G-Metadata++: Rich Keyword Search Enhanced with a GIS

[1]Len Shapiro, [2]Lois Delcambre, [3]Tim Tolle, [2]Mathew Weaver, [2]David Maier, [4]Dale Guenther, [1]Jason Brewster, [1]Afrem Gutema

[1]Computer Science
Department,
Portland State University
P.O. Box 751
Portland, Oregon 97207
len, jbrew, gutemaa
@cs.pdx.edu

[2]OGI School of Science
& Engineering
Oregon Health & Science
University
20000 NW Walker Road
Beaverton, OR 97006
lmd, mweaver, maier
@cse.ogi.edu

[3]Strategic Planning Office,
[4]Regional Ecosystem Office,
Region 6,
USDA Forest Service
Box 3623,
333 SW First Avenue
Portland, OR 97208
ttolle, dguenther @fs.fed.us

Introduction

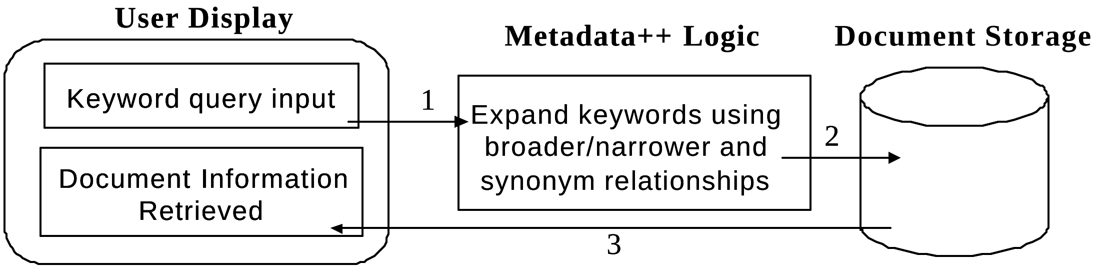
A Geographical Information System (GIS) can be viewed as a software tool that uses associations between metadata (feature, location, and other attribute data) and geographical features, to respond to queries by displaying maps.

We have developed a software tool, Metadata++, that uses associations between metadata and documents, to respond to queries by retrieving documents. For Metadata++ the metadata is keywords and rich associations between them. See Figure 1.

In this abstract we describe our work enhancing Metadata++, by adding a GIS component so that Metadata++ can make sophisticated use of location metadata. We will see that the addition of GIS will enhance each of Metadata++'s query input, processing and output capabilities.

We believe that this project illustrates new ways in which feature-centric systems such as GIS can be used to enhance document-centric systems such as Metadata++.

Figure 1: Metadata++



User Display
Keyword query input
Document Information Retrieved
1
Metadata++ Logic
Expand keywords using broader/narrower and synonym relationships
2
Document Storage
3

Metadata++

Metadata++ is being developed for use by the USDA Forest Service, USDI Bureau of Land Management, and USDI Fish and Wildlife Service to support natural resource managers' needs for document retrieval. We are focusing on "agency approved" documents – that is, documents that are produced by agency personnel as part of their day-to-day work. These documents include such things as Environmental Assessments, Environmental Impact Statements, subject matter expert reports and Decision Notices, as well as more formal reports such as result from research or monitoring.

Our interactions with candidate agency and public users made it clear that they want more than a traditional Boolean expression keyword interface to make a document retrieval system useful. They need richly structured *controlled vocabularies* (CVs) to specify their needs. Each CV consists of keywords related to a particular domain, such as climate, vegetation, wildlife, or hydrology. Metadata then helps a user to formulate queries with keywords from a CV in their domain of interest, thus providing a context to the keywords in that domain. For example, the term *landscape*, in the *ecology* CV, has a scale connotation that differs from the same term used in landscape architecture (*recreation* CV) and both uses differ from the connotations associated with *social forestry*. There is extensive work in progress to identify the domains that are important to natural resource managers, with approximately 30 domains identified so far.

Users also requested that the relationships between keywords be structured, so Metadata++ captures metadata relationships such as broader term/narrower term (e.g., animal/fox) and synonyms (Douglas-fir/ *Pseudosuga Menziesii)*. Metadata++ uses these relationships between CV keywords to retrieve documents that the user may have wanted to see but that may not be associated with the keywords that appear in the user's query. For example, if the user queries with the keyword Douglas-fir, Metadata++ will retrieve documents with the keyword *Pseudosuga Menziesii*. See the middle box in Figure 1.

A third relationship between and among terms, called 'associations,' can greatly enhance browsing based on expert knowledge, even when the individual did not enter the search with that knowledge. For example, one may wish to find documents about serpentine soils and execute a search for such documents. With Metadata++, the knowledge that 'certain ferns are always associated with ultra-mafic soils and that serpentine soil is one such ultra-mafic soil' can be represented. Several documents with those 'fern species' as key words but without 'serpentine' or 'ultramafic soil' as key words exist. The person searching could then be alerted that these documents exist and why that might be important to his/her search.

To summarize, the main functions of Metadata++ are:

- Promote powerful document search – where a user can search for documents by formulating a query with keywords from the CV of interest to them. This is pictured in Figure 1.
- Metadata attachment – where CV keywords represented in Metadata++ can be associated with documents by the user, thus promoting searches for relevant documents.
- Metadata loading – where new CVs and relationships among their keywords can be added to Metadata++, thus making the tool extensible as knowledge grows.

We have constructed prototypes with these features, and these prototypes have been well received by users.

Enhancing Metadata++ with GIS

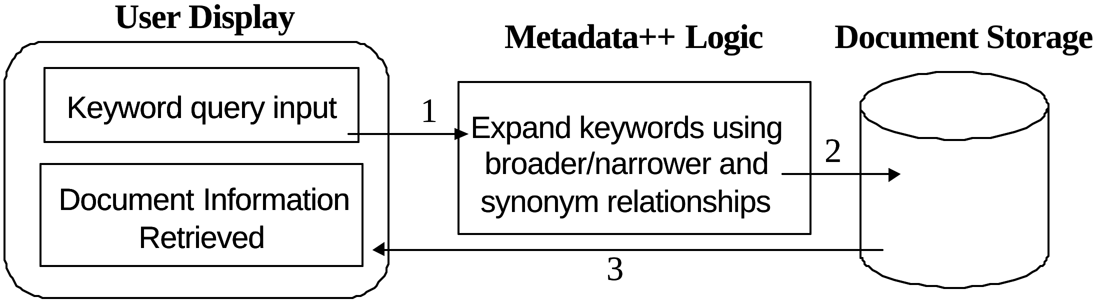
The "location domain" is central to most searches needed by natural resource people. Users need to retrieve documents about specific places of interest, and a GIS can be particularly helpful here. For example, a user planning a vegetation-related project in the Wenatchee National Forest would normally request documents about vegetation in Wenatchee. A GIS can identify all features that are similar to or contained in the Wenatchee National Forest. These related features will correspond to keywords that can retrieve documents of significant interest to the proposed project, documents that would have been missed if only the keyword "Wenatchee" were used in the search. Relationships can be defined differently depending upon the questions, e.g., similar climate and topography or similar sized town located remotely from a large city.

About a dozen schemes for describing location are commonly used, including:

- Political description – the US consists of states, divided into counties, with cities
- Administrative – USDA Forest Service divides the United States into Regions, each Region contains National Forests, each forest is divided into Ranger Districts
- Watersheds – the USGS organizes the world into watersheds, with multiple levels of detail, i.e., watersheds within watershed.

Metadata++ uses a location CV but is not able to do spatial reasoning[1] about keywords for geographic locations. A GIS is ideal for capturing relationships between keywords in the CV of location keywords. For example, a GIS can identify different names for the same geographic location, can determine that one city is in one county, and even that one feature is synonymous (at least within some user-specified parameter, such as 90%) with another. Furthermore, a GIS can display the input and output of queries with respect to their location keywords on maps. We call the enhanced system *G-Metadata*++, pictured in Figure 2, and we describe its architecture below.

Figure 2: G-Metadata++


User Display
Metadata++ Logic
Document Storage
Keyword query input
Document Information Retrieved
1
Expand keywords using broader/narrower and synonym relationships
2
3

GIS as an Input Device in G-Metadata++

Agency users often want to specify their needs in terms of geographic areas. Sometimes those areas do not have names, so the areas cannot be expressed as simple keywords. Furthermore, it is often simpler for a user to identify them visually. Therefore, G-Metadata++ allows users to specify a geographic area and to indicate an interest in documents from that area. In Figure 2, a user can input an area from either predefined choices (within Metadata++) at step 1, or by drawing on a map interface (within the GIS system). G-Metadata++ at step 1' will send that area to the GIS component for expansion into the features it contains. The GIS component will then translate those features into keywords, and then send them at step 2' to Metadata++, which will use them as part of its normal query processing algorithm.

For example, if a user enters "spotted owl" using Metadata++ and identifies a project area by dragging a rectangle in the Northwestern corner of Oregon on a county map of

[1] Metadata++ is able to reason with and thus exploit the broader-term/narrower-term relationship, the synonym relationships, and the general associations that it represents.

Oregon as input, the resulting query might be "retrieve documents related to spotted owls in Clatsop County and in the Coast Range Physiographic Province."

GIS as an Output Device in G-Metadata++

A Metadata++ query may retrieve a huge number of documents. We can help the user visualize the kinds of documents retrieved, by highlighting the features that are contained in retrieved documents, shading the features according to the number of documents corresponding to each.

For example, if the result of the previous query returns 100 documents with keyword Wallowa and 20 documents with Union, then the output would show features Wallowa and Union highlighted, with Wallowa shaded more heavily. This might help the user to reformulate the query into a query focusing on a more manageable set of documents.

GIS as a Relationship Server in G-Metadata++

We have noted above that Metadata++ can retrieve documents keyed under scientific names as well as common names, using its synonym relationship. Similarly, if a user identifies the Wilson River Watershed as a keyword (in the text interface, not on a GIS input screen) search term to a query, the GIS component may recognize that the Wilson watershed and Tillamook County are 90% identical, and add Tillamook County to the keyword search terms. The 90% is a parameter that can be set by a user depending on the precision desired, and reset depending on the number of documents retrieved.

Conclusion: Could this all be done in a GIS?

It is tempting to consider whether all of the functionality of G-Metadata++ could be accomplished within a GIS. We feel that this is not the case because GIS is "feature"-centric while Metadata++ is document-centric. The relationships of synonym, broader/narrower and association are well supported by a GIS when the keywords to be related are geographic features. However, a GIS is not well suited to manage relationships among non-geographic terms, and such searches are often needed, e.g., searches about species, habitats or activities for background to a research proposal. Metadata++ allows keywords to be associated with documents and for keywords to be directly related. Although many of the documents of interest to natural resource managers have location(s) associated with them, some documents do not. The focus of our work is to use both technologies to our advantage. We are using GIS to manage and reason about location metadata and we are using Metadata++ to track documents and managing non-location metadata, i.e., metadata in the remaining twenty-nine domains.

Acknowledgements

This work is supported in part by the National Science Foundation, grant number EIA 9983518. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.