Controlled Vocabularies for Metadata Interoperability

What is a Controlled Vocabulary?

In many professional communities, the terminology used to communicate is often community-specific. For example, consider the use of the word 'altitude'. Altitude refers to the height of something above a reference point (e.g., sea level). Although altitude is a common term, if we were examining a set of blueprints for a building we probably would not think of using the word 'altitude' to describe the level of the rooftop. Most likely, we would associate 'altitude' with aviation; while we would describe the building's roof using the word 'height'. Similarly, if we were in a boat looking downward into the water, we probably wouldn't use 'altitude' or 'height' to describe the position of the bottom; rather we would use the word 'depth'.

The three words – altitude, height, and depth – are all similar in that they represent measures of distance relative to a fixed level; but they are all used differently. As well, they can be associated with different communities. In this simple example, those communities include aviation, architecture and oceanography. What's more, the same term might be used differently in a different community, as for example when oceanographers use 'altitude' to mean the distance above the ocean floor.

Collectively, all the terminology for a community represents a specialized vocabulary for that community. In some cases, these vocabularies are formally managed. Here, managed means the terms are stored and maintained using agreed upon procedures. Procedures might exist for adding terms, modifying terms and in the rarest cases, for deleting terms. When vocabularies are formally managed, they become a controlled vocabulary.

Usage vs. Discovery Vocabularies

Previously, we noted the term 'altitude' as describing part of the spatial position of something. We may complete the spatial description by including the terms 'latitude' and 'longitude'. The term 'latitude' typically refers to a value that describes the y-coordinate of something on the earth (more generally a spheroid, but let's keep this discussion to our own planet). With the term 'longitude' to describe the x-coordinate, we can fully specify the 'position' of something on the earth.

Now consider a data asset that contains altitude, latitude and longitude values. We can think of the asset as being a database table, spreadsheet or text file. The asset will likely have names for the columns of numbers. The names could be 'altitude', 'latitude' and 'longitude'. Alternately, the names could be cryptic codes such as ALT, LAT and

LONG. The terms (or names) used within the asset represent what we refer to as a *usage vocabulary.*

A usage vocabulary is important when clients want to utilize the data within the asset. Software applications, or people, have to understand these terms in order to effectively access and use the data within the asset.

However, discovering the content of the asset is different from utilizing the content. In the discovery process, the usage vocabulary may or may not be useful. In the case where cryptic codes are used to identify the data column, the usage vocabulary is not useful. This is because the search software (or the people) will not likely think of using the exact cryptic codes that are used within the asset.

In this case, we introduce another vocabulary – the *discovery vocabulary*. The discovery vocabulary uses terminology to identify the data that are common to the subject community. Terms in the discovery vocabulary are very diverse; thus making the vocabulary itself difficult to define.

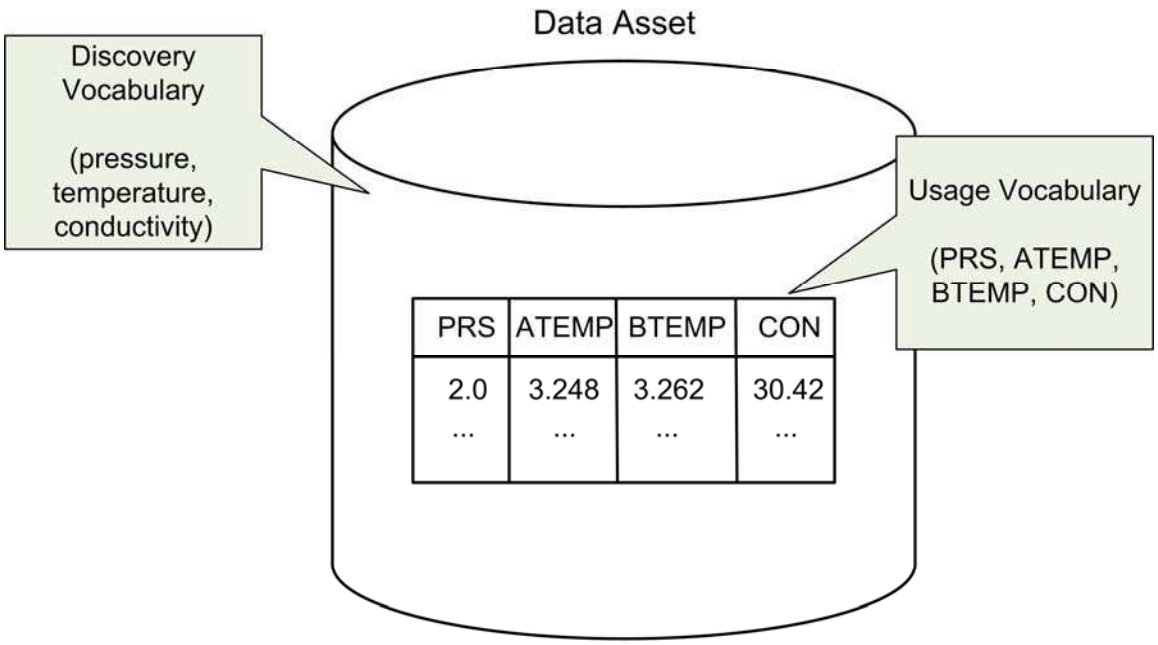
Terms in the discovery vocabulary often represent an aspect of the data asset that has a common description in the subject community. These terms can take a variety of forms.

1. Terms in the discovery vocabulary may be identical to terms in the usage vocabulary. This is the situation when the data asset uses common language terminology to identify the data. An example would be a data asset containing data values identified as 'temperature' or 'salinity'. Both of these terms are part of the usage vocabulary, and since they are natural search terms, they also would be terms in the discovery vocabulary.

2. Terms in the discovery vocabulary may represent groups of terms in the usage vocabulary. This is a common situation for legacy assets, where cryptic codes have been used to identify similar data from multiple sources. As an example, consider a legacy data asset that contains temperature values from sensors A, B, and C. Suppose these data are identified within the asset as ATEMP, BTEMP, CTEMP (i.e., terms in the usage vocabulary). The discovery vocabulary term that encapsulates all three usage terms would be 'temperature'. In this case, the 'temperature' term in the discovery vocabulary represents a group of terms from the usage vocabulary.

3. Terms in the discovery vocabulary may represent groups of data values. In this case, the discovery vocabulary terms identify particular subgroups of the data, rather than all of the data. As an example, if the data asset contains geology data then certain geological time periods (e.g., Mesozoic Era) may be identified in the discovery vocabulary. In physical oceanography, a discovery term may identify a particular water mass (e.g., Labrador Sea Water) which has particular characteristics (i.e., physical or chemical).

Discovery vocabularies aid the client in finding the data asset, while the usage vocabulary aids in utilization of the asset. Both vocabularies can pertain to data-related topics such as parameters, platforms, sensors, geographic areas, etc.



Semantic vs. syntactic Vocabularies

The semantics and syntax of a vocabulary are also important.

Semantics provide meaning to the terms, in a way understandable to a person (e.g., Altitude refers to the vertical position of a flying object).

Syntax is related to the format instructions for storage of values in computers. Syntax might include information on values such as float, real, ASCII or binary (e.g., Altitude data values are measured in feet, float 8.6).

As noted previously, usage vocabularies provide information on the terminology for using the data values. Thus, a usage vocabulary would include both semantic metadata (e.g., the term 'latitude' and the definition of what this term means) and syntactic metadata (e.g., the data value for latitude is float 8:6).

A discovery vocabulary does not typically contain syntactic metadata. This is because the discovery vocabulary describes collections of usage terms or data values that have

meaning to the community. For example, 'North Atlantic Ocean' represents a certain grouping of latitude/longitude values; and represents a specific meaning to the oceanographic community. The term 'North Atlantic Ocean' has syntax within itself (e.g., capitalization, allowed spaces) but does have syntax associated with the content; because there is no direct content. The term 'North Atlantic Ocean' doesn't explicitly contain values, but rather is a description of a collection of values.

What is the relation between a Metadata Standard and a Controlled Vocabulary?

Metadata is used to describe the aspects of 'something'. In the MMI community, the 'something' could be most anything related to the marine community, such as a data set, a marine oriented service, etc.

A metadata standard, also known as a content standard, is used to define the 'containers' for the metadata. The metadata standard is like a list of possibly important items for describing 'something'. The metadata standard describes this all-encompassing list of containers. Since the list is all-encompassing, any specific container may or may not be of use for the particular thing being described.

A controlled vocabulary could be used as content for the metadata containers. The controlled vocabulary, which is a managed list of acceptable values, will predetermine all the possible 'stuff' in the 'container'. If a value is not included in the 'stuff', it will not stay in the 'container'. This trial and error is also a mechanism for controlled vocabulary maintenance. In this case, the controlled vocabulary could be defined internal or external to the metadata standard.

An example may help to distinguish the metadata containers from the controlled vocabulary. Suppose we are going to purchase a vehicle. In this example, the vehicle represents the 'something' we are going to describe. A hypothetical metadata standard to assist us in vehicle selection might consist of an all-encompassing list of important containers that describe any possible vehicle. These containers could include: model name, color, number of doors, number of passengers and whether or not there is a real spare tire (i.e., as opposed to the donut spare). Note that the descriptors in the all-encompassing list do not apply to every possible vehicle. For example, a motorcycle has no spare tire at all and thus the descriptor for 'real spare' doesn't really apply.

Now consider the number of doors. In this case, the controlled vocabulary may be represented by the numbers 0, 2, 3, and 4. This vocabulary allows for all passenger vehicles produced by all manufactures. The zero case accounts for motorcycles; the two door case for two-door cars; and the three and four door case for the older and newer style vans (note, we are ignoring the gate or hatchback). No other value for the number of doors is possible in passenger vehicles.

Metadata Standard	Controlled Vocabulary		
Model			
Name			
Color	Red Blue Black Green Gray	Blue	Blue
Number of Doors	0 2 3 4	4	0
Maximum Occupants			
Real Spare	Yes No	Yes	

In this example, the metadata standard represents the structure and metadata descriptors, while the controlled vocabulary is the content (e.g., the 0, 2, 3, 4 list) of the descriptors.

Why are Controlled Vocabularies Important?

The vehicle example above helps illustrate the importance of controlled vocabularies. The controlled vocabulary maintains the listing of defined and agreed upon terminology. The vocabulary is managed in the sense that it has maintenance procedures for the creation, updating or modification of the terms.

Controlled vocabularies are very useful. The vocabulary:

- establishes the permissible language to be used by the community;
- maintains the proper and agreed upon spelling of the terms within the vocabulary;

- clarifies terms for those who are new to the community; and
- helps the community avoid the use of arbitrary terms that often cause inconsistencies and confusion.

Avoiding misspellings and inconsistencies always promotes understanding. However, in the world of computers the controlled vocabulary offers enhanced capabilities because in this case, the vocabulary can be incorporated into automated procedures.

In a data system, such a vocabulary can become part of the input and quality control aspects of the system. The vocabulary could be part of the input by providing users or other systems with a list of allowed input for the metadata descriptions. Similarly, the vocabulary can be used to check existing or imported metadata descriptions for consistency in content, including such things spelling.

A Last Resort: Developing a Local Vocabulary

Developing Controlled Vocabularies

How to determine the terms

To identify the terms of the controlled vocabulary, you need to first examine the descriptions of your assets, looking for discrete (i.e., non continuous) content. Things that are measured are usually continuous, while things that have specific descriptions are usually discrete. Also, if you can count the total number of possible descriptions, it is likely to be discrete.

If the possible content of the metadata descriptor is found to be discrete, then it is a likely candidate for a controlled vocabulary. For example, if the descriptor was *ocean_name* and the content was the name of the ocean, then the ocean names could be added to the system as terms in a controlled vocabulary. In this case, the controlled vocabulary contains the five ocean names.

Once you have identified those descriptors that contain discrete terms, you must identify all possible terms to be contained in the descriptors as values. This is the list of terms for your vocabulary. You should be able to provide a definition of each term, such that its definition is unique to that term. This definition development is a process of building a dictionary of terms for the controlled vocabulary.

Scalability - allowing for additions

The scalability of a controlled vocabulary is an important aspect. The vocabulary should not be limited by the initial terms in the list. To avoid this, you need to examine the terms and think about the general class of things that all the terms are describing. Don't think about an individual term (or an individual car, to extend the vehicle example). Rather, think about the general class of things. Now, attempt to define attributes of the general class. This may not be an easy process. However, if you are successful your vocabulary will be scalable. This process is also an excellent step towards the development of an ontology.

Tips and Tricks

Here are a few tips when developing your controlled vocabulary.

- don't have vocabulary terms with embedded information

  Don't encode information within the vocabulary terms. As an example, a term that contains encoded information may have certain characters as meaning certain facts about the term. For example, a single term like XT07aa might indicate an XBT temperature from a T-7 computed using coefficient set aa. Such a term contains information on the type of sensor, the model of sensor, the parameter being measured and processing information. This type of information should be split out of the single term, into multiple terms.

- think about future grouping of terms

  At some point, you may have to start grouping values associated with the terms in the usage vocabulary; effectively creating a discovery vocabulary. Allowing for such grouping will help in the management of both vocabularies and in the user discovery of terms. Your vocabulary management should be capable of adding this grouping with minimal impact on the management system.

- don't allow users to manage the vocabulary

  Users need a mechanism to suggest new terms for the controlled vocabulary but they cannot be given the ability to add new terms. A controlled vocabulary is just that – controlled. It is controlled to avoid confusion among terms and to avoid the introduction of errors. Additions, deletions or corrections must be managed by the person responsible for the vocabulary.

- Units are important

  Your usage vocabulary may or may not contain explicit units. For example, the data terms in the usage vocabulary may have a direct association with the unit (i.e., one term can only have one unit). A more

preferred method is to allow multiple units for a single data term (e.g., distance can have units of metres or kilometres). By allowing multiple units you effectively introduce another type of vocabulary that your system must support – a unit vocabulary.

- the same syntactic rules

  The terms used in the vocabulary will be created using a set of syntactic rules that may involve capitalization, the use of underscores, or the use of other special characters. The vocabulary must be developed with consistent application of these rules throughout the vocabulary terms.

- use natural terms

  Whenever possible, natural terms that are commonly used within the community, should be used within the vocabulary. However, if these terms introduce ambiguity, then consider other terms. Unambiguous terms and definitions are the cornerstone of the vocabulary.

- unambiguous definition

  The terms used in your vocabulary should be associated with rigorous definitions. These definitions should be unambiguous to the community using the vocabulary.

Developing Controlled Vocabularies for Legacy Data

How to approach this project

It is important that you approach this task with a long term vision and commitment to that vision. To do this task properly, will first and foremost require knowledge in the field from which the data originate. Knowledge in the field will provide you with the ability and credibility to talk to the people who really know the data – the scientists who collected or produced it. This knowledge will require time to build.

The first thing you will need is a plan for vocabulary development. This plan may actually be a subcomponent of a larger plan – a plan for metadata capture and management. However, here we only deal with the vocabulary subcomponent.

You should start with a small subset of the data. Divide the entire data set that you will be dealing with into logical pieces (logical from your point of view). If your data asset is a collection of many data sets collected from oceanographic cruises, start with a single cruise. Then subdivide further, perhaps by topic (physics, chemistry, biology, or geology). Start with the topic you know the most about.

Now examine the data. You should be looking for the different types of data that were collected, the different instrumentation or procedures that were used and different units that may be possible for the data types. You can start with compiling a list of names that refer to the data types. Also make a list of allowed units for those names. Finally, start to document the procedures followed to collect or process the data type (if you are lucky, there will be existing documentation on procedures). These lists will form the basis of your vocabularies. For example, the list of data types will form the starting point for your usage vocabulary.

The usage vocabulary will require a bit of extra work on your part. You should investigate the provenance or history of the data names and values associated with these names. During this process you should examine the various data quantities and the names affixed to these quantities. Ask yourself if two quantities with different names are actually the same, or alternately, if two with the same name are actually different. This terminology evolution should be documented, as it will be extremely useful in the development of a thesaurus, metadata mappings, and general documentation. You should think about different procedures for acquiring or processing the data. Finally, don't forget units and don't underestimate units. A considerable amount of complexity exists in the domain of units – and if the units are abbreviated differently, they are different (e.g., don't think for a second that oxygen content in mg/l is equivalent to ml/l; even if the values are similar).

In this process, no detail is too small. The research environment is full of cases where multiple procedures exist to measure the same data type. For example, two different biological incubation setups may produce measurements of the same data quantity. These different procedures represent important metadata that needs to be associated with the data quantity. However, the usage vocabulary needs to indicate the same data term is being measured. Another vocabulary notes the differences in the measurement procedures.

At this point you should start to realize that your job as data custodian has been morphed into a combination of data system designer, scientist, investigative police officer and investigative news reporter.

Tips and Tricks

The "Tips and Tricks" from the previous section also apply to legacy data. As well, the following apply.

- Archived reports are valuable

  For legacy data, there will likely be a limited source of information in documents produced during data collection, processing, or the reporting of results. These documents are valuable resources and will need to be searched for metadata relevant to the data set. However, if you find

planning documents that outline planned data collection, keep in mind that these represent the planned collection – actual collection will likely be different.

- Different procedures that quantify the same data type

  In a research setting (e.g., a chemistry or biology lab, or instrumentation development shop), you should be looking for different procedures that quantify the same data type. For example, procedure A and procedure B (or instrument A and instrument B) may both be used to measure values for a single data type. Both procedures (or instruments) measure the same parameter, and as such, the name associated with the data from both procedures should be the same.

  However, the information detailing the different procedures must be documented and maintained with the data value. Thus, a separate vocabulary detailing the procedures should be created. This vocabulary is effectively an instrumentation/procedure vocabulary that notes for each data type, the procedures used to determine the data value.

Use Case

Now we consider the development of a controlled vocabulary. For this particular case, we consider a legacy data rescue project.

A data rescue project is an effort directed towards recovering data that are presently inaccessible. Such an effort could result from a situation where data have been collected over many years by individual scientists. Perhaps these scientists are nearing the end of their careers – perhaps some have already retired. The data they have collected during their careers is now in jeopardy of being lost.

Suppose these scientists have been collecting data during their 30+ year career. Since these scientists started in an era before common computer use, many of the data sets exit in paper form. In all likelihood, the data sets are in office filing cabinets or in the basement of the building where they work. These data may not exist in electronic form. Nevertheless, the data represent a wealth of historic information and are in fact, irreplaceable. The data may also contribute to long-terms data sets, a particularly important topic for helping understand long-term, global trends. These data need to be rescued and placed in managed databases at the organisational and national levels.

The first step in this process is to begin collecting information from the individual scientists. The scientist, or possibly those involved in the field programs, may have documentation on data collection plans that pertain to individual data sets. There may also be log books, or field journals that were used for notes during the field activity. As well, reports may exist that describe the actual field program that resulted in the data set.

Collect examples of the data, paper listings or plots. Enquire as to availability of digital data. Start to make notes as to the types of collected data. These notes should include the sampling procedures or instruments used for the particular data. Keeps notes on when the data were collected and a general idea as to where it was collected. For oceanographic data, you might want to approximately locate the data sampling using Marsden Squares. The spatial information may be useful for prioritizing the rescue effort. Determine if there are hardcopy or softcopy records. A particularly difficult problem is when the "data" exists as a physical sample (we won't deal with that here). How are the data stored and are there backup copies? Also, is there any activity currently underway to rescue these data? You will need all of this information to help you understand the data you are rescuing.

The terminology you use for the collected data will be the starting point for your usage vocabulary. Initially when you are scanning the documentation and legacy data, don't be concerned about building a vocabulary. Rather, you should be concerned about building your own knowledge as to the collected data.

After you have reviewed numerous data sets from various scientists, review your notes on the data that was collected. There will be different, but similar, terminology used for the collected data. As well, review the procedures or instruments looking for similar instruments collecting data that has been named differently. Using this type of information, revisit the scientists and attempt to clarify if the data names you have noted as different, are in fact the same (or alternately, if the same names are really different parameters).

This process will likely reduce the number of terms in your list of parameter types, as you will find different terms that are in fact referring to the same parameter. With this reduced list, define the other important attributes for the terms. For example, the date you are formally creating the term and the limits on the values associated with the term. If you are storing this information in a database, make sure you assign a unique identifier to each term. A short description of the term, and longer more detailed description as to what the terms means should also be noted.

This list of terms forms your controlled usage vocabulary. Now you need to consider how your terminology matches with the organizational and national terminology. As well, you need to decide if the organisational and national terminology meets your needs. If it does, then use the organisational or national usage vocabularies. Also, identify if you need to suggest updates to these vocabularies.

These steps should help you form the initial parts of a controlled vocabulary. Since you are likely the person with the most knowledge on this vocabulary, you should be the person responsible for its management.