# Issues in Georeferenced Digital Libraries

Greg Janée
Alexandria Digital Library Project
University of California at Santa Barbara
gjanee@alexandria.ucsb.edu

James Frew
Bren School of Environmental Science and Management
University of California at Santa Barbara
frew@bren.ucsb.edu

Linda L. Hill
Alexandria Digital Library Project
University of California at Santa Barbara
lhill@alexandria.ucsb.edu

## Abstract

Based on a decade's experience with the Alexandria Digital Library Project, seven issues are presented that arise in creating georeferenced digital libraries, and that appear to be intrinsic to the problem of creating any library-like information system that operates on georeferenced and geospatial resources. The first and foremost issue is providing **discovery** of georeferenced resources. Related to discovery are the issues of **gazetteer integration** and specialized **ranking** of search results. **Strong data typing** and **scalability** are implementation issues. Providing **spatial context** is a critical user interface issue. Finally, sophisticated **resource access** mechanisms are necessary to operate on geospatial resources.

## Introduction

In creating a digital library of any kind, issues that must be addressed—problems to be surmounted, special features that must be provided, considerations to be taken into account—inevitably arise due to the specific characteristics of the resources being stored in the library and the functionality being provided over those resources. This article is concerned with the issues that arise when creating a digital library of georeferenced resources.

A *georeferenced resource* is a granule of information that is relevant to an identifiable subset of the Earth's surface; we call the referenced subset the resource's *spatial coverage*. Any kind of document that is relevant to a particular geographic place (e.g., Dickens' *A Tale of Two Cities*, Gibbon's *History of the Decline and Fall of the Roman Empire*) is an example of a georeferenced resource. There is a large subclass of georeferenced resources that includes maps, aerial photography, and remote-sensing imagery, whose content is distributed over the extent of its spatial coverage. We call

resources in this subclass *geospatial resources* to emphasize their spatial character. A *georeferenced digital library* is an information system that stores georeferenced resources, and moreover provides a spatial orientation to those resources in terms of discovery, browsing, viewing, and access.

In 1994 the Alexandria Digital Library (ADL) Project [1] began creating a georeferenced digital library that would reproduce and extend the content and functionality of a traditional map library, specifically, the Map & Imagery Laboratory at the University of California, Santa Barbara. Starting with this objective, the Project has developed four successive library architectures:

- a "rapid prototype" system [2], comprising a relational database of map and imagery metadata, accessed through a desktop geographic information system (GIS);
- a "web prototype" system [3], which replaced the standalone GIS with an HTTP server, generating an HTML forms-based user interface accessible via the World Wide Web;
- the "ADL-3" system [4], which extended the HTTP server into full-fledged middleware, supporting HTTP interfaces to multiple clients, and connections to multiple catalog databases; and currently
- the Alexandria Digital Earth ProtoType (ADEPT) system [5], second-generation middleware that functions as a generic framework for managing, querying, and accessing distributed, heterogeneous georeferenced resources and other types of structured scientific data.

As these systems have been developed, used, and evaluated, we have noticed that, regardless of the architecture or technological approach, the same issues have arisen. We believe these issues to be intrinsic to the problem of creating any library-like information system that stores and operates on georeferenced and geospatial resources.

The remainder of this article discusses seven issues. The first and foremost issue is providing **discovery** of georeferenced resources. Related to discovery are the issues of **gazetteer integration** and specialized **ranking** of search results. **Strong data typing** and **scalability** are implementation issues. Providing **spatial context** is a critical user interface issue. Finally, sophisticated **resource access** mechanisms are necessary to operate on geospatial resources.

## The issues

### 1. Discovery

The first and most basic issue is: how can georeferenced resources be discovered? The traditional approach to discovery in digital libraries is to use text-based information retrieval techniques on the metadata associated with the resources. Applied to georeferenced resources, this would mean basing discovery on matching of textual placenames. Using this technique, a user desiring a map of *Boulder, Colorado,* would use

this phrase as a query, and the search service would return resources whose metadata "matched" (partially or exactly) the query phrase.

This technique works well for certain classes of information and in certain contexts. Clearly, such a technique would return a resource whose title is "Downtown street map of Boulder, Colorado." As a general technique, however, it suffers from two serious drawbacks. First, there is a whole class of georeferenced resources having *no* associated placenames, namely, data gathered from moving sensors. Examples include satellite imagery and aerial photography, which carry only source metadata such as the coordinates and attitude of the camera or sensor at the time of exposure.

Second, discovery based on textual placename matching is unreliable in the large. Consider for a moment a user desiring a map of the *Flatirons* rock formations in the *Front Range* of the *Rocky Mountains* just outside of *Boulder*, *Arapahoe County*, *Colorado*. The highlighted words in the previous sentence all describe the area of interest with varying degrees of specificity, but it turns out that the most specific phrase (and the only *useful* phrase) that will retrieve data from the United States Geological Survey (USGS) for this area is *Eldorado Springs* (!), which happens to be the name of the topographic map quadrangle in the USGS's standardized grid system that covers the Flatirons. The USGS picks one prominent feature within each quadrangle to serve as the name for that quadrangle. If the USGS had included *Flatirons* in its cataloging, our hypothetical problem would be solved, but the USGS can't be faulted, for there is no easy or manageable way to associate every possible placename with every library resource. This is but one instance of a very general problem, namely, that there are *many* names—administrative, physiographic, technical (e.g., telephone and postal codes), temporal (e.g., earthquake and hurricane names)—for any given place, and a resource that is about a place can be cataloged by, at best, a handful of those names.

To address the general problem of reliable georeferenced discovery, some kind of structured (non-textual) search technique must be employed. Defining a controlled vocabulary of places (so that, for example, "Boulder, Colorado" would be treated not as a text string but as one of a set of discrete terms) is one structured approach often employed by libraries. This approach resolves the placename multiplicity problem to a certain extent (assuming a user interface to the vocabulary can guide the user in selecting the proper term), but it places a considerable burden on both library catalogers and users to understand, agree upon, and use the vocabulary. Gaining agreement on a vocabulary is a classic problem; it requires designating one among many possible names as the preferred name for a place, which is similar to the USGS choice of one placename to represent a quadrangle, but it still leaves the problems of anticipating all of the other names that users will associate with the place and adding them as alternative or related names.

A more general and powerful technique is to support range searching (i.e., searching that employs relational operators such as BETWEEN and OVERLAPS) over a continuous domain such as numeric latitude/longitude coordinates. This technique places a large burden on libraries and users to express spatial coverages and query regions in terms of coordinates. But the advantage is that it allows a user to discover information without the user and library having to agree on anything except the coordinate system.

## 2. Gazetteer integration

We argue that structured discovery techniques are required to implement georeferenced discovery that is reliable in the large, and that discovery based on textual placename matching can be problematic. Nonetheless, placenames play a critical role in georeferenced discovery, and in georeferenced digital libraries in general, because human spatial cognition relies on relationships to and among known features, and to the extent that those features are named, to placenames [6]. As proof of this, consider how you would answer the question, "Where are you right now?" Would you answer in terms of a coordinate system ("I'm at 35º N…"), or in terms of relationships to symbolically-named features ("I'm at the intersection of Hollywood and Vine…")?

The answer is almost always the latter, and consequently a *gazetteer* (a dictionary that supports translation between placenames and coordinates [7], [8]) must be integrated into a georeferenced digital library to help translate between the user's mental model of geographical space and the library's discovery system.

Ideally, the gazetteer would be so integrated into the library that the translation from user queries to the library's discovery system would be hidden and seamless: the user enters "Flatirons" and the library automatically performs an appropriate coordinate-based search and returns data for the Eldorado Springs quadrangle. Unfortunately, at the time of this writing, publicly available gazetteer data is able to support this level of functionality only to a very limited degree, because most of it has been created by national- and state-level toponymic authorities whose primary purpose is to disambiguate one named place from another. For this purpose, general point locations for places are sufficient. Additional problems with automated use of gazetteer lookups include incomplete data and non-uniqueness of names.

But while seamless gazetteer integration may remain elusive, *human-mediated* gazetteer integration is effective, and easily achieved by treating gazetteer lookup services as an extension of the library's user interface. In this model, the user enters "Flatirons" and interacts with a gazetteer and possibly a map to locate the desired place (in case there are multiple Flatirons, as there almost always are). From there the library can form an appropriate coordinate-based spatial query region.

## 3. Ranking

Another issue related to discovery of georeferenced resources is the need for specialized ranking methods. Text-based ranking methods, such as those based on word frequencies in documents, do not apply to georeferenced discovery. Yet some form of ranking is required with georeferenced discovery to avoid the common problem of an atlas of the world (with its entire-world spatial coverage) being returned as the first answer to every query.

One method the authors are investigating is ranking library resources based on the spatial similarity of their spatial coverages to the user's query region; *spatial similarity* of two geographic regions being a function of the regions' sizes, shapes, and locations [9]. Thus, given the state of Colorado as a query region, spatial similarity would rank highest those

resources that approximately match Colorado as a whole (e.g., a map of the state of Colorado, a Colorado statewide dataset); would rank lower a resource such as a map of the southwest United States; and would rank lowest both very large (e.g., a world atlas) and very small (a city street map) resources.

More sophisticated ranking methods are certainly possible [10]. Regardless of the particular ranking scheme employed, a georeferenced digital library will require some form of spatial ranking if it contains a large number of resources whose spatial coverages differ in extent by several orders of magnitude or more.

**4. Strong data typing**

Many, if not most, digital library implementations are text-based in that the functionality they provide related to metadata, discovery, and ranking is based on finding and counting occurrences of words, whether in resource metadata or in the resources themselves. If we think in terms of data typing in programming languages, then *text* in these digital libraries is effectively an untyped quantity. Although in human terms text is highly structured and laden with semantics, in information systems text is often treated as a simple stream of bytes; that is, *any* stream of bytes is interpretable as text, with words being simply delimited sequences of bytes within the stream. A consequence of this simplified view of text is that text-based discovery systems can be (though are not necessarily) relatively lightweight, easy to implement, scalable, and, significant for the purposes of this discussion, extremely forgiving. Digital libraries can generally operate on *any* text (whether or not the text is syntactically or grammatically correct or comes from well-formed markup), and their basic text operations can be applied to arbitrary aggregations of text (from a single metadata field, to all of a resource's metadata, to all metadata in the library) with equal ease.

Unfortunately, these desirable characteristics of text largely fade away when a structured discovery technique, such as range searching over continuous geographic coordinates, is added to a digital library. Such a technique requires that the library support describing and operating on *geographic regions*. In the simplest case, the library may support only box-shaped regions, but the library may also want to support other kinds of regions such as polygons, polylines for describing linear features, and region collections for describing non-contiguous regions. In any case, geographic regions are necessarily a non-textual data type and, in programming language terms, require strong data typing. Some of the ramifications of such an addition are:

- Input validation is required. With text, a digital library can take advantage of any structure it recognizes (metadata fields and other markup, for example), or it can fall back to a base level interpretation of text as a byte stream. With geographic metadata (descriptions of the spatial coverages of library items, for example) there is no such fallback: either the metadata is recognized and in a usable form, or it is not.

- Relatively complex internal structures and external representations are required to describe spatial coverages and geographic query regions, particularly if higher-order kinds of regions such as polygons are supported.

- Syntactic and semantic heterogeneity of external representations becomes an issue. There are multiple kinds of geographic regions, multiple ways of describing regions, and multiple applicable metadata standards in use. Even for the simplest kind of geographic region, boxes, there are both syntactic and semantic differences between the FGDC [11], GML [12], Dublin Core [13], and ADL [14], [15] metadata formats. In addition, there are the classical cartographic problems of projection and datum conversion [16]. A digital library that strives to accommodate this heterogeneity will have to provide mapping mechanisms at some level.

- The library must provide the means to express and execute different types of query constraints (spatial, textual, etc.) as well as Boolean combinations of different types of constraints. The latter are particularly valuable, but can be particularly difficult to implement, as discussed in the next section.

- Specialized user interface components are required to input and to view any kind of strongly typed data. In the case of geographic coordinates, this means integrating an interactive map browser into the library's user interface, which in turn requires underlying map and gazetteer servers.

As discussed in the previous section, geographic discovery may require specialized ranking mechanisms. In addition, the library may have to provide mechanisms for combining different types of rankings such as spatial and textual rankings.

The overall implication of these strong data typing issues is that a digital library that supports a structured discovery technique such as coordinate-based geographic discovery will require an implementation that is substantially different from a digital library that supports textual discovery only. Once a decision has been made to support structured discovery, a second decision must be made as to what kinds of geographic regions will be supported. Supporting boxes is fairly straightforward; supporting higher-order kinds of regions, polygons in particular, requires a significant increase in implementation complexity and difficulty.

## 5. Scalability

It is very easy to accumulate geospatial data since it is often generated by automatic means. Many Earth sensors generate periodic reports; a single flight of an aerial survey can result in hundreds and even thousands of images; and satellite-based sensors such as MODIS [17] operate continuously, producing multiple terabytes of data per day. In addition, the size of the Earth is quite large relative to human scale and the scales on which most phenomena are studied. The USGS's 7.5-minute grid over the conterminous United States, by which it catalogs its 1:24,000-scale products, contains over 55,000 grid cells; extended to the entire Earth, it would result in over 4 million grid cells. Some geospatial data products such as DOQQs [18] have 1-meter resolution or less; compare this to the Earth's surface area of 500 trillion square meters. Thus scalability in terms of accommodating large numbers of resources can quickly become an issue for georeferenced digital libraries that have geospatial data in their collections.

Georeferenced discovery techniques are nicely scalable in theory (the multidimensional index structures in common use today [19] offer logarithmic search time), but in practice, scalability is more limited. At the time of this writing, commercial georeferenced search engines reach the limits of practical use (say, one day of compute time utilizing several gigabytes of RAM and unlimited disk space) when indexing as few as one million items. More items can be accommodated, of course, but only with custom engineering and exponentially increasing expense. Microsoft's TerraServer [20] provides a nice example of what can be accomplished with custom engineering, but it also demonstrates just how much engineering is required to handle large amounts of geospatial data.

A more significant problem is the difficulty of *joining* (in the relational database sense [21]) Boolean combinations of spatial constraints with textual and other types of constraints. A typical and reasonable gazetteer query such as "find a city named Boulder near the Rocky Mountains" is a conjunction of a discrete constraint (find a city, where "city" is one of a set of discrete types), a textual constraint (named "Boulder"), and a spatial constraint (near the Rocky Mountains). When large numbers of items are involved, generating efficient query plans becomes paramount, but query plans are unavoidably sensitive to the queries and to the distribution of the data, and the data is inevitably not uniformly distributed. Whether the library is working within the framework of a single relational database or across multiple, distributed search indexes, the heterogeneous join problem is a significant practical problem, and a research problem as well [22].

## 6. Spatial context

To use any information space (such as a digital library) effectively, users need context to understand where they are in that space, and how the resources they are currently viewing or operating on relate to the information space and to other resources in the space. In the case of georeferenced digital libraries, the relevant context is geographic, i.e., the context is the Earth's surface and recognizable landmarks on that surface.

Geographic context is critical in formulating georeferenced queries. Users typically want to express georeferenced queries by making symbolic references to known landmarks ("Boulder"), by expressing spatial relationships to those landmarks ("east of Boulder and south of the Flatirons"), or by graphically indicating a region of interest on a context-providing background map. Thus georeferenced digital libraries that support coordinate-based discovery will need to provide an interactive background map that is integrated with both the library's user interface and a gazetteer lookup service, and that has sufficient detail to allow users to relate the query region to the landmarks they recognize. Systems that require the user to input numeric geographic coordinates are generally impossible to use without some kind of external, context-providing device.

Context is also critical in interpreting and evaluating both query result sets and individual library resources. Suppose for a moment that we have gotten past the first spatial query that everybody does ("let's see if I can find my house"), an exercise that is highly misleading because we are intimately familiar with our own surroundings. Instead, suppose that the aerial photograph in Figure 1 has been returned as a response to the previous query about the Flatirons rock formations near Boulder, Colorado, an area with which we may not be familiar. Does this give us the information we want? More to the point, where *are* the Flatirons in this image?


**Figure 1. An image that shows… what?**

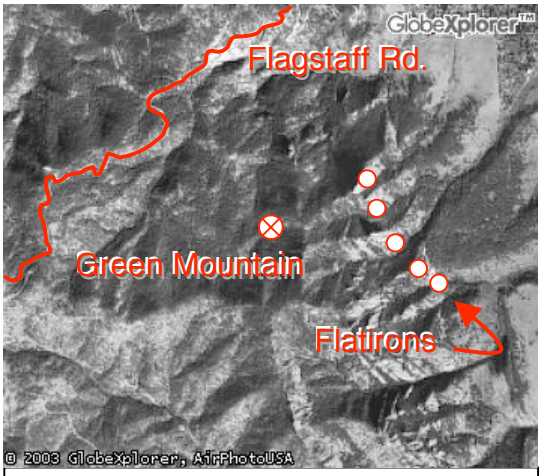


**Figure 2. Same image, with context.**

Adding context to the image by layering map and placename data over it, as demonstrated in Figure 2, makes it possible for the user to (1) interpret the resource and evaluate it for relevance to the query, and then (2) utilize it. Such context-providing functionality has long been available in geographic information systems (GIS), but as of this writing GIS applications are relatively expensive and uncommon, and require a high level of sophistication on the part of the user. To be useful to the general user, we argue that a fully developed georeferenced digital library must provide a certain amount of its own (lightweight) GIS functionality. At a minimum, a digital library must make it possible for users to evaluate georeferenced resources, and particularly geospatial resources, for query relevance.

### 7. Resource access

Geospatial resources often have complex structure and can be quite large. A resource can consist of multiple parts, require specialized viewing tools, and be accessible via multiple formats, protocols, and interfaces. For example, geospatial images are often distributed as pairs of files: an image proper (say, a TIFF file) together with a text file containing the georeferencing information. A library client may have to access either of these files independently or both files to use the image successfully. To take another example, the industry-standard ESRI "shapefile" format is in fact a set of 4–7 coordinated files of different types [23]. Geospatial resources are also often accessible via programmatic services (or protocols) such as the OpenGIS Web Coverage Service (WCS) [24] and OPeNDAP [25] protocols. Access via a service becomes more critical as the size of the

geospatial resource increases, because service-based access typically allows navigation to and interaction with just a desired subset of the resource. For example, both the WCS and OPeNDAP protocols support server-side cropping, subsampling, and layer selection of a resource, valuable capabilities when individual resources can be many gigabytes in size.

Geospatial resources are also closely tied to geographic information systems and other types of data exploration environments. In some situations, GIS tools are simply a common means of viewing and working with such resources; in others, the resources are entirely unusable outside the appropriate analysis or visualization tool. Either way, the ability to process geospatial information strictly within the confines of a standard Web browser is limited.

These characteristics of geospatial resources—their complex structure and close ties to analysis environments—call out the need for georeferenced digital libraries to be able to describe, in detail and for the benefit of both programmatic clients and human users, the components that make up a library resource and the different modes by which it may be accessed. It is insufficient to describe the content of a geospatial resource by, say, a single, unadorned URL ("click here"). A more detailed description allows the user to make an informed decision about the best way to access the resource, and can facilitate a seamless handoff of the resource from the library to a more capable analysis or visualization tool.

## Conclusion

Seven issues have been presented that arise in the creation of digital libraries and other information systems that operate on georeferenced and geospatial resources.

The first and foremost issue is providing **discovery** of georeferenced resources. To provide a form of georeferenced discovery that works reliably in the large, we have found that a structured discovery technique is required, e.g., range searching over latitude/longitude coordinates. Implementing such coordinate-based search immediately brings up two other issues: the need for **strong data typing**, to support description of and operation on geographic regions, and **gazetteer integration**, to support mapping between placenames and coordinates. Another issue related to discovery is the need for a specialized **ranking** method that returns the most geographically relevant resources first. Geospatial resources bring up their own issues, including provision for **scalability**, since geospatial resources are often created automatically and in large numbers, and the need for **resource access** mechanisms that make it possible to operate on those resources. There are a number of user interface implications to the above issues; another user interface issue is the need for **spatial context**, to allow users to formulate georeferenced queries and interpret and evaluate geospatial resources.

## References

[1] Alexandria Digital Library homepage, http://www.alexandria.ucsb.edu/.

[2] James Frew, Larry Carver, Chistoph Fischer, Michael Goodchild, Mary Larsgaard, Terence Smith, and Qi Zheng. The Alexandria Rapid Prototype: building a digital library for spatial information. *Proceedings of the 1995 ESRI International User Conference* (Palm Springs, California; May 22-26, 1995).
[3] James Frew, Michael Freeston, Randall B. Kemp, Jason Simpson, Terence Smith, Alex Wells, and Qi Zheng. The Alexandria Digital Library Testbed. *D-Lib Magazine* **2**(7/8) (July/August 1996). Available at doi: 10.1045/july96-frew.
[4] James Frew, Michael Freeston, Nathan Freitas, Linda Hill, Greg Janée, Kevin Lovette, Robert Nideffer, Terence Smith, and Qi Zheng. The Alexandria Digital Library Architecture. *International Journal on Digital Libraries (IJODL)* **2**(4) (July 2000): 259–268.
[5] Greg Janée and James Frew. The ADEPT Digital Library Architecture. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (Portland, Oregon; July 14-18, 2002): 342–350.
[6] Rob Kitchin and Mark Blades. *The Cognition of Geographic Space*. London: I.B. Tauris, 2002.
[7] Linda L. Hill, James Frew, and Qi Zheng. Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. *D-Lib Magazine* **5**(1) (January 1999). Available at doi: 10.1045/january99-hill.
[8] Greg Janée and Linda L. Hill. The ADL Gazetteer Protocol. 2001. Available at: http://www.alexandria.ucsb.edu/gazetteer/protocol/specification.html.
[9] Greg Janée. Spatial Similarity Functions. 2003. Available at: http://www.alexandria.ucsb.edu/~gjanee/archive/2003/similarity.html.
[10] Marc van Kreveld, Iris Reinbacher, Avi Arampatzis, and Roelof van Zwol. Distributed Ranking Methods for Geographic Information Retrieval. *Proceedings of the Twentieth European Workshop on Computational Geometry* (Seville, Spain; March 24-26, 2004) (to appear).
[11] Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata. 1998. Available at http://www.fgdc.gov/metadata/contstan.html.
[12] Open GIS Consortium (OGC), Inc. OGC 02-023r4. OpenGIS Geography Markup Language (GML) Implementation Specification. 2003. Available at: http://www.opengis.org/docs/02-023r4.pdf.
[13] Simon Cox. DCMI Box Encoding Scheme: specification of the spatial limits of a place, and methods for encoding this in a text string. 2000. Available at http://dublincore.org/documents/dcmi-box/.
[14] XML document type definition for the ADL query language, http://www.alexandria.ucsb.edu/middleware/dtds/ADL-query.dtd.
[15] XML document type definition for ADL bucket reports, http://www.alexandria.ucsb.edu/middleware/dtds/ADL-bucket-report.dtd.
[16] John P. Snyder. *Map Projections— A Working Manual*. U.S. Geological Survey professional paper 1395. Washington, D.C.: USGPO, 1987.
[17] Moderate Resolution Imaging Spectroradiometer. See http://modis.gsfc.nasa.gov/.
[18] Digital Orthophoto Quarter-Quadrangle. See http://www.usgsquads.com/prod_doqq.htm.

[19] Volker Gaede and Oliver Günther. Multidimensional Access Methods. *ACM Computing Surveys* **30**(2) (June 1998): 170–231.
[20] Tom Barclay, Jim Gray, and Don Slutz. Microsoft TerraServer: A Spatial Data Warehouse. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (Dallas, Texas; May 16-18, 2000): 307–318.
[21] Jeffrey D. Ullman and Jennifer Widom. *A First Course in Database Systems* (second edition). Upper Saddle River, New Jersey: Prentice-Hall, 2002.
[22] Laura M. Haas, Donald Kossmann, Edward L. Wimmers, and Jun Yang. Optimizing Queries Across Diverse Data Sources. *Proceedings of the Twenty-Third International Conference on Very Large Databases (VLDB)* (Athens, Greece; August 25-29, 1997): 276–285.
[23] Environmental Systems Research Institute (ESRI), Inc. ESRI Shapefile Technical Description. 1998. Available at: http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf.
[24] Open GIS Consortium (OGC), Inc. OGC 03-065r6. Web Coverage Service (WCS), Version 1.0.0. 2003. Available at: http://www.opengis.org/docs/03-065r6.pdf.
[25] Open-source Project for a Network Data Access Protocol. See: http://opendap.org/.