Search Service Interoperability

Committee Draft, 26 January 2004

Document identifier:

wd-egov-searchservice-CD

Editor:

Eliot Christian, U.S. Geological Survey <echristi@usgs.gov>

Abstract:

Governments are recommended to enhance interoperability among their networked systems by adopting a common search service. The search service should be based on the ISO 23950 international standard that features a high degree of interoperability across many communities of practice and types of data and information holdings. Governments should implement the search service as a supplement to other search mechanisms, as these may be required for reasons other than broad scale interoperability.

Status:

This Committee Draft has been approved by the e-Gov Technical Committee.

Committee members can send comments on this specification to the egov@lists.oasis-open.org list. Others should subscribe to and send comments to the egov-comment@lists.oasis-open.org list. To subscribe, send an email message to egov-comment-request@lists.oasis-open.org with the word "subscribe" as the body of the message.

For information on whether any patents have been disclosed that may be essential to implementing this specification, and any offers of patent licensing terms, please refer to the Intellectual Property Rights section of the E-Government TC web page (http://www.oasis-open.org/committees/egov/ipr.php).

Table of Contents

1 Requirements

1.1 Government Stake in Search Interoperability

Governments at all levels worldwide are major producers and consumers of data and information. Governments and the publics they serve have long invested heavily in enhancing the discovery and use of government data and information resources, thereby serving goals such as government transparency and accountability, efficiency of commerce, education, scientific research, and a range of other societal objectives. As the Internet becomes ever more essential to the dissemination of data and information resources held by governments, interoperability of information search mechanisms is a major issue.

Broad scale, standards-based interoperability is especially critical for governments in that they must depend on and foster a competitive intermediary market for information dissemination and service delivery. Governments must offer to intermediaries an information search interface that is non-proprietary, fair, and stable with respect to clearly defined processes and technical standards. By the choice of open standards, governments encourage competition and maximize customer choice.

1.2 Search Service Defined

Current technology is continuing its evolution toward modularization of complex systems into components that interoperate primarily through the passing of structured messages at interfaces designed for networking. Each set of operations available at a component network interface is defined as a “service”. This overall approach to interoperability is known as a component-based, service-oriented architecture.

In a component-based, service-oriented architecture, interoperability of search implies the definition of a common search service. The broad scale of government interoperability requires that this search service be based on widely implemented international standards, and that it supports a high degree of interoperability across many communities of practice and types of information holdings. Yet, such an interoperable search service cannot supplant the many other search mechanisms optimized for particular technologies or communities of practice. Rather, the common search service for government data and information resources should be implemented typically as a supplement to other search mechanisms.

An interoperable search service must define how query request messages are handled at the service interface and what response messages can be returned. It must define how to handle Boolean query requests combining matches against abstract concepts such as Title, Author, Subject, and Date. (The abstraction here is important as it avoids the necessity for searchers to know the particular schemas of every collection being searched.) The search service would then respond with matching documents, available in a selection of original formats or a construction represented in an eXtensible Markup Language (XML) format.

1.3 Communities of Practice

1.3.1 Libraries

The most severe constraint on an international standard search service is the huge installed base of public mechanisms for search of government data and information supported among the world’s libraries. These communities have traditionally shared a common orientation on what could be loosely termed a “bibliographic catalog record”. On a global basis, it is clear that a few commonalities of search have become widely accepted: citation of items by title, author, date of publication, place of publication; and, cataloging of items by title, author, and subject. Since 1990, the international community of libraries has built on this base and achieved an extraordinary degree of consensus on the standard search service for library catalogs accessible over network technologies.

1.3.2 Online Information Services

Online information services (e.g., Lexis/Nexis, Chemical Abstracts Service, Dow Jones News Retrieval) represent another major community of practice. These services typically provide fee-for-service search access and for obvious commercial reasons they have less incentive than libraries to support open search standards. Yet, online information services are often major intermediaries for government holdings and most are already supporting the same international standard search service adopted by libraries.

1.3.3 Government Information Locators

Every government organization holds a wide variety of data and information resources and maintains a wide range of directories and other data and information locators. Data and information may be in the form of paper or electronic documents, budget tables, e-mail files, audio and video files, databases, and data systems of all kinds. U.K agencies maintain a public Information Asset Register of unpublished information holdings, i.e., information or collections of information, held electronically or in hard copy, which may not be publicly available. Whether for the sake of efficiency within the owning organization or to comply with requirements of public access, interoperability among such locators is essential. This requirement is rooted in public policy interests for government transparency, accountability, and protection of privacy. These communities of practice require an international standard search service addressing all types of government information over the long term.

1.3.4 Spatial Data Management

The international standard search service promulgated primarily by the library community has been adopted by all major vendors of geospatial products ("geospatial" refers to maps referenced to places on the Earth). More than 50 national governments operate national "clearinghouses" of geospatial data, and these distributed Clearinghouses are a key feature of the Global Spatial Data Infrastructure. Because all of these use the Geospatial Profile of the international standard search service **[GEO],** they enjoy full interoperability at the search service level with national, regional, thematic, and other geospatial data clearinghouses. A further profile refinement supports biological diversity communities of practice **[Bio]**.

1.3.5 Internet Search Engines

The development of Internet search engines can be traced to the advent of Web crawling technology. Because Web pages were constructed using HyperText Markup Language (HTML) and contained a high proportion of unstructured, "document-like" information, content was mostly indexed for search using full-text search technologies. As Web sites become more interactive, a smaller proportion of content is available to Web crawlers that deal only with static Web pages. The interactively generated content (sometimes known as "deep Web" or "hidden Web") is only searchable through a search service tailored to the particular collection at the Web site. Such site-specific search services for Web content often use search technology designed for Internet-wide search services.

For some years, debates raged over the idea that full-text search engines offered an unbeatable price/performance ratio in comparison to more traditional cataloging techniques. Today, most Internet search technologies offer a combination of wholly automated and machine-aided cataloging techniques, and treat Web content as semi-structured information. This responds to the user requirement for good "precision" as well as "recall"--especially important for Intranet and data mining applications. Commercial competition among Internet search engine vendors has impaired convergence on international standard search services. Yet, the range of search engine vendors with significant market share today is rather small, and leading technologies can readily accommodate an international standard search service.

1.4 Data and Information Types

1.4.1 Tangible Information (Documents, Artifacts, CD's and other media)

Although many people perform casual searching on their own, much of the world's public continues to rely on trained searchers and librarians to provide essential services in access to government information,

and a significant portion of that information is in paper and other tangible media. Public access is supported by specialized training in library schools and by a massive and pervasive infrastructure. For instance, the United States has more than 120,000 libraries, including over 1,300 Federal Depository Libraries. Libraries worldwide use a common record format to interchange bibliographic catalog records, the ubiquitous Machine-Readable Cataloging **[MARC]**. (Note: The Library of Congress has proposed a standard XML schema for MARC records.) The international standard search service is among the relatively narrow class of standards that support MARC. Support for this legacy interchange format is essential to meeting public policy goals of long-term access and customer-oriented design.

1.4.2 Metadata Associated with Digital Documents

Those involved in early development of the HTML standard anticipated the need to carry bibliographic citation information within an HTML page. The “title” metadata element is defined in HTML itself and a provision was made for additional metadata that could be defined elsewhere (i.e., the HTML “meta” sub-element within the “head” element). Embedded HTML metadata is akin to the “catalog in print” practice wherein bibliographic catalog citations are embedded in the front matter of a printed book. A profile of meta element names and definitions for interoperability purposes is provided by the Dublin Core Metadata Initiative **[DC]**, defined in international standard ISO 15836. A more elaborate mechanism for metadata embedding is provided by Encoded Archival Description **[EAD]**, a standard for encoding archival finding aids using ISO 8879, Standard Generalized Markup Language (SGML).

Of course, catalog records using these and other metadata conventions to characterize digital documents can also be stored separately. This approach is used in the Open Archives Initiative **[OAI]**, originally focused on access to e-print archives of scholarly communication. The recommended international standard search service has long been in production use for searching these metadata variants.

1.4.3 Catalogs, Directories, and Metadata Databases

Although rooted in bibliographic cataloging practice, the international standard search service features a general-purpose design that is independent of the actual data model supported by a compliant server. Because of this abstraction of the search function, the international standard search service has been adapted to a wide range of catalogs, directories and databases. For instance, the service has been used with Lightweight Directory Access Protocol (LDAP) for resource directories, and Structured Query Language (SQL) for relational databases, among others. In the case of databases, a typical application of interoperable search operates on databases of metadata, such as a specimen collection catalog.

1.4.4 File System Attributes and Object Properties

Data processing systems and most office systems have rarely attended to bibliographic cataloging practice, but the systems typically include semantically equivalent metadata structures (e.g., “file name” is roughly equivalent to “title”, “file owner” is roughly equivalent to “author”, etc.). Given that the international standard search service is independent of the particular data model, it has been fairly straightforward to provide a “semantic map” to surface the equivalences. The search service has been used with Internet Anonymous FTP Archive (IAFA) for file system catalogs and Distributed Authoring and Versioning for the Web (WebDAV). The service also adopted readily to the underlying data model of named properties and property sets that is defined for objects addressable by Microsoft software.

1.4.5 Service Registries and Semantics Registries

A gateway approach was used to demonstrate interoperability between services registries using Universal Description, Discovery, and Integration (UDDI) and the international standard search service. A similar approach is being pursued for interoperable searching of an ebXML registry and an ISO 11179 registry. These are of particular interest from an infrastructure architecture perspective in that such registries can also hold the key semantic concepts that are essential to further evolution of the international standard search service. As new communities of practice converge on interoperable searching using their key semantic concepts, these can be placed into the searchable semantic registry.

1.4.6 Extensibility, and Scalability

The query structure employed in the international standard search service is a nested Boolean structure that is adaptable to virtually all search tasks. The international standard search service has the usual sets of data *structures* (word, phrase, date, etc.) and *relations* (equal, greater than, etc.), but additional structure and relation semantics can be defined through profiles. For example, for searching by latitude and longitude, the Geospatial Profile **[GEO]** defines a “coordinates” *structure* and “enclosed within” *relation*. With the definition of appropriate semantics, the standard has been used for locating musical passages by the actual sounds **[music]**, finding pictures or faces by image characteristics, matching genes in gene sequence databases, and identifying people by their fingerprints.

Chemical Abstracts Services uses the standard for finding chemicals by formula or atomic structure, where a single chemical search can operate on up to 10,000 terms. Optimization of any particular implementation is not much constrained by the service itself; the standard service has been employed in the huge Mormon genealogical database that serves millions of searches per hour. Tests have shown that 250 parallel searches across Internet servers supporting this standard are no slower than the slowest server itself.

1.4.7 Internationalization

The international standard search service has long been used worldwide in many languages. It supports negotiation between client and server as to each other’s capabilities for the session. Character set and language negotiations are also supported.

2 Recommended Search Service Standard

The recommended international standard search service is **[ISO23950]** (identical to ANSI Z39.50). ISO 23950 is implemented as a client/server application using a request/response communications model. The ISO 23950 standard is used with many kinds of access control. Being a protocol with an abstract data model, servers can exert access at any level: services, sessions, distributed resources, databases, records, and/or fields.

One ISO 23950 profile specifies operation over TCP/IP with Abstract Syntax Notation (ASN.1) encoding **[RFC1729]**. Another ISO 23950 profile specifies operation as a Web Service **[SRW/SRU]** (the SOAP version is known as “SRW” while the REST version is known as “SRU”).

ISO 23950 application profiles (some examples are noted elsewhere in this document) define subsets of operation and additional semantics for higher degrees of interoperability or simplified implementations.

3 Implementation Issues

3.1 Justifying Standardization

For open societies worldwide, there is a strong public policy interest in highly interoperable search services supporting public access to government information. In addition to enhancing the effectiveness and transparency of government, adoption of a standard search service can also be justified on efficiency grounds. Government-wide efficiencies result from increased sharing of information and lowered costs for developing customized bridges as are otherwise required to merge information from multiple government sources. Also, within any single organization, adoption of a search service based on international standards provides a degree of “future-proofing” against changes in search technologies. In effect, a standard service provides interoperability across time—minimizing disruptions expected with migration to new technologies and ensuring continued access to holdings supported by legacy technologies.

3.2 Minimal Imposition in Markets

Because governments wisely avoid undue imposition in marketplaces, the adoption of any standard by governments must be carefully considered. The public policy interest in this case prompts adoption of the standard government-wide but does not extend to search services offered outside of the government sector. All search services supported by government would support the standard but intermediaries would not be so required, even when offering mediated access to government information.

Governments would continue to rely primarily on off-the-shelf technology as offered by the wide range of commercial and open-source vendors of search services worldwide. Product and service offerings to government would be required to support the standard search service. In most cases, this requirement entails minimal cost to search service vendors as they already support alternative service interfaces within current technology. ISO 23950 gateways are available for the major search technologies that publish an applications program interface. Research products and tools supporting this standard service are also available. Also, the standard supports extension mechanisms to nurture innovation in areas not yet ready for the broadest level of standardization.

3.3 Implementation Costs

Governments expend massive resources on disseminating government data and information. Support of a common search standard would entail additional cost, but that addition would be a small percentage of the overall cost. For example, the cost of an Internet portal for government information may be on the order of millions of dollars per year while the software cost for supporting the standard search service may be on the order of thousands of dollars per year.

There is an ongoing operational cost to government in supporting any search service. For every major type of information resource offered through the search service, someone familiar with the holdings must identify what equivalences exist between the international standard search concepts and the locally held information. This one-time “semantic mapping” task is typically handled by a system administrator, and is usually the same function as that required for setting up a proprietary search interface anyway.

For an existing non-standard search interface, costs can be minimized by implementing the standard search service as a query translation gateway. Such a translation gateway interacts with the existing search interface just as would any user at an Internet browser, and it can be located anywhere on the Internet. Through its support of standards-based search clients, the gateway expands the accessibility of an existing search service without any re-engineering cost.

3.4 Government Standardization Mechanisms

Governments have often instituted law or policy with regard to standardization. In the United States, public law (United States Code Chapter 44, Section 3511) defines responsibilities for establishing the

U.S. Federal Government Information Locator Service. Federal policy (OMB Memorandum 98-5) points to the Federal Information Processing Standard (FIPS Pub 192-1) that adopts the international standard search service **[GILS]**. A similar set of law, policy, and standards is in place for geospatial data in the United States, and corollary examples exists in many States of the U.S. Such policies and standards are cited specifically in the procurement documents issued by government, and vendors certify compliance.

Some governments and other organizations have established policies requiring certain metadata to be associated with Government-operated Internet sites. In the U.K, the e-Government Metadata Standard **[e-GMS]** is concerned with the particular facets of metadata intended to support resource. (The European Commission has similarly issued the **[MIReG]** metadata standard for use by all Member States of the European Union.) A standard search service policy that references that metadata standard would allow for building actual interoperable systems for resource discovery.

3.5 Intermediation Policies

Even within a single nation there may be millions of government organizations, and most of these may provide a search service. Intermediaries, within or external to government, often aggregate services across government organizations that share a management hierarchy, geographic area, or focal theme. A common example is the “referral service” portal wherein a common search mechanism is used to help users find appropriate information. Like other intermediaries, these aggregated search services may be able to bring to bear superior expertise and resources.

Anyone attempting aggregation stands to gain from having a standard search service across the organizations they aggregate. However, aggregation of content through technology must address certain public policy concerns. Public trust in the provenance and quality of information provided by a government source is crucial to the functioning of open societies. Although “seamless” aggregation of information across sources may be a convenience, agreed mechanisms and inter-organization policies to assure correct attribution and context are essential. It is also essential that more specialized search mechanisms available from the originating institutions are not supplanted by the intermediate service. Otherwise, rather than enhancing public access to diverse information sources across government, the ultimate affect might be to collapse that diversity and trivialize information search techniques generally.

3.6 Organizing Content

The interoperable search service precisely defines the way in which searches are expressed and communicated between the client and server on a network. In addition to a common syntax for expressing search criteria, the interoperable search service operates at the semantic level through reference to abstract search concepts. These abstract search concepts must be semantically mapped to appropriate components of the content. Often, a search concept can be directly mapped to one or more existing fields in a database or to existing elements and attributes in an XML Schema. But, a search concept may also be semantically mapped to some characteristic that does not pre-exist. For example, if the place name “London” is referenced to a geographic extent, the searcher can find maps that include London even though the literal name “London” does not appear on the map. There are many approaches to semantic mapping and research is very active in this area, especially in situations where structural characteristics are inferred rather than designated by a human designer. It is important to note that the interoperable search service itself does not standardize on any particular mechanism and so is quite compatible with many intriguing and powerful new techniques of semantic mapping.

In addition to organizing content within a searchable collection, designers of search facilities are often faced with compiling information as a logical collection. For example, the Open Archives Initiative promotes a Protocol for Metadata Harvesting that can be a useful way to maintain a collection of Web resources. Also, some “meta-search” facilities (sometimes called “Clearinghouses”), allow for distributed search of many collections with target collections selected by referral records that characterize the coverage of each collection. Here again, it is important to note that the interoperable search service itself does not standardize on any particular approach and so is compatible with many and diverse approaches to compiling collections of information.

4 References

4.1 Normative

[ISO23950] *ANSI/NISO Z39.50-1995, Information Retrieval (Z39.50): Application Service Definition and Protocol Specification,* http://lcweb.loc.gov/z3950/agency , 1995.

[RFC2119] S. Bradner, *Key words for use in RFCs to Indicate Requirement Levels*, http://www.ietf.org/rfc/rfc2119.txt, IETF RFC 2119, March 1997.

4.2 Informative

[Bio] The Species Analyst, *Darwin Core V1*, http://tsadev.speciesanalyst.net/documentation/ow.asp?DarwinCoreV1 , December 2001.

[DC] OCLC Research, *Dublin Core Metadata Element Set, Version 1.1*, http://dublincore.org/documents/dces/, June 2003.

[e-GMS] U.K. Government, *e-Government Metadata Standard (e-GMS)*, http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=768, June 2003.

[EAD] Encoded Archival Description, http://www.loc.gov/ead/ , 2002.

[GILS] E. Christian, *Application Profile For The Government Information Locator Service (GILS)*, http://www.gils.net/prof_v2.html, April 1997.

[GEO] D. Nebert, *Z39.50 Application Profile for Geospatial Metadata or "GEO"*, http://www.blueangeltech.com/Standards/GeoProfile/geo22.htm, May 2000.

[MARC] American National Standards Institute, *American National Standard Z39.2-1985 Bibliographic Information Interchange*. New York: American National Standards Institute, 1985.

[MIReG] {to be completed}

[music] M. Dovey, *Online Music Recognition and Searching*, http://www.omras.org , 2003.

[OAI] Open Archives Initiative, http://www.openarchives.org/ , 2003.

[RFC1729] C. Lynch, *Using Z39.50-1992 Directly over TCP*, ftp://ftp.loc.gov/pub/z3950/profiles/tcp.txt, IETF RFC 1729, March 1997.

[SRW/SRU] U.S. Library of Congress, *Z39.50-International: Next Generation*, http://www.loc.gov/z3950/agency/zing/zing-home.html, January 2003.

Appendix A. Acknowledgments

The following individuals were members of the committee during the development of this specification:

- Eliot Christian (chair), U.S. Geological Survey
-

In addition, the following people made contributions to this specification:

- John Borras, Office of the e-Envoy, U.K.
- Bob Colomb, University of Queensland, Australia
- Matthew Dovey, Oxford University, U.K.
- Ralph LeVan, Online Computer Library Center, U.S.
- Andy Powell, UKOLN, University of Bath, U.K.
- Rob Sanderson, University of Liverpool, U.K.
- Mike Taylor, Index Data, U.K.

Appendix B. Revision History

Rev	Date	By Whom	What
searchservice-01	2003-11-17	Eliot Christian	Initial version
searchservice-02	2003-12-31	Eliot Christian	Revised per public comments (see First Revision Note below) as follows: - In section 1.3.5, changed 'This other content' to 'This interactively generated content' - Revised text in section 1.4.3 for clarity - Added text to section 2 concerning access control in ISO 23950 - Added text to section 3.2 concerning extension mechanisms in ISO 23950 - Added section 3.6 discussing mechanisms for organizing content (these are not part of the search service per se) - Deleted text throughout that had referred to 'archives' and 'records management' - Updated the citation for e-GMS
searchservice-CD	2004-01-26	John Borras	Committee Draft approved by TC

First Revision Note: On November 17, 2003, the OASIS e-Government Technical Committee publicly solicited comments on the initial version of this document. Comments were to be sent to the author by 16 December, 2003. A summary of comments through 18 December 2003 and draft responses are at http://www.oasis-open.org/apps/org/workgroup/egov/download.php/4651/wd-egov-searchservice-01-comments.html

Appendix C. Notices

OASIS takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on OASIS's procedures with respect to rights in OASIS specifications can be found at the OASIS website. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementors or users of this specification, can be obtained from the OASIS Executive Director.

OASIS invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to implement this specification. Please address the information to the OASIS Executive Director.