



Search Service Interoperability

Working Draft 01, 17 November 2003

Document identifier:

wd-egov-searchservice-01

Location:

<http://www.oasis-open.org/egov/docs/>

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 Working Draft document is updated periodically on no particular schedule. Send comments to the editor.

Committee members should 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>).

29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table of Contents

1	Requirements	3
1.1	Government Stake in Search Interoperability.....	3
1.2	Search Service Defined	3
1.3	Communities of Practice	3
1.3.1	Libraries, and Archives	3
1.3.2	Online Information Services	4
1.3.3	Government Information Locators and Records Management	4
1.3.4	Spatial Data Management.....	4
1.3.5	Internet Search Engines.....	4
1.4	Data and Information Types	5
1.4.1	Tangible Information (Documents, Artifacts, CD's and other media).....	5
1.4.2	Metadata within Digital Documents	5
1.4.3	Catalogs, Directories, and Metadata Databases.....	5
1.4.4	File System Attributes and Object Properties	5
1.4.5	Service Registries and Semantics Registries	5
1.4.6	Extensibility, and Scalability	6
1.4.7	Internationalization	6
2	Recommended Search Service Standard	7
3	Government Issues	8
3.1	Justifying Standardization.....	8
3.2	Minimal Imposition in Markets	8
3.3	Implementation Costs	8
3.4	Government Standardization Mechanisms.....	8
3.5	Intermediation Policies.....	9
4	References	10
4.1	Normative.....	10
4.2	Informative	10
	Appendix A. Acknowledgments	11
	Appendix B. Revision History.....	12
	Appendix C. Notices.....	13

61 1 Requirements

62 1.1 Government Stake in Search Interoperability

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

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

74 1.2 Search Service Defined

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

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

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

93 1.3 Communities of Practice

94 1.3.1 Libraries, and Archives

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

103 **1.3.2 Online Information Services**

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

109 **1.3.3 Government Information Locators and Records Management**

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

121 **1.3.4 Spatial Data Management**

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

129 **1.3.5 Internet Search Engines**

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

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

146 **1.4 Data and Information Types**

147 **1.4.1 Tangible Information (Documents, Artifacts, CD's and other media)**

148 Although many people perform casual searching on their own, much of the world's public continues to
149 rely on trained searchers and librarians to provide essential services in access to government information,
150 and a significant portion of that information is in paper and other tangible media. Public access is
151 supported by specialized training in library schools and by a massive and pervasive infrastructure.
152 For instance, the United States has more than 120,000 libraries, including over 1,300 Federal Depository
153 Libraries. Libraries worldwide use a common record format to interchange bibliographic catalog records,
154 the ubiquitous Machine-Readable Cataloging [**MARC**]. (Note: The Library of Congress has proposed a
155 standard XML schema for MARC records.) The international standard search service is among the
156 relatively narrow class of standards that support MARC. Support for this legacy interchange format is
157 essential to meeting public policy goals of long-term access and customer-oriented design.

158 **1.4.2 Metadata within Digital Documents**

159 Those involved in early development of the HTML standard anticipated the need to carry bibliographic
160 citation information within an HTML page. The "title" metadata element is defined in HTML itself and a
161 provision was made for additional metadata that could be defined elsewhere (i.e., the HTML "meta" sub-
162 element within the "head" element). Embedded HTML metadata is akin to the "catalog in print" practice
163 wherein bibliographic catalog citations are embedded in the front matter of a printed book. A profile of
164 meta element names for interoperability purposes is provided by the Dublin Core Metadata Initiative [**DC**],
165 defined in international standard ISO 15836. This style of metadata is also used in the Open Archives
166 Initiative [**OAI**], originally focused on access to e-print archives of scholarly communication. A more
167 elaborate mechanism for metadata embedding is provided by Encoded Archival Description [**EAD**], a
168 standard for encoding archival finding aids using ISO 8879, Standard Generalized Markup Language
169 (SGML). The recommended international standard search service has long been in production use for
170 searching each of these and other metadata conventions for digital documents.

171 **1.4.3 Catalogs, Directories, and Metadata Databases**

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

179 **1.4.4 File System Attributes and Object Properties**

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

188 **1.4.5 Service Registries and Semantics Registries**

189 A gateway approach was used to demonstrate interoperability between services registries using
190 Universal Description, Discovery, and Integration (UDDI) and the international standard search service.
191 A similar approach is being pursued for interoperable searching of an ebXML registry and an ISO 11179

192 registry. These are of particular interest from an infrastructure architecture perspective in that such
193 registries can also hold the key semantic concepts that are essential to further evolution of the
194 international standard search service. As new communities of practice converge on interoperable
195 searching using their key semantic concepts, these can be placed into the searchable semantic registry.

196 **1.4.6 Extensibility, and Scalability**

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

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

211 **1.4.7 Internationalization**

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

215 **2 Recommended Search Service Standard**

216 The recommended international standard search service is **[ISO23950]** (identical to ANSI Z39.50).
217 ISO 23950 is implemented as a client/server application using a request/response communications
218 model.

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

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

224 **3 Government Issues**

225 **3.1 Justifying Standardization**

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

235 **3.2 Minimal Imposition in Markets**

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

241 Governments would continue to rely primarily on off-the-shelf technology as offered by the wide range of
242 commercial and open-source vendors of search services worldwide. Product and service offerings to
243 government would be required to support the standard search service. In most cases, this requirement
244 entails minimal cost to search service vendors as they already support alternative service interfaces
245 within current technology. ISO 23950 gateways are available for the major search technologies that
246 publish an applications program interface. Research products and tools supporting this standard service
247 are also available.

248 **3.3 Implementation Costs**

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

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

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

264 **3.4 Government Standardization Mechanisms**

265 Governments have often instituted law or policy with regard to standardization. In the United States,
266 public law (United States Code Chapter 44, Section 3511) defines responsibilities for establishing the
267 U.S. Federal Government Information Locator Service. Federal policy (OMB Memorandum 98-5) points to

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

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

279 **3.5 Intermediation Policies**

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

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

295 4 References

296 4.1 Normative

- 297 [ISO23950] *ANSI/NISO Z39.50-1995, Information Retrieval (Z39.50): Application Service*
298 *Definition and Protocol Specification*, <http://lcweb.loc.gov/z3950/agency> , 1995.
- 299 [RFC2119] S. Bradner, *Key words for use in RFCs to Indicate Requirement Levels*,
300 <http://www.ietf.org/rfc/rfc2119.txt>, IETF RFC 2119, March 1997.

301 4.2 Informative

- 302 [Bio] The Species Analyst, *Darwin Core V1*,
303 <http://tsadev.speciesanalyst.net/documentation/ow.asp?DarwinCoreV1> ,
304 December 2001.
- 305 [DC] OCLC Research, *Dublin Core Metadata Element Set, Version 1.1*,
306 <http://dublincore.org/documents/dces/>, June 2003.
- 307 [e-GMS] U.K. Government, *e-Government Metadata Standard (e-GMS)*,
308 <http://www.e-envoy.gov.uk/assetRoot/04/00/09/73/04000973.pdf>, April 2002.
- 309 [EAD] Encoded Archival Description, <http://www.loc.gov/ead/> , 2002.
- 310 [GILS] E. Christian, *Application Profile For The Government Information Locator Service*
311 *(GILS)*, http://www.gils.net/prof_v2.html, April 1997.
- 312 [GEO] D. Nebert, *Z39.50 Application Profile for Geospatial Metadata or "GEO"*,
313 <http://www.blueangeltech.com/Standards/GeoProfile/geo22.htm>, May 2000.
- 314 [MARC] American National Standards Institute, *American National Standard Z39.2-1985*
315 *Bibliographic Information Interchange*. New York: American National Standards
316 Institute, 1985.
- 317 [MIReG] {to be completed}
- 318 [music] M. Dovey, *Online Music Recognition and Searching*, <http://www.omras.org> ,
319 2003.
- 320 [OAI] Open Archives Initiative, <http://www.openarchives.org/> , 2003.
- 321 [RFC1729] C. Lynch, *Using Z39.50-1992 Directly over TCP*,
322 <ftp://ftp.loc.gov/pub/z3950/profiles/tcp.txt>, IETF RFC 1729, March 1997.
- 323 [SRW/SRU] U.S. Library of Congress, *Z39.50-International: Next Generation*,
324 <http://www.loc.gov/z3950/agency/zing/zing-home.html>, January 2003.
- 325

326 **Appendix A. Acknowledgments**

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

328 • Eliot Christian (chair), U.S. Geological Survey

329 •

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

331 • John Borrás, Office of the e-Envoy, U.K.

332 • Maewyn Cumming, Office of e-Envoy, UK

333 • Bob Colomb, University of Queensland, Australia

334 • Matthew Dovey, Oxford University, U.K.

335 • Ralph LeVan, Online Computer Library Center, U.S.

336 • Andy Powell, UKOLN (UK Office for Library Networking), U.K.

337 • Rob Sanderson, University of Liverpool, U.K.

338 • Mike Taylor, Index Data, U.K.

339

Appendix B. Revision History

Rev	Date	By Whom	What
searchservice-01	2003-11-17	Eliot Christian	Initial version

340

341 **Appendix C. Notices**

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

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

353 Copyright © OASIS Open 2003. *All Rights Reserved.*

354 This document and translations of it may be copied and furnished to others, and derivative works that
355 comment on or otherwise explain it or assist in its implementation may be prepared, copied, published
356 and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice
357 and this paragraph are included on all such copies and derivative works. However, this document itself
358 does not be modified in any way, such as by removing the copyright notice or references to OASIS,
359 except as needed for the purpose of developing OASIS specifications, in which case the procedures for
360 copyrights defined in the OASIS Intellectual Property Rights document must be followed, or as required to
361 translate it into languages other than English.

362 The limited permissions granted above are perpetual and will not be revoked by OASIS or its successors
363 or assigns.

364 This document and the information contained herein is provided on an "AS IS" basis and OASIS
365 DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY
366 WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR
367 ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.