Representing, Exploiting, and Extracting Metadata using Metadata++

Mathew Weaver	Bill Howe	Lois Delcambre	Tim Tolle	Dave Maier
mweaver@cse.ogi.edu	bill@cse.ogi.edu	lmd@cse.ogi.edu	ttolle@fs.fed.us	maier@cse.ogi.edu

1 Introduction

Metadata++ is a superimposed metadata model that provides enhanced flexibility and functionality for creating, maintaining, and searching document metadata. The Metadata++ model has its roots in a traditional thesaurus.

The Merriam-Webster dictionary defines *thesaurus* as:

> **a :** a book of words or of information about a particular field or set of concepts; *especially* **:** a book of words and their synonyms
> **b :** a list of subject headings or descriptors usually with a cross-reference system for use in the organization of a collection of documents for reference and retrieval

Metadata++ provides a more flexible and usable mechanism for the organization and retrieval of documents within an application domain. The motivating domain for developing Metadata++ is natural resource management, with the USDA Forest Service as our government partner.

There are two main building blocks used in the Metadata++ model: terms and document proxies. A term is a word or a phrase that represent some thing or concept within the application domain. Terms such as "Wenatchee National Forest", "air pollution", and "potential vegetation" exist within the forestry domain. The basis of Metadata++ is a finite set of terms.

As in a traditional thesaurus [1], terms can be related to one another in various ways. The first type of relationship is hierarchical. A hierarchical relationship may be used to relate broader terms and narrower terms, classes and their instances, and collections and their members. Hierarchical relationships would be used to express a taxonomy of species, as well as the geographic hierarchy among place names (i.e. "Cle Elum Ranger District" is part of "Wenatchee National Forest").

The second type of relationship is the synonym relationship. This relationship is used between terms that have very similar meaning or connotation. For example, the taxonomy of species includes both scientific names and common names. The scientific names and common names could be separated into different hierarchies, where each term has a synonym relationship with the corresponding term in the other hierarchy. This type of relationship allows Metadata++ to find documents attached to scientific species names even if the search is specified using common names (and vice versa).

The third type of relationship is the associative relationship. The associative relationship exists between terms that are not synonymous, but that have some association or correlation within the application domain. The ANSI Thesaurus standard [1] describes an associative relationship as a "relationship [that] covers associations between descriptors that are neither equivalent nor hierarchical, yet the terms are semantically or conceptually associated to such an extent that the link between them should be made explicit in the thesaurus" and goes on to say that "whenever one term is used, the other should always be implied. … Moreover, one of the terms is often a necessary component in any explanation or definition of the other." Metadata++ more broadly defines associative relationships to capture any arbitrary correlation

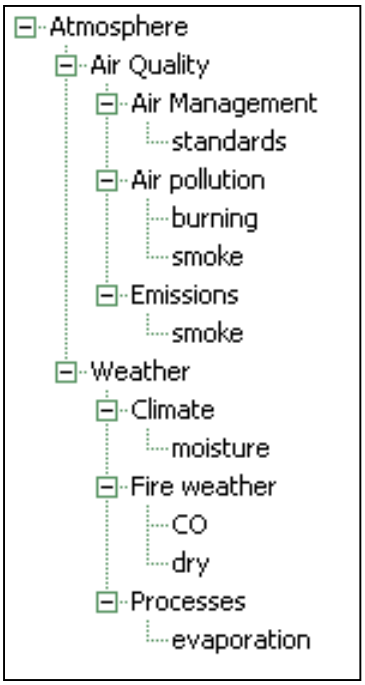

Figure 1: Climatologist Perspective

between terms specified by the user. For example, an associative relationship may be used to relate "Wenatchee National Forest" and "Western Hemlock" because Western Hemlock trees are found in the Wenatchee National Forest.

The second type of building block in Metadata++, document proxies, are objects that represent documents. These objects exist within the Metadata++ model, while the documents represented may exist elsewhere (an external web site, a document repository, as hard copy on a book shelf, etc.). The important content of a document proxy is some document identifier and some way of addressing (i.e., finding) the document. This may be a URL or an email address or a telephone number and a person's name. The metadata of the document is represented by relating the document proxy with terms found in Metadata++. Using document proxies provides two main benefits: isolating the document metadata from the format and location of the document itself, and allowing metadata to be created at sub-document granularity (i.e. multiple proxies can reference a single document).

One of the goals of a traditional thesaurus (see Z39.19) is to capture the well-defined vocabulary (including structure) of a particular domain. Much time and effort is required to construct a well-formed thesaurus that contains terms and relationships that adequately meet the needs of the majority of the intended users.

Metadata++ builds on a traditional thesaurus by allowing different individual thesauri to be constructed (and used) from the same set of terms. Each thesaurus – called a perspective – includes some subset of the entire set of terms, as well as zero or more relationships. These relationships may be common (i.e. shared by multiple perspectives) or specific to only one perspective. By integrating different perspectives, Metadata++ makes it possible for different types of experts (within an application domain) to construct and use their own specific thesauri – but have the benefit of sharing metadata (and documents) with each other. For example, a climatologist may construct a thesaurus as shown in Figure 1, while a Forester may construct a thesaurus as shown in Figure 2. Similarly, a single user may use different perspectives for different tasks.

These experts use similar (but not identical) sets of terms, but create different relationships between the terms. By allowing – and integrating – multiple thesauri, Metadata++ lets users share and search data from a large application domain using a vocabulary that is constructed specifically for their expertise.

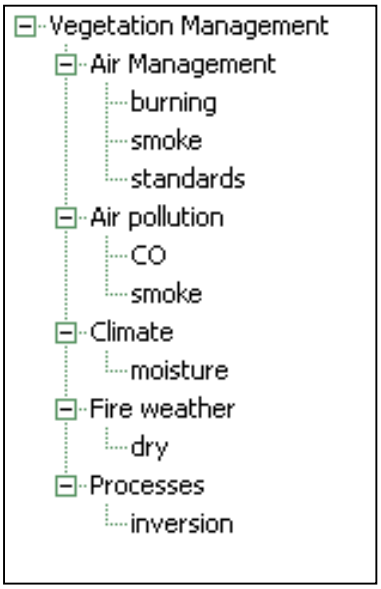

Figure 2: Forester Perspective

2 Perspectives

A perspective is a set of terms and a set of hierarchical, synonym, and associative relationships – each of which relating two terms. A user creates a new perspective by creating new relationships or selecting from existing relationships. Defining each relationship individually provides great flexibility – but could also be a burdensome task for perspectives containing large controlled vocabularies. Instead, a user may want to build a new perspective by combining pieces of existing

perspectives. Metadata++ allows perspectives to consist of individually defined relationships and relationships that exist in other perspectives.

In addition to simplifying the perspective creation process, referencing other perspectives allows different users to easily share common vocabularies. For example, consider the perspective that represents the taxonomy of species. The hierarchy within this perspective is large and complex – and rarely customized by individual users. Instead of making different users manually create the same taxonomy within their custom perspectives, the taxonomy perspective could be created by one expert and shared by others. Suppose a new species is discovered and added to the taxonomy perspective. All perspectives that share the taxonomy will automatically include the new species.

2.1 Metadata Attachment

One of the primary objectives of Metadata++ is to support document retrieval. Before documents can be retrieved, they must be put into the system. Once a document has been put into the system (i.e. one or more document proxies have been created to reference the document), metadata for the document can be created by relating the document proxy with terms. Creating metadata involves more than simply relating a document proxy object with a term object. Metadata is created with respect to a perspective (i.e. specified set of hierarchical, synonymous, and associative relationships). For example, a climatologist who is creating metadata could use the climatologist perspective, whereas a forester could use the forester perspective for creating metadata. Within a perspective, the user browses to one or more terms that are related to the document. For each term that the user selects, an attachment relationship is created. When the user selects a term and attaches it to the document proxy, Metadata++ stores the entire path (within the hierarchy of the perspective) used to select the term. This absolute path – starting at a root node and ending at the selected term – is referred to as a *trail*. The trail consists of the list of terms that lead from the root through the hierarchy to the selected term. For example, Figure 1 shows the climatologist perspective. Figure 3(a) shows the trail that would be used if the user selected the term "smoke" to associate with a document (D1). Figure 3(b shows the trail that would be used if the user attached document D2 to the term "smoke" using the forester perspective. Metadata for the document are created as the user continues to select terms and create attachment relationships.

2.2 Extracting Metadata to Simplify Attachment

Metadata attachment is a time-consuming process, and thorough and accurate results are difficult to achieve. Even users who are extremely familiar with a document's contents can omit important terms that would help others locate the document in the future. Providing controlled vocabularies is an improvement since they prevent the use of non-standard terms and prevent typos. However, navigating the perspectives can still take time and does not help prevent omissions. These problems motivated us to explore more creative solutions.

MetaData eXtract (MDX) is a tool to be used with Metadata++ to augment users' ability to attach thorough and accurate metadata to documents. The following are the requirements such a system should meet.

- *The candidate terms suggested by MDX should be selected from the same semantically rich set that users have defined.*
- *The use of MDX should reduce the time required to attach metadata.* Researchers at the Forestry Service have even observed a phenomenon that they call the '15 minute rule:' if the metadata attachment task takes longer than fifteen minutes, it won't be done!
- *The reasons why MDX suggested a term should be user-accessible.* Users should be able to question the relationship MDX found between a term and a document.

- *MDX should be interactive.* We wish to augment the user's own ability to attach metadata, rather than supplant the user's role as a subject matter expert. This involves being able to explain *why* candidate terms were selected so that they might make informed decisions on which terms they accept or reject.
- *MDX should exhibit reasonable response times.* Our researcher should not be required to upload the document on his way home and check the results in the morning.
- *The number of terms returned should not overwhelm the user.* Our researcher will be no better off if a 200-item list that takes 25 minutes to review is returned. The fifteen-minute rule above will be a good guideline for this requirement as well.

The MDX procedure consists of four tasks designed to help meet these requirements. These tasks are depicted in Figure 3 and are described below.

- *Access the Database:* Currently, our entire database of terms fits in memory, so the first step is performed using straightforward techniques. As our database grows, we will need a more sophisticated approach.
- *Parse the Document:* Documents can be divided into semantic units or structural units. The boundaries of a semantic unit carry meaning with the reader; examples include sentences, paragraphs, and sections. Structural units are physical features of the document that do not carry meaning such as lines, pages, or word sets of a fixed size. We use sentences as our semantic unit for simplicity and to increase the significance of partial matches. A partial match within a short semantic unit is more significant than a partial match within a long semantic unit, so we chose to divide our document into sentences as our initial approach.

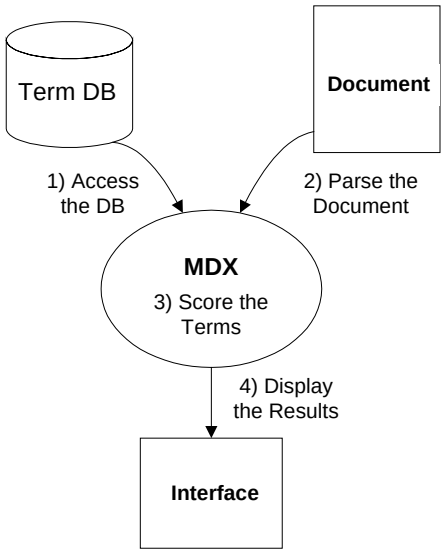

Figure 3: Simplified MDX Architecture

- *Score the Terms:* Scoring the terms consists of measuring the similarity between each term and each sentence. We apply a simplified vector model to calculate similarity. Each term and each sentence is translated into a vector in a common vector space. The similarity of these vectors is measured by the cosine of the angle between them.
- *Display the Results*: Our results are displayed to the user along with the original document context. Presenting the original sentences allows the user to make an informed decision whether or not to include the suggested term.

2.3 Search

Users find documents by specifying and executing a search within the context of a perspective. As with metadata attachment, a user specifies a search by selecting one or more terms within the hierarchy. Each search term is part of a trail (the path traversed by the user from the root node to the search term). The search trail plays an important role in executing the search. The trail is compared with the attachment relationships of document proxies within the system. Those documents that have attachment relationships matching the search trail are included in the results of the search. The user can specify additional matching criteria about each search trail. Each term in the search trail may be specified as mandatory, optional, or wildcard. If the term is designated as mandatory, then it must exist in the attachment relationships used in the results of the search. If a term is designated as optional, then it may or may not exist in the attachment relationships used in the results of the search. A wildcard designation is similar to optional, only wildcard also allows any set of terms to match the wildcard. For example, consider the search trail shown in Figure 5.

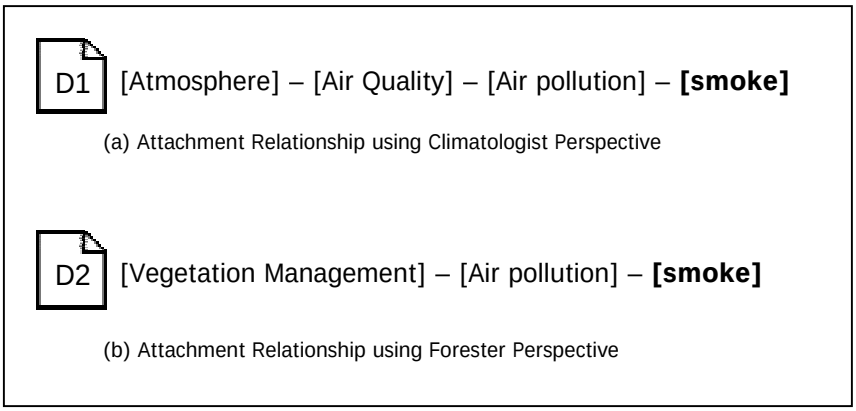

Figure 4: Documents with Attachment Relationships

[*] – [*] – [Air pollution] – **[smoke]**

Figure 5: Search Trail specified in climatologist perspective

In this example, the user has specified a search (within the climatologist perspective) for documents related to "smoke". The term "Air pollution" is designated as mandatory (underlined). This means that the search should only find documents attached to "smoke" where "Air pollution" immediately precedes "smoke" in the attachment relationship. However, terms "Atmosphere" and "Air Quality" have been designated as wildcard (asterisk) – so any terms can precede "Air pollution" in the attachment relationship(s) of the documents returned by the search. The search specified in Figure 5 would return both documents (D1 and D2) shown in Figure 4. Both documents are attached to "smoke" using a trail that satisfies the search criteria.

In addition to matching the specified search trail, Metadata++ exploits hierarchical relationships to find documents related to any descendent term as defined in the current perspective. For example, the Forester perspective (shown in Figure 2) contains a term "Air pollution". This term has two descendent terms, "CO" and "smoke". Suppose Heber specified a search trail of "[Vegetation Management] - [Air pollution]". Metadata++ would find documents attached explicitly to the term "Air pollution", but would also find documents attached to "CO" or "smoke". However, John Taylor may use the climatologist perspective (shown in Figure 1) to specify a search trail of "[Atmosphere] - [Air Quality] – [Air pollution]". In this case, Metadata++ would find documents explicitly attached to "Air pollution" and documents attached to "smoke" or "burning". Unlike Heber's search, however, this search would *not* find documents attached to "CO", because "CO" is not a descendent of "Air pollution" in the climatologist perspective. This example shows how perspectives can be used to customize and exploit a specific user's conceptual understanding of the application domain.

2.4 Metadata Templates

Domain experts often produce documents about similar concepts and places. For example, a large majority of documents produced by a forester, Heber Grant, working in the Wenatchee National Forest would be about different Ranger Districts within that National Forest. To help assist the user in creating metadata, Metadata++ supports Metadata Templates. After a user defines a template, the template can be used for metadata attachment or search. A metadata template consists of one or more trails (a pre-selected path within a specific perspective). For example, consider the perspective that outlines the geographic entities of the USDA Forest Service <show illustration>. Without a template, Heber would browse to "Wenatchee National Forest" and then select the appropriate Ranger District for each document. Instead, Heber browses to "Wenatchee National Forest" (within the forester perspective) and saves that trail in a template. The next time he needs to create metadata for a document, Heber first selects the template. The trail defined in the template pre-selects the "Wenatchee National Forest" term, so Heber can immediately select the appropriate Ranger District and instantiate the attachment relationship – without having to browse through the entire perspective.

When saving a trail as part of a template, the user can select some subset of the trail as the display name. The main benefit of using templates is to avoid repeatedly browsing for commonly used terms. These are often specific terms that appear several levels deep in the hierarchy. Selecting such a term can generate a lengthy trail. In order to make the template more useable and less cluttered, the user may not wish to see the entire trail (i.e. every term along the trail) when viewing the template. In the example above, Heber may not want to see "[Location] - [USDA Forest Service] - [Region 6] - [Wenatchee National Forest]" as the label displayed in the template. Instead, Heber can select some subset of the trail – such as "Location" – as the display name. The template will display only the display name – but will still use the entire trail. So Heber will see "Location", but browse only those terms that are descendants of "Wenatchee National Forest". From a user interface perspective, the trail display names in a template appear as fields or properties – and the user then fills in the corresponding values by selecting an appropriate term.

A user may define a template to match commonly known metadata content standards. For example, Dublin Core [2] metadata may be represented by a template. Each property within the template may reference a term that corresponds to an element of Dublin Core as illustrated below. By using such a template, the user can quickly generate Dublin Core style metadata without forfeiting the benefits of Metadata++. The user is presented with the screen listing all of the properties within the template. The user may then (optionally) augment the paths found in the template by browsing to an appropriate term in the corresponding perspective.

In addition to metadata attachment, templates are useful for searching. The pre-selected paths within a template act as properties in traditional property-value search. Using the template, the user may specify search terms by browsing for the terms within the hierarchy. The system uses the paths to the specified search term(s) to find matching attachment

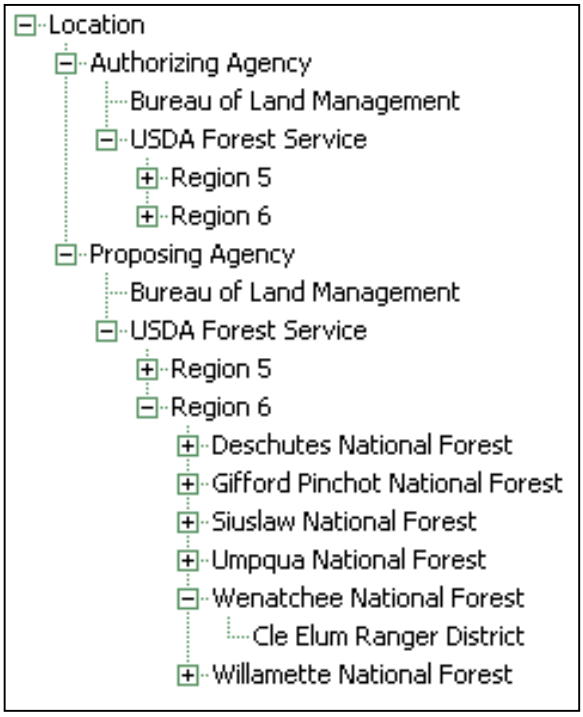

Figure 6: Location perspective (Martin)

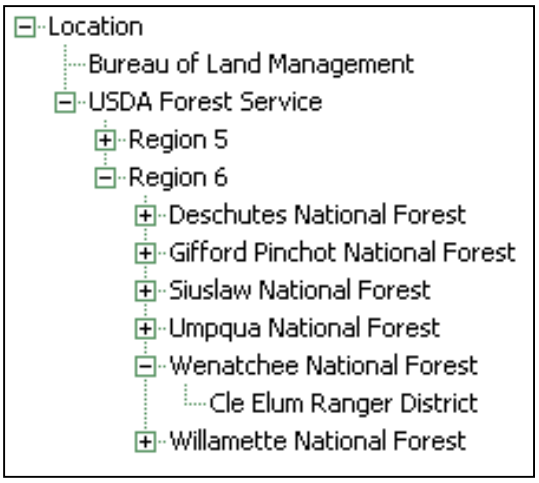

Figure 5: Location Perspective (Heber)

relationships – which are then used to retrieve documents.

3 Scenario

This section describes a brief scenario that illustrates the benefits of using Metadata++. Our forester Heber Grant (from earlier example) works mostly in Wenatchee National Forest and is concerned with documents about locations in that area. Heber may use a location perspective as shown in Figure 5. Suppose Martin Harris works in the regional Forest Service office and deals with documents from all other. In addition, Martin needs to keep track of more detailed metadata. Whenever a document is complete, Martin needs to note the agency that proposed the work (such as a Ranger District) and the agency that authorized the work (such as a National Forest or Region). In order to do this, Martin creates a new location perspective (shown in Figure 6) by adding two additional terms – "Proposing Agency" and "Authorizing Agency." Suppose Martin uses the new perspective to create an attachment relationship, such as:

[Location] – [Proposing Agency] – [USDA Forest Service] – [Region 6] – [Wenatchee National Forest] – [Cle Elum Ranger District]

Even though Heber's perspective does not contain the term "Proposing Agency", he can still find Martin's document by specifying a search such as:

[*] – [Wenatchee National Forest] – [Cle Elum Ranger District]

By using different perspectives, both Heber and Martin can create and use metadata appropriate for their specific tasks, but still share metadata with each other.

4 Conclusion

Metadata++ is a simple and powerful model for capturing and exploiting domain knowledge for document retrieval. Perspectives allow users (or groups of users) to define and use specialized thesauri for creating and searching metadata. Templates provide a convenient mechanism for quickly and easily creating metadata and specifying a search. We have already received positive feedback from potential users. Metadata++ is a unique and useful mechanism for capturing, presenting, and searching metadata.

5 References

1) ANSI/NISO Z39.19 – 1993. *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*. NISO Press, 1994.

2) *Dublin Core Metadata Initiative*. http://www.dublincore.org/