Content Access Characterization in Digital Libraries

Greg Janée
Dept. of Computer Science
U. of Calif., Santa Barbara
gjanee@alexandria.ucsb.edu

James Frew
School of Environmental Science and Management
U. of Calif., Santa Barbara
frew@bren.ucsb.edu

David Valentine
Davidson Library
U. of Calif., Santa Barbara
valentine@library.ucsb.edu

Abstract

To support non-trivial clients, such as data exploration and analysis environments, digital libraries must be able to describe the access modes that their contents support. We present a simple scheme that distinguishes four content accessibility classes: ***download*** *(byte-stream retrieval),* ***service*** *(API),* ***web interface*** *(interactive), and* ***offline****. These access modes may recursively nest in alternative (semantically equivalent) or multipart (component) hierarchies. This scheme is simple enough to be easily supported by DL content providers, yet rich enough to allow programmatic clients to automatically identify appropriate access point(s).*

1. Background

The Alexandria Digital Library (ADL) Project has created a distributed DL architecture [1] that allows data providers to publish geospatial and other types of highly structured, metadata-rich data. A key feature of this architecture is a middleware layer that allows clients to perform federated collection- and item-level searches over multiple libraries. The middleware uses XML-based communication, and supports HTTP, Java and Java RMI interfaces that make all library functionality available to programmatic clients.

Our first ADL clients focused on discovery of library content – access to and use of library content were only superficially addressed. Users wishing to access a DL item were presented with a "click here to download" hyperlink and left to their own devices. A better approach is to exploit the programmatic nature of ADL's library services by embedding ADL in several data exploration environments, and making discovery and use of library content seamless within those environments.

This effort has brought up several access-related issues. First, any data exploration environment inevitably has constraints on the content formats and protocols it understands. If a corresponding DL doesn't support format- and/or protocol-related query constraints, then it must at least be possible to filter query results by these kinds of criteria.

Second, a data exploration environment needs to understand the structure of an item and its available access options. This is particularly necessary for geospatial data, which is often multipart and typically accessed via multiple formats, protocols, and interfaces. For example, a georeferenced image is often distributed as a pair of files: an image array and a file containing the georeferencing information. A client may have to independently access one or both of these files to successfully use the item. Geospatial data is also often accessible via programmatic services such as WMS, ArcIMS, and DODS.

Note that the issue here is not so much describing the content as it is describing *access* to the content. We wish to semi-formally characterize the methods by which DL items may be accessed. We're particularly trying to capture basic distinctions like: is the item online or offline? If online, can it be directly downloaded? If so, what is the expected format and size? Are there programmatic services associated with the item? If so, how does one interact with them? Is the item decomposable into constituent parts? Are there alternative representations? All this information must be usable by programmatic clients, as well as by human users deciding how to best access the item.

2. Related Work

Metadata standards generally focus on descriptive metadata such as author and title, not structural metadata. The FGDC Content Standard for Digital Geospatial Metadata [2] allows an item to have alternative typed "distributions", but there is no support for describing a multipart item's structure or item-related services. Dublin Core [3] provides essentially no means of describing structure or access. A DC "best practice" is to use URLs as identifiers, but DC has no standard way to describe what such a URL actually refers to.

The Metadata Encoding and Transmission Standard (METS) [4] allows elaborate descriptions of the structure of scanned or transcribed written works. However, support for other content types is limited; more

significantly, METS has no simple way to distinguish between components and alternative representations.

Fedora [5] provides an object-oriented framework for describing and invoking item-related services, but its structural metadata is very limited. The Fedora project is exploring integrating Fedora access into METS.

3. The ADL Content Access Model

We describe a DL item's accessibility via zero or more **access points**. Each access point describes how to access a single, independent representation of the DL item. Different types of access points reflect fundamentally different modes of accessing content.

A **download** access point simply returns a byte-stream representation of an entire DL item. A typical download access point is a static file made accessible via HTTP. A download access point's attributes include a URL and (optionally) a high-level format description, a MIME type and encoding(s), and an approximate length. (All access points have optional, human-readable title and description attributes.)

A **service** access point allows a DL item to be accessed by interacting with a programmatic service. A typical service access point is an OpenGIS Web Map Service (WMS) [7] returning selected portions of a map. A service access point's attributes include the service's URL and (optionally) the name of the service's protocol and a pointer to a formal (e.g., WSDL) or informal (e.g., English) description of the service.

A **web interface** access point is a URL that may require additional human interaction before access is granted. (Or, more information about the access point may simply be unknown.) A typical web access point is a license agreement that in turn points to the actual content. The only attribute of a web interface access point is a URL.

An **offline** access point refers to an offline representation (digital or physical) of the DL item. A typical offline access point is a library call number.

Access points may be recursively grouped into hierarchies. An **alternatives** access point describes two or more equivalent representations (for example, HTML and PDF versions of the same document). A **multipart** access point describes constituent parts (e.g., a TIFF image and its accompanying "world" file). A multipart access point's optional format attribute describes the format of the access point as a whole (e.g., a "zip" archive).

From a programmatic client's perspective, the most useful access points are download and service. The web interface access point exists to cover those situations where an item is online but cannot be accessed programmatically, either because human intervention is required, or because the interface is not well specified. Thus a web interface access point represents a kind of fallback with respect to the other types of access points.

The ADL content access model is formally described (and documented with examples) by an XML DTD [6].

4. Conclusion

We have presented a simple yet effective way to characterize and describe DL content access. Programmatic DL clients can automatically access recognized content formats and protocols, while human readers can make informed choices between alternative download and interactive service options. We are using this model to interface two programmatic clients – the ESRI ArcGIS geographic information system and a Distributed Oceanographic Data System (DODS) data viewer – to the ADL library. Our access model is a good match for these environments.

One outstanding issue is the interaction, if any, between content access and authorization and rights schemes. A DL item whose access points are all hidden behind authorization pages or license agreements is currently forced to describe those access points as web interfaces, not as more programmatically useful downloads and/or services. An improvement might allow authorization to be an attribute of an access point.

We would also like to extend our scheme to describe and automatically invoke conversion services. This would allow a client to access content outside its natively supported domain.

5. References

[1] Janée, G. and Frew, J. The ADEPT digital library architecture. Second ACM/IEEE-CS Joint Conference on Digital Libraries (Portland OR, June 2002). ACM Press, 342-350.

[2] Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C. http://www.fgdc.gov/metadata/csdgm/

[3] Weibel, S., Kunze, J., and Lagoze, C. Dublin Core Metadata for Resource Discovery. RFC 2413 (September 1998), Internet Engineering Task Force. http://www.ietf.org/rfc/rfc2413.txt

[4] Library of Congress. Metadata Encoding and Transmission Standard (METS). http://www.loc.gov/standards/mets/

[5] Fedora Project. Mellon Fedora Technical Specification Version 1.1 (December 2002) http://www.fedora.info/documents/master-spec.rtf

[6] http://www.alexandria.ucsb.edu/middleware/dtds/ADL-access-report.dtd

[7] de La Beaujardière, J. (ed.) Web Map Service Implementation Specification. OpenGIS Implementation Specification OGC 01-068r3. Open GIS Consortium Inc. http://www.opengis.org/techno/specs/01-068r3.pdf