of a unique identifier for a person or an organization

MODS

Missing Mandatory Concepts

Concept	Description
Author / Originator Identifier	A unique identifier for a resource author or originator
Author / Originator Identifier Type	The type of unique identifier for a resource author or originator

Missing Recommended Concepts

Concept	Description
Responsible Party Identifier	A unique identifier for a person or an organization
Responsible Party Identifier Type	The type of a unique identifier for a person or an organization

Missing Optional Concepts

Concept	Description
Rights	Information about rights held in and over the resource
Transfer Size	The size of the digital resource

netCDF

Missing Mandatory Concepts

Concept	Description
Author / Originator Identifier	A unique identifier for a resource author or originator
Author / Originator Identifier Type	The type of unique identifier for a resource author or originator
Resource Identifier Type	The type of identifier used to uniquely identify the resource.

Missing Recommended Concepts

Concept	Description
Responsible Party Identifier	A unique identifier for a person or an organization
Responsible Party Identifier Type	The type of a unique identifier for a person or an organization
Related Resource Identifier	Identifier for a resource related to the resource being described.
Resource Type	A resource code identifying the type of resource; e.g. dataset, a collection, an application for which the metadata describes.

Missing Optional Concepts

Concept	Description
Resource Language	The language of the resource.
Resource Version	Version of the cited resource
Transfer Size	The size of the digital resource
Resource Format	Format of the resource

EOL

Missing Mandatory Concepts

Concept	Description
Author / Originator Identifier	A unique identifier for a resource author or originator
Author / Originator Identifier Type	The type of unique identifier for a resource author or originator
Resource Identifier Type	The type of identifier used to uniquely identify the resource.
Publisher	Publisher of the cited resource

Missing Recommended Concepts

Concept	Description
Responsible Party Identifier	A unique identifier for a person or an organization
Responsible Party Identifier Type	The type of a unique identifier for a person or an organization
Keyword Vocabulary	If you are following a guideline or using a shared vocabulary for the words/phrases in your 'keywords' attribute, put the name of that guideline here.
Resource Type	A resource code identifying the type of resource; e.g. dataset, a collection, an application (See MD_ScopeCode) for which the metadata describes.

Missing Optional Concepts

Concept	Description
Resource Version	Version of the cited resource

CGD

Missing Mandatory Concepts

Concept	Description
Author / Originator Identifier	A unique identifier for a resource author or originator
Author / Originator Identifier Type	The type of unique identifier for a resource author or originator
Resource Identifier Type	The type of identifier used to uniquely identify the resource.

Missing Recommended Concepts

Concept	Description
Responsible Party Identifier	A unique identifier for a person or an organization
Responsible Party Identifier Type	The type of a unique identifier for a person or an organization
Related Resource Identifier	Identifier for a resource related to the resource being described.
Keyword Vocabulary	If you are following a guideline or using a shared vocabulary for the words/phrases in your 'keywords' attribute, put the name of that guideline here.
Contributor Role	The role of any individuals or institutions that contributed to the creation of the data.

Missing Optional Concepts

Concept	Description
Resource Language	The language of the resource.
Rights	Information about rights held in and over the resource
Transfer Size	The size of the digital resource
Resource Language	The language of the resource.
Resource Version	Version of the cited resource

Metadata Analysis – How Complete are My Metadata?

This section presents the results of an analysis of the completeness of a collection of metadata records in a dialect or a set of dialects with respect to the recommendation(s) being reported on. A collection is a group of metadata records, commonly organized by data center, organization or project and often stored in a database or web accessible folder. Collections are composed of metadata records of the same dialect.

Sample metadata were obtained from RAL, Unidata, CISL, EOL, HAO, MMM, and IIS after meetings with labs to determine the state of the metadata for the assets the lab wanted to have analyzed. CGD and ACOM had a database and ASCII headers respectively. These two assets are considered incomplete. They are not ready to support the DataCite Recommendation. The samples from the other assets are highly variant in size, from 4 to 1300 records. This section presents the results of an analysis of the completeness of these metadata collections with respect to the DataCite recommendation. Completeness is measured by determining how many concepts from each DataCite recommendation are contained in the metadata records.

These results are presented as counts of records with identical completeness scores with respect to the recommendation(s). The completeness scores are given in terms of the **number of elements that are missing** from a record, so **low scores are good**. When a recommendation includes multiple levels (e.g. Mandatory, Recommended, and Optional), the scores are given as a series of numbers, one for each level. These are termed signatures (see Glossary). Typically, many records are missing the same concepts and, therefore, have identical signatures. The signature 2 3 1 indicates a metadata record that has been tested for three levels and is missing 2 mandatory, 3 recommended, and 1 optional concepts. This record is less complete than a record with a signature of 1 1 1 and more complete than a record with a signature of 3 4 3.

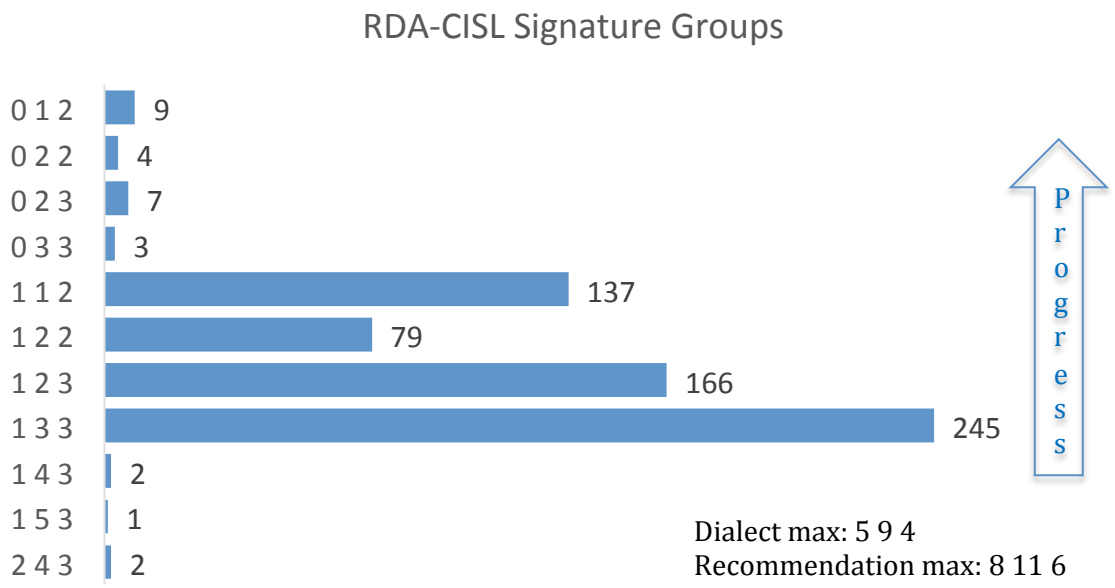
Completeness Results for NCAR Metadata Dialects

The differences between the completeness scores in the following chart reflect concepts that are present in the more complete records and missing from the less complete ones. By organizing records based on their scores across all three levels we create a set of signatures for the collection, which allows us to identify groups of records that typically contain the same concepts.

We report completeness for the three DataCite recommendation levels. The order of the levels is mandatory, recommended, and optional. A score of "0 0 0" indicates that the record is as complete as possible with respect to the DataCite recommendations. There are no complete records in any of the collections analyzed in this report.

Additionally, the concepts missing in each signature group are given. Concepts that do not appear in any records but are contained in the given dialect are listed and described as "unused concepts". All "incomplete concept" signatures also include the unused concepts.

RDA-CISL



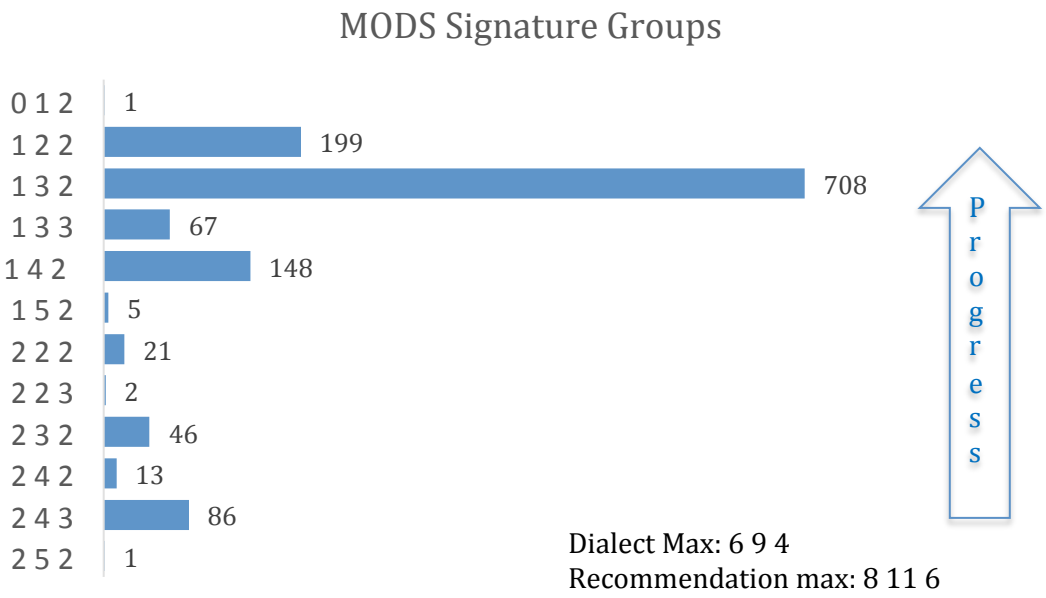
Unused Concepts

Concept	Description
Resource Format	Format of the resource
Contributor Role	The role of any individuals or institutions that contributed to the creation of the data.
Rights	Information about rights held in and over the resource

Incomplete Concepts by Signature

Signature	Concepts
0 1 2	The unused concepts described above.
0 2 2	Spatial Extent
0 2 3	Related Resource Identifier, Transfer Size
0 3 3	Related Resource Identifier, Spatial Extent, Transfer Size
1 1 2	Author / Originator
1 2 2	Author / Originator, Spatial Extent
1 2 3	Author / Originator, Related Resource Identifier, Transfer Size
1 3 3	Author / Originator, Related Resource Identifier, Spatial Extent, Transfer Size
1 4 3	Author / Originator, Contributor Name, Related Resource Identifier, Spatial Extent, Transfer Size
1 5 3	Author / Originator, Contributor Name, Resource Type, Related Resource Identifier, Spatial Extent, Transfer Size
2 4 3	Author / Originator, Resource Creation/Revision Date(2), Related Resource Identifier, Spatial Extent, Transfer Size

MODS



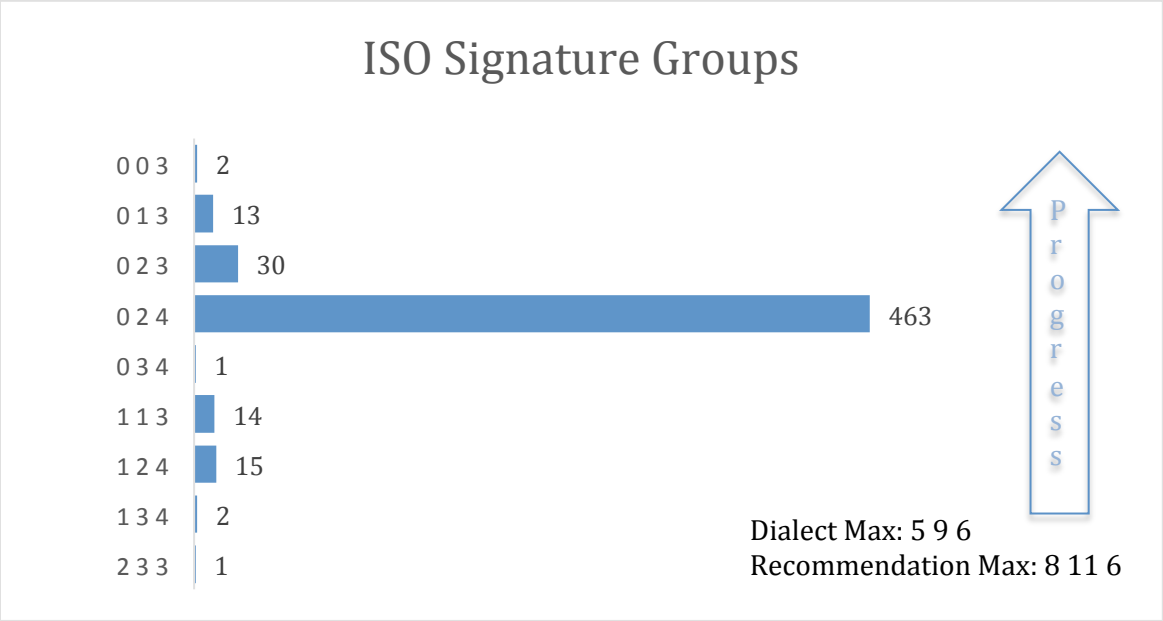
Unused Concepts

Concept	Description
Spatial Extent	The spatial extent of the resource.
Resource Format	Format of the resource
Resource Version	Version of the cited resource

Incomplete Concepts By Signature

Score	Concepts
0 1 2	The unused concepts described above.
1 2 2	Resource Creation/Revision Date(2)
1 3 2	Resource Creation/Revision Date(2), Keyword Vocabulary, Related Resource Identifier
1 3 3	Publisher, Contributor Name, Contributor Role, Resource Language
1 4 2	Resource Creation/Revision Date(2), Keyword Vocabulary, Related Resource Identifier, Abstract
1 5 2	Resource Creation/Revision Date(2), Keyword Vocabulary, Related Resource Identifier, Abstract
2 2 2	Publisher, Resource Creation/Revision Date(2)
2 2 3	Publisher, Resource Creation/Revision Date(2), Resource Language
2 3 2	Publisher, Resource Creation/Revision Date(2), Keyword Vocabulary, Related Resource Identifier
2 4 2	Publisher, Resource Creation/Revision Date(2), Keyword Vocabulary, Contributor Name, Contributor Role, Related Resource Identifier, Abstract
2 4 3	Publisher, Resource Creation/Revision Date(2), Contributor Name, Contributor Role, Resource Language
2 5 2	Publisher, Resource Creation/Revision Date(2), Keyword Vocabulary, Related Resource Identifier, Abstract, Resource Language

ISO



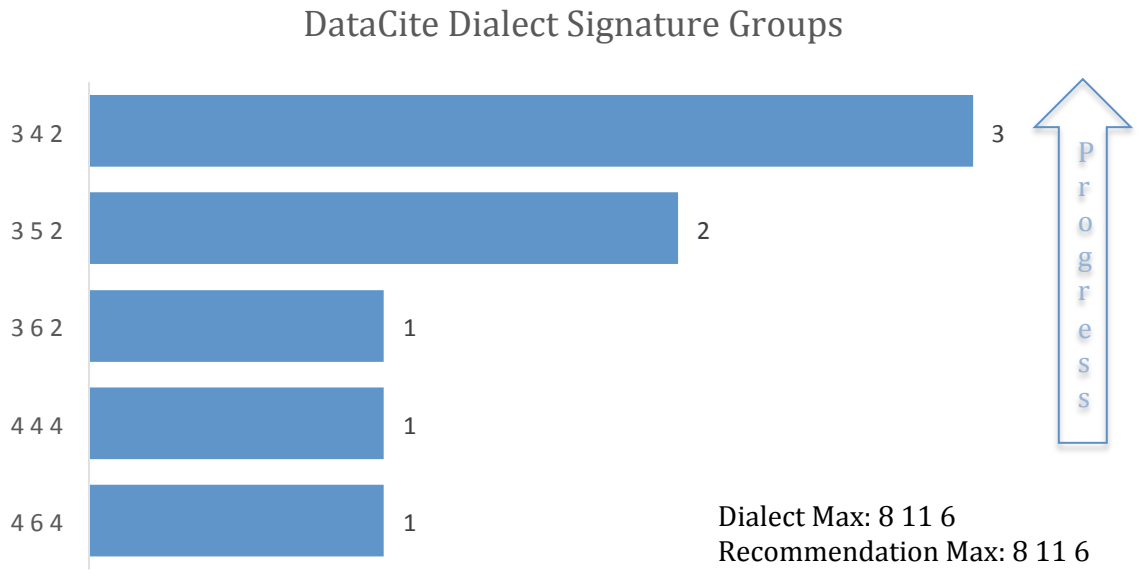
Unused Concepts

Concept	Description
Resource Format	Format of the resource
Transfer Size	The size of the digital resource

Incomplete Concepts by Signature

Score	Concepts
0 0 3	Resource Version
0 1 3	Resource Version, Related Resource Identifier
0 2 3	Keyword Vocabulary, Related Resource Identifier, Rights
0 2 4	Keyword Vocabulary, Related Resource Identifier, Resource Version, Rights
0 3 4	Keyword Vocabulary, Related Resource Identifier, Rights
1 1 3	Resource Creation / Revision Date, Resource Version
1 2 4	Author / Originator, Keyword Vocabulary, Related Resource Identifier, Resource Version, Rights
1 3 4	Author / Originator, Keyword Vocabulary, Related Resource Identifier, Resource Version, Rights, Abstract
2 3 3	Author / Originator, Publisher, Related Resource Identifier, Resource Version

DataCite



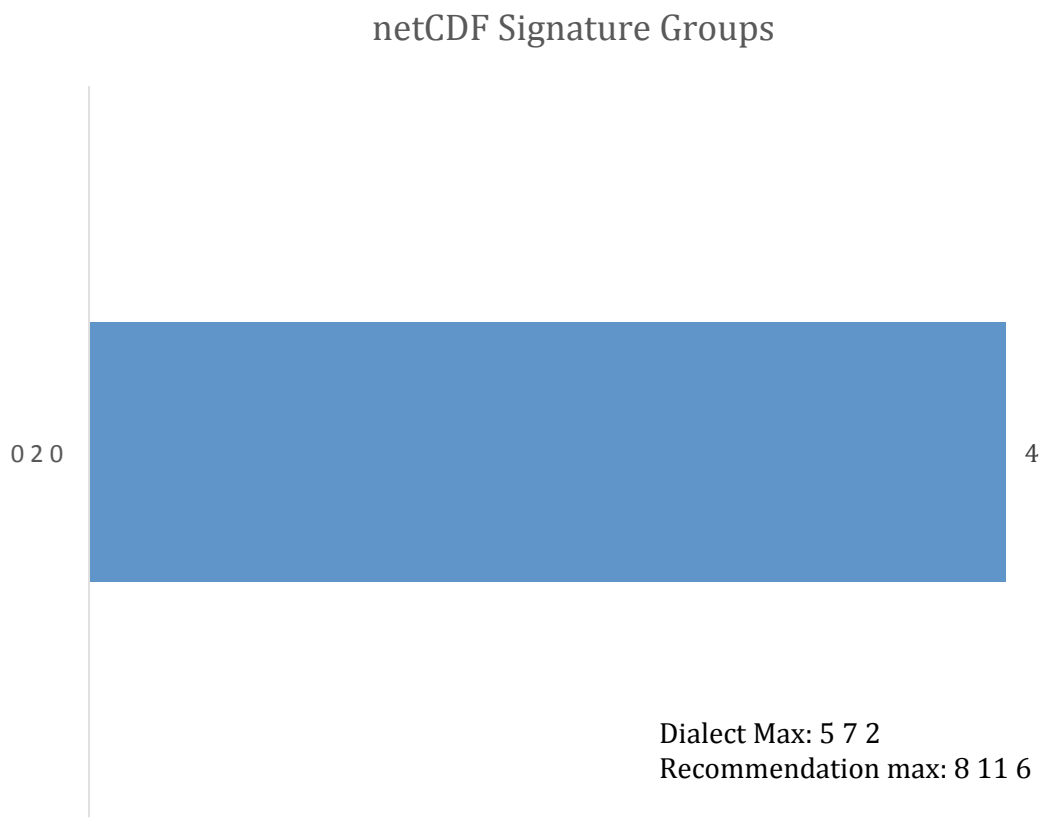
Unused Concepts

Concept	Description
Resource Identifier Type	The type of identifier used to uniquely identify the resource
Author / Originator Identifier	A unique identifier for a resource author or originator
Author / Originator Identifier Type	The type of unique identifier for a resource author or originator
Contributor Role	The role of any individuals or institutions that contributed to the creation of the data.
Responsible Party Identifier Type	The type of a unique identifier for a person or an organization
Responsible Party Identifier	A unique identifier for a person or an organization

Incomplete Concepts by Signature

Score	Concepts
3 4 2	The unused concepts described above plus Spatial Extent, Transfer Size, Resource Version
3 5 2	Related Resource Identifier, Spatial Extent, Resource Version, Rights
3 6 2	Keyword Vocabulary, Related Resource Identifier, Spatial Extent, Transfer Size, Rights
4 4 4	Resource Creation/Revision Date(2), Resource Language, Transfer Size, Resource Format, Resource Version
4 6 4	Resource Creation/Revision Date(2), Keyword Vocabulary, Spatial Extent, Resource Language, Transfer Size, Resource Format, Resource Version

netCDF



Unused Concepts

Concept	Description
Keyword Vocabulary	If you are following a guideline or using a shared vocabulary for the words/phrases in your "keywords" attribute, put the name of that guideline here.
Contributor Role	Format of the resource

Specific Guidance – How to Improve the Shared Metadata

The analysis above identifies specific concepts that are missing from NCAR metadata records, but are included in their respective dialects. This section provides specific guidance on how to write metadata for those concepts in a variety of dialects. A positive and straightforward first step is to assess what some NCAR records already include and implement them collection wide. The information is presented in three ways

- A table to describe the dialects usage at NCAR in relation to the recommendation.
- An incomplete concepts chart for each dialect
- Guidance links for incomplete and unused concepts

The table below is comprised of rows for each DataCite recommendation concept and columns for each dialect. Cells are filled with a color or a percentage. The percentage is how many records in the sample set contain that concept. Green represents 100%. Yellow represents 0%, a concept that the dialect contains but is not in any record in the sample set for that dialect. Red represents a concept missing from the dialect. The table is intended to show not only how complete a dialect is for the DataCite recommendation, as well as how complete the records are with respect to the dialect maxima. An important use of the table is also determining if the currently used dialect is the best for the purposes of labs at NCAR.

We provide charts for each dialect showing how many records are missing a concept and what level the concept belongs to. These charts are intended to help identify the most important and most attainable goals to maximize results for each iteration towards improvement. If the concept is missing from 95% of records but only from the Optional level of the DataCite recommendation it should be lower priority than a Mandatory concept missing in only 50% of records. The Mandatory level is red, the Recommended level is green, and the optional level is blue.

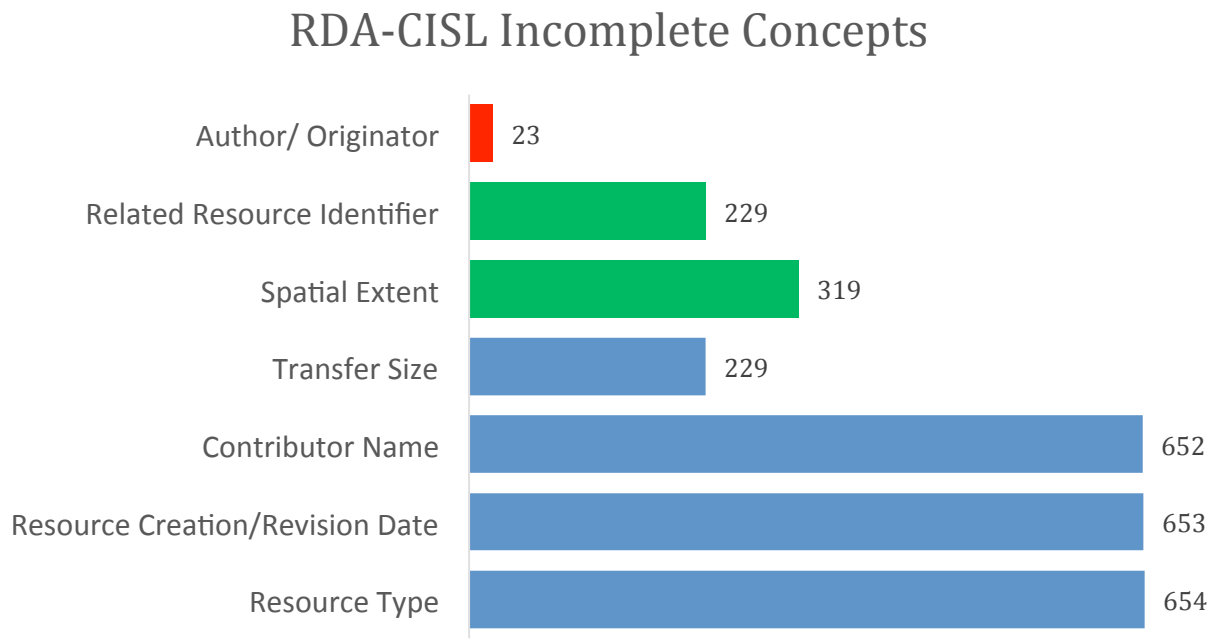
The guidance links resolve to pages on the Earth Science Information Partners wiki. These pages describe the concept as well as dialect specific XPaths to describe how the concept can be contained in a record. They are given for for the concepts that are missing from some records and concepts contained in the dialect but unused in the collection. These links also contain XML samples of how the concept is shared using DIF, ECHO, ISO, and CSDGM dialects. Not every concept has a hyperlink because all of the guidance pages are not yet created.

NCAR Labs Usage of Concepts in the DataCite Recommendation

	DCITE	ISO	MODS	RDA-CISL	netCDF
Total Number of Records	**8**	**541**	**1297**	**655**	**4**
Resource Identifier					
Resource Identifier Type					
Author / Originator		97%		4%	
Author / Originator Identifier					
Author / Originator Identifier Type					
Resource Title					
Publisher		99.8%	81%		
Resource Creation/Revision Date	75%	97%	8%		
Theme Keyword					
Keyword Vocabulary	75%	6%	71%		
Contributor Name		99.8%	89%		
Contributor Role		99.8%	89%		
Responsible Party Identifier Type					
Responsible Party Identifier					
Resource Creation/Revision Date	75%	97%	9%		
Resource Type					
Related Resource Identifier	63%	3%	60%	35%	
Abstract		99.5%	89%		
Spatial Extent	13%			49%	
Resource Language	75%		89%		
Resource Identifier					
Transfer Size	25%			35%	
Resource Format	75%				
Resource Version	13%	6%			
Rights	63%	6%			

RDA-CISL Evaluation

A mapping of the native dialect XML representation of the Research Data Archive to the DataCite metadata concepts was created. When the dialect rubric was created it was applied to all of the records in the RDA. All of the records had an Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) wrapper that needed to be removed. Additionally the records all declare the OAI-PMH namespace for the default namespace, but use the RDA schema. This should be addressed by declaring the default namespace to the schema that is used to create the records, not the OAI schema.



Metadata Improvement

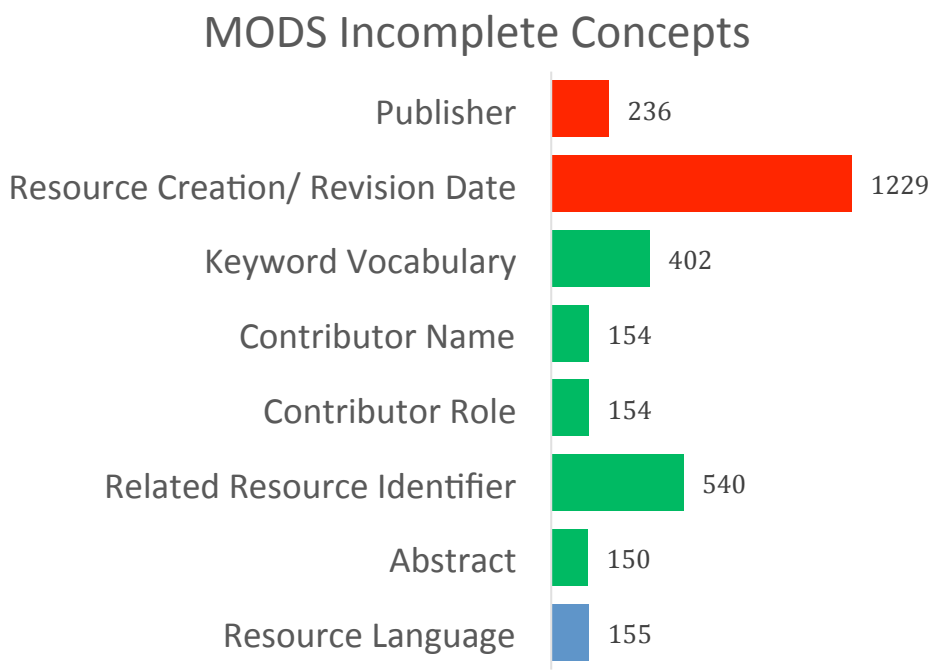
The concepts in the table below are either not contained in every record (incomplete), or in any record (unused). All of the concepts listed below can be contained in Research Data Archive native dialect records. Click on the concept below to access online guidance for writing the concept in a variety of dialects.

Guidance Links

Incomplete	Author/ Originator	Related Resource Identifier	Spatial Extent	Transfer Size	Resource Creation /Revision Date(2)	Contributor Name	Resource Type
Unused	Resource Format	Contributor Role					

MODS Evaluation

The MODS User Guidelines version 3 was used to map the dialect to the concepts found in the DataCite recommendation. The user guide comes from the Library of Congress and can be found here. There were 48 records that used a namespace from NSDL. These records were removed from the analysis. This is the largest collection analyzed.



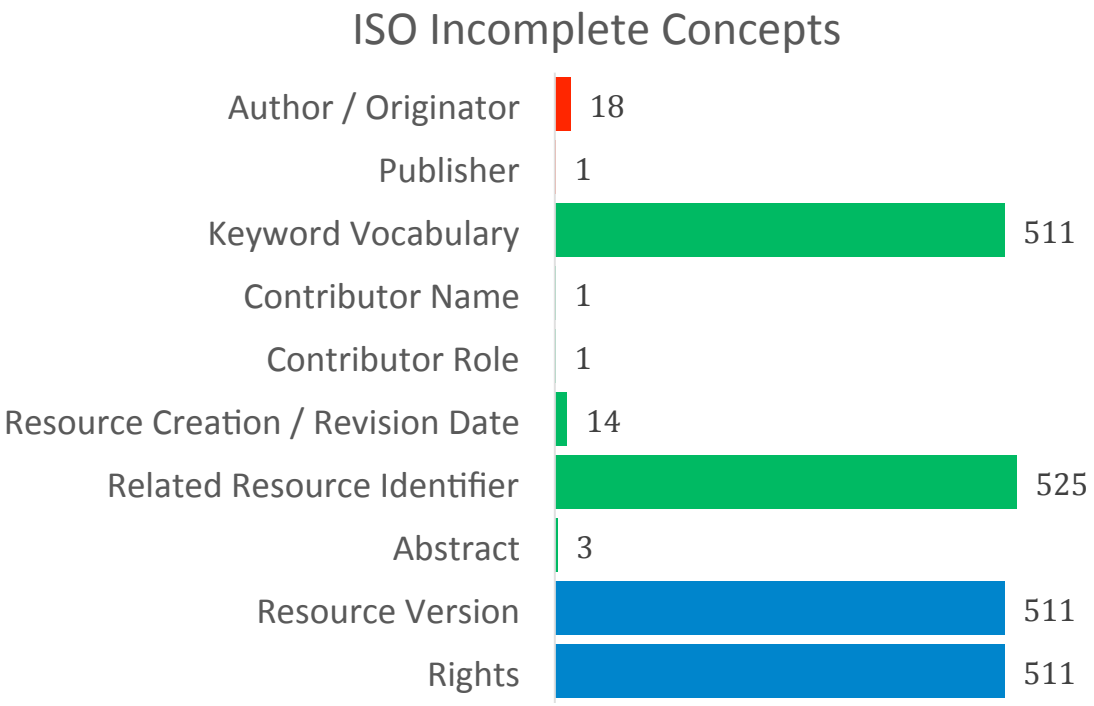
Metadata Improvement

The concepts in the table below are either not contained in every record (incomplete), or in any record (unused). All of the concepts listed below can be contained in MODS dialect records. Click on the concept below to access online guidance for writing the concept in a variety of dialects.

Guidance Links

Incomplete	Publisher	Resource Creation/ Revision Date	Keyword Vocabulary	Contributor Name	Resource Language
	Contributor Role	Related Resource Identifier	Abstract		
Unused	Spatial Extent	Resource Format	Resource Version		

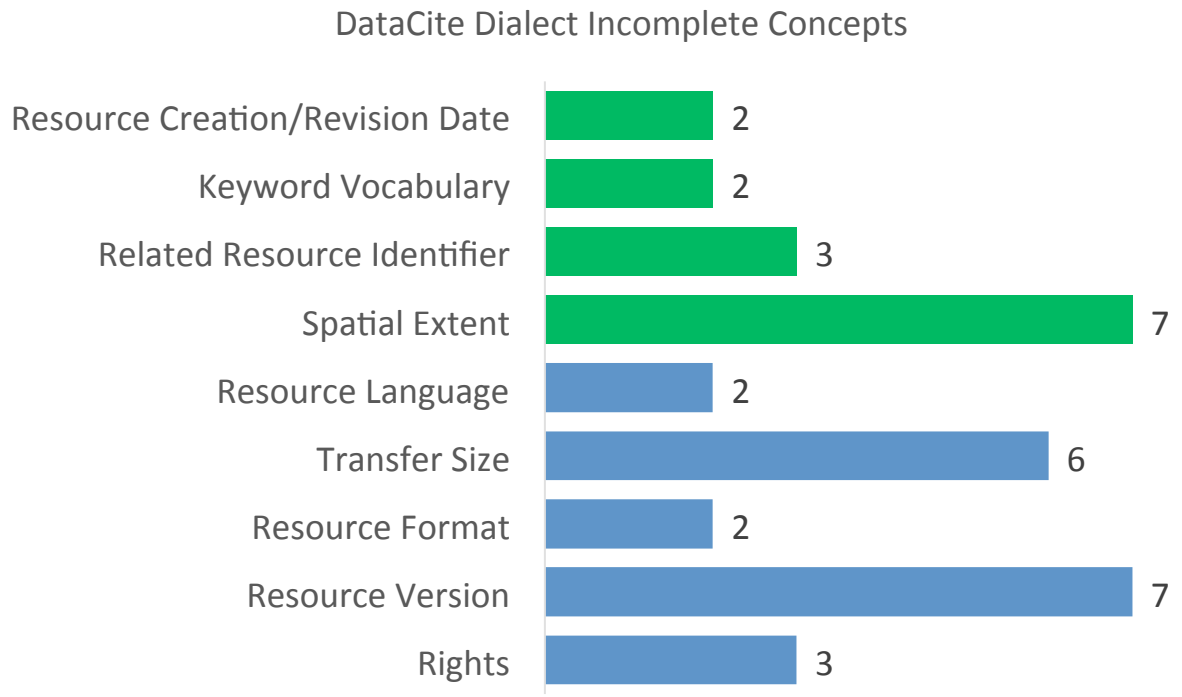
ISO Evaluation
The ISO records came from collections at EOL and Unidata as well as the DCERC ISO sample set of EOL records.



Metadata Improvement Guidance
The concepts in the table below are either not contained in every record (incomplete), or in any record (unused). All of the concepts listed below can be contained in ISO dialect records. Click on the concept below to access online guidance for writing the concept in a variety of dialects.
Guidance Links

Incomplete	Author/ Originator	Publisher	Keyword Vocabulary	Contributor Name	Contributor Role	Resource Creation/ Revision Date
	Related Resource Identifier	Abstract	Resource Version	Rights		
Unused	Transfer Size	Resource Format				

DataCite Evaluation
The DataCite sample set is the most diverse, coming from MMM, HAO, Unidata, and CISL. While the sample size is quite small there are many differences as to what is included in a record. It is worth noting that the records leave recommended concepts out even though the recommendation is from the same organization.



Metadata Improvement
The concepts in the table below are either not contained in every record (incomplete), or in any record (unused). All of the concepts listed below can be contained in ISO dialect records. Click on the concept below to access online guidance for writing the concept in a variety of dialects.
Guidance Links

Incomplete	Resource Creation/ Revision Date(2)	Keyword Vocabulary	Resource Version	Related Resource Identifier	Spatial Extent	Rights
	Resource Language	Transfer Size	Resource Format			
Unused	Resource Identifier Type	Author / Originator Identifier	Author / Originator Identifier Type	Contributor Role	Responsible Party Identifier Type	Responsible Party Identifier

netCDF Evaluation

The NcML files in the sample set were extracted from RAL NetCDF files. There are no concepts that appear in some files and not others.

Metadata Improvement

The concepts in the table below are not contained in any record (unused) while they do exist in the ISO-1 dialect. Click on the concept below to access online guidance for writing the concept in a variety of dialects.

Guidance Links

Unused	Keyword Vocabulary	Contributor Role

Glossary

Collection: A group of metadata records commonly organized by a data facility, organization or project and often stored in a database or web accessible folder.

Concept: General term for describing a documentation entity. Concepts can occur in many dialects where they are typically represented (in XML) by an element.

Dialect: A particular form of the documentation language that is specific to a community.

Dialect Maximum: The maximum number of concepts from a particular recommendation that are included in a particular recommendation. Note: the dialect maximum is always less than or equal to the recommendation maximum.

Element: An item providing a value for a concept, typically in an XML representation. Elements depend on dialects. They are the instantiation of a concept in a dialect.

Level: Recommendations may have different degrees of necessity associated with a concept's occurrence in a record. These subsets of concepts within a recommendation are called levels.

Recommendation: A set of concepts that an organization identifies for achieving a documentation goal.

Recommendation Maximum: The number of concepts included in a particular recommendation. Note that the recommendation maximum is the maximum completeness score available for a metadata record being evaluated with respect to that recommendation. The recommendation maxima are always greater than or equal to all dialect maxima for that recommendation.

Signature: A series of numbers that give the number of concepts/elements missing from a metadata record (or a group of metadata records) in a series of levels. Signatures with low numbers indicate fewer missing elements and a signature made up completely of 0's indicates a record or group of records that is complete with respect to a particular recommendation/dialect combination. A signature of 2 3 indicates that 2 elements are missing from the first level and 3 are missing from the second. The sum of the numbers in a signature is the total number of elements missing from a record or group of records.