DataCite Recommendation Analysis for the National Center for Atmospheric Research

2015-12-09 - Draft

**Table of Contents**

Executive Summary 2

DataCite – What is it? 3

Recommendation Dialect Comparison – How Does My Dialect Fit? 3

Recommendation/Dialect Maximum Graph 4

Recommendation/Dialect Comparison Report 5

DataCite Concepts missing from NCAR dialects 6

Metadata Analysis – How Complete are My Metadata? 7

Completeness Results and Missing Elements for USGS Metadata Records 7

Signature Groups with Concept Occurrence 9

Specific Guidance – How Do I Improve My Metadata 10

Mandatory Level 10

Recommended Level 11

Optional Level 11

Appendix 12

Quick Metadata Evaluation 12

Glossary 15

# 

# Executive Summary

This report presents the results of a NCAR metadata completeness evaluation 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.

The National Center for Atmospheric Research has many ways of sharing the data they produce and archive. The DSET was charged with a mission to unify these ways of sharing. Currently there are 9 labs, each with unique assets. There are a variety of metadata employed, some groups use xml standards from external groups, 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, RDA-CISL and EOL 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 currently hoping to improve the identification of their metadata’s dataset through DOIs.

One important observation is that all of the metadata dialects do not include all of the concepts prescribed by the DataCite recommendation, save for the DataCite dialect. As one might expect, 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.

Our metadata sample included X records from Q NCAR labs. Of those, Y (~Z%) included all of the metadata elements in the DataCite recommendation that are contained in the dialect the record is written in. Other groups of records were missing information about Sean, Ted, Lindsay, and John.

Terminology used in this report is defined in the Glossary.

# DataCite – What is it?

[DataCite](https://www.datacite.org) is an organization founded to help make data more accessible and usable. 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, and their members assign them. They are also known as the originators of Digital Object Identifiers (DOIs).

In the context of the terminology we use (see [Glossary](#_Glossary)), DataCite is an organization that created a set of recommendations at three levels, mandatory, recommended, and optional (described in the [DataCite Metadata Schema](http://schema.datacite.org/meta/kernel-3.1/doc/DataCite-MetadataKernel_v3.1.pdf)) 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](http://wiki.esipfed.org/index.php/Data_Discovery_(DataCite)). The dialect is currently being used in the DataCite [search portal](http://search.datacite.org/ui) and in creating DOI landing pages. The recommendations are useful for communities looking for expert guidance about metadata concepts that are useful for 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](http://schema.datacite.org/meta/kernel-3.1/doc/DataCite-MetadataKernel_v3.1.pdf) 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 data outputs.

This assessment of a sampling of X collections from the National Center for Atmospheric Research is based on the DataCite3.1 recommendation. The NCAR collections use 5 xml dialects, ISO, MODS, DataCite, netCDF, RDA-CISL.

# Recommendation Dialect Comparison – How Does My Dialect Fit?

Recommendations are created in order to address metadata needs 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. It presents these in several ways:

1. A recommendation comparison report
2. A chart comparing the number of concepts in the DataCite recommendation and the NCAR dialects.
3. A Recommendation/Dialect comparison that lists all concepts in the DataCite recommendation and NCAR dialects.
4. Tables that describe the concepts in the DataCite recommendation that are missing in the NCAR dialects, and where those concepts are implemented in other dialects.

## Recommendation Comparison Report

The purpose of the Recommendation 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 for achieving a documentation goal. The recommendation levels included in this study are: DataCite 3.1 Mandatory, DataCite 3.1 Recommended, DataCite 3.1 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 | 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 |

Translation Version: 0.1 (Feb. 9, 2015)

## 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 concepts) include 8, 11, 6 concepts 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. 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. The lower lines in the Figure shows these dialect maxima.

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. These numbers are reflected in the gaps between the two lines in the Figure.

## Recommendation/Dialect Comparison Report

This report (link TBD) shows all of the concepts included in the DataCite recommendation and verifies their existence in the NCAR dialects with an “X”.

**Mandatory Level**

| Concept | Score | Description | RDA-CISL | ISO | MODS | netCDF | DCITE | ISO-1 | EOL |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Author / Originator | 7 | The principal author of the resource | X | 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 | 6 | Publisher of the cited resource | X | X | X | X | X | X |  |
| Resource Creation/Revision Date | 7 | The date the resource was created | X | X | X | X | X | X | X |
| Resource Identifier | 7 | Identifier for the resource described by the metadata | X | 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**

| Items | Score | Description | RDA-CISL | ISO | MODS | netCDF | DCITE | ISO-1 | 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 | 7 | The role of any individuals or institutions that contributed to the creation of the data. | X | X | X | X | X | X | X |
| Keyword Vocabulary | 6 | 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 | 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 | 7 | The date the resource was created | X | X | X | X | X | X | X |
| Resource Type | 5 | A resource code identifying the type of resource; e.g. dataset, a collection, an application (See MD\_ScopeCode) for which the metadata describes. | X | X | X |  | X | 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 |  |  |
| Spatial Extent | 7 | The spatial extent of the resource. | X | 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 | RDA-CISL | ISO | MODS | netCDF | DCITE | ISO-1 | EOL |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Resource Format | 6 | The physical or digital manifestation of the resource | X | X | X |  | X | X | X |
| Resource Identifier | 7 | Identifier for the resource described by the metadata | X | X | X | X | X | X | X |
| Resource Language | 5 | The language of the resource. |  | X | X |  | X | X | X |
| Resource Version | 4 | Version of the cited resource |  | X | X |  | X | X |  |
| Rights | 6 | Information about rights held in and over the resource | X | X |  | X | X | X | X |
| Transfer Size | 5 | The size of the digital resource | X | X |  |  | X | X | X |

## DataCite Concepts missing from NCAR Dialects

The Tables below provide lists of the DataCite concepts that are missing from the NCAR dialects along with a description of the concept. The DataCite dialect contains all 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, ISO-1

**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 |

# 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 downloaded from 25 USGS metadata collections that are included in the [USGS ScienceBase](https://www.sciencebase.gov/catalog/) using the ScienceBase OAI-PMH service. We requested 50 records using a single OAI-PMH request for each collection. In some cases the collections include less than 50 records, so we have the complete collection. For larger collections (>50 records) we hope that these provide a representative sample. 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 and Missing Elements for USGS Metadata Records

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 CSDGM record is as complete as possible with respect to the DataCite recommendations. There are 17 complete records in this sample of 617 (top row in Figure below). The largest groups of records have scores of 1 1 2 (one missing mandatory and recommended concept and two missing optional concepts, 109 records) and 0 0 1 (one missing optional concept, 105 records).

The following Table shows the number of records in the collection for each signature score. Each row in the Table corresponds to a group of records with the same set of scores (termed a signature), given in the first column. The second column gives the total number of records in each group, e.g. 17 records have the signature “0 0 0”. Subsequent columns give the number of records that include specific concepts given in the titles. The DataCite recommendation concepts that do not occur within the CSDGM dialect have been removed.

Each of the 17 records with a signature 0 0 0 (in the first row) contain all of the concepts, so each column has a 17. It is often the case that metadata records with a particular signature score share the same concepts. The records in the second group (0 0 1) are all missing the Resource Format concept, indicated by a zero in the Resource Format column (shaded) for the 0 0 1 signature row. This homogeneity, identical record counts across entire rows, is shared with 11 out of 14 signature groups. The remaining three signature groups: 0 1 2, 1 1 1, and 1 1 2, have records that do not contain the same concepts as the rest of the signature group (bold cells). The percentage of records that deviate from their signature group is just under 5%, and every record that deviates from it’s signature group deviates in the same way as other deviant records in that signature group.

The 17 records in the sample collection that are complete with respect to the DataCite recommendation are bright spots that can be utilized as examples to inform the process of improving the rest of the collection. They can also help identify expert knowledge within the organization that can serve as a resource for the metadata improvement process.

## Signature Groups with Concept Occurrence

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Signature** | **Record Count** | **Author/ Originator** | **Publisher** | **Resource Creation/ Revision Date** | **Contributor Name** | **Resource Type** | **Spatial Extent** | **Transfer Size** | **Resource Format** | **Rights** |
| 0 0 0 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 |
| 0 0 1 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 0 | 105 |
| 0 0 2 | 56 | 56 | 56 | 56 | 56 | 56 | 56 | 0 | 0 | 56 |
| 0 0 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 0 1 1 | 58 | 58 | 58 | 58 | 58 | 0 | 58 | 58 | 0 | 58 |
| 0 1 2 | 91 | 91 | 91 | 91 | **90** | **3** | **89** | 0 | 0 | 91 |
| 0 2 2 | 12 | 12 | 12 | 12 | 12 | 0 | 0 | 0 | 0 | 12 |
| 1 0 0 | 22 | 22 | 0 | 22 | 22 | 22 | 22 | 22 | 22 | 22 |
| 1 0 1 | 96 | 96 | 0 | 96 | 96 | 96 | 96 | 96 | 0 | 96 |
| 1 0 2 | 23 | 23 | 0 | 23 | 23 | 23 | 23 | 0 | 0 | 23 |
| 1 1 1 | 19 | 19 | 0 | 19 | 19 | **1** | **18** | **1** | **18** | 19 |
| 1 1 2 | 109 | **84** | **25** | 109 | 109 | 0 | 109 | 0 | 0 | 109 |
| 1 2 2 | 6 | 6 | 0 | 6 | 6 | 0 | 0 | 0 | 0 | 6 |
| 2 2 1 | 2 | 2 | 0 | 0 | 2 | 0 | 2 | 0 | 2 | 2 |
| Total | 617 | 592 | 365 | 615 | 616 | 324 | 596 | 299 | 59 | 616 |

*red = mandatory, blue = recommended, green = optional*

Using the counts in the table above, it is possible to determine the number of occurrences of each concept and to inform decisions about which concepts to improve upon first for the greatest impact. The following graph helps visualize these “bang for the buck” opportunities. It shows that the majority of concepts listed occur in most records, having a missing concept count that is close to zero. However, there are four concepts that have much higher counts. Improving utilization of these concepts is the quickest improvement that can be made to collections, as organizational knowledge already exists on how to collect these concepts and include them in the CSDGM dialect.

*red = mandatory, blue = recommended, green = optional*

# Specific Guidance – How Do I Improve My Metadata

The analysis above identifies specific concepts that are missing from USGS metadata records, but are included in the CSDGM dialect. This section provides specific guidance on how to write metadata for those concepts in the CSDGM dialect. A positive and straightforward first step is to assess what some USGS records already include and implement them collection wide.

## Mandatory Level

Since the mandatory level is the most important to the recommendation, it should be addressed first. Most of the concepts appear in almost all of the records. The concept Publisher appears about 60% of the time. Adding a publisher to the remaining 40% of records will result in all but 27 records or 4% of the sample set at the dialect max. This remaining 4% not at dialect max can be improved by ensuring that the Author and Resource Creation/ Revision Date elements are added to the records that do not contain them.

|  |  |  |  |
| --- | --- | --- | --- |
| Concept | Description | XPath | Missing % |
| Publisher | Publisher of the cited resource | /metadata/idinfo/citation/citeinfo/pubinfo/publish | 40.84 |
| Author | The principal author of the resource | /metadata/citation/citeinfo/origin | 4.05 |
| Resource  Creation /  Revision Date | The date the resource was created | /metadata/idinfo/citation/citeinfo/pubdate | 0.32 |

## Recommended Level

In the recommended level the scores range from 2 to 0. The Resource Type concept occurs about half the time. Once this concept is included in all records, the sample set will only have 14 records or 2.3% that are only one concept away from dialect max. This 2.3% can be improved by addressing the remaining concepts listed below.

|  |  |  |  |
| --- | --- | --- | --- |
| Concept | Description | XPath | Missing % |
| Resource Type | A resource code identifying the type of resource e.g. dataset, a collection, an application (see MD\_ScopeCode) which the metadata describes | /metadata/distinfo/resdesc | 47.49 |
| Spatial Extent | The spatial extent of the resource | /metadata/idinfo/spdom/bounding | 3.40 |
| Resource  Creation / Revision Date | The date the resource was created | /metadata/idinfo/citation/citeinfo/pubdate | 0.36 |
| Contributor Name | Contributor to the resource | /metadata/idinfo/datacred | 0.16 |

## Optional Level

In the Optional level we have two concepts that are included in some, but not the majority of the records. Only 10% of records include a Resource Format. Improving documentation of this concept, as well as Transfer Size, will result in all but 1 of the records in the sample set reaching the dialect max for the Optional Recommendation. The one record is missing Rights.

|  |  |  |  |
| --- | --- | --- | --- |
| Concept | Description | XPath | Missing % |
| Transfer Size | The size of the digital resource | /metadata/distinfo/stdorder/digform/digtinfo/transize | 48.46 |
| Resource Format | The physical or digital manifestation of the resource | /metadata/distInfo/distributor/distorFormat/formatName | 90.44 |
| Rights | Information about rights held in and over the resource | /metadata/idinfo/accconst  /metadata/idinfo/useconst | 0.16 |

# 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 spirals. 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 spiral 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.

**Spiral**: A set of concepts required to support a particular documentation need or use case.