# Efficient and Effective Information Retrieval and Sharing (EEIRS) Request For Information (RFI) Response Analysis

December 2005

# 1. Introduction

In September and October 2005, U.S. General Services Administration's (GSA) Office of Governmentwide Policy surveyed industry, academia, and government agencies with an instrument that covered a wide range of information retrieval, categorization, dissemination, and sharing needs and capabilities. The intent was to identify and promote the most cost-effective means to search for, identify, locate, retrieve, and share information, and assess the net performance difference (including cost-benefits) of assigning metadata and/or a controlled vocabulary to various types of information versus not doing so. Specifically, this instrument tested the following hypothesis:

> "For the majority of government information, exposing it to indexing with commercial search technology is sufficient to meet the information categorization, dissemination, and sharing needs of the public and as required by law and policy."

GSA received 47 detailed responses to its survey. These responses overwhelmingly supported the hypothesis. Specifically, analysis of the responses found that, in over 56% of the cases, respondents favored the use of search technologies over other solutions requiring human investment in the advance preparation of content, such as metadata tagging, the creation of controlled vocabularies and other structured information models, and/or explicit cataloging of information. By contrast, respondents documented the need to perform *significant* advance preparation of content to facilitate information retrieval and sharing in only 14% of the cases. In 30% of the cases, respondents documented the need to perform *some* advance preparation of content to enable efficient searching. This document describes these findings in more detail, specifically covering the following topics:

- A description of the survey instrument used to test the hypothesis;
- A description of the sample providing data for analysis;
- A detailed analysis of responses, providing justification for the findings described above.

# 2. The Survey Instrument

GSA in conjunction with OMB and agency subject matter experts developed the Efficient and Effective Information Retrieval and Sharing Request for Information (RFI) (GSA # GS00V05PDC0062). The primary objective of GSA was to gather and evaluate various approaches and carefully consider factors such as cost, ease of implementation, interoperability and sustainability of operations in order to identify ways to promote greater public access to and Federal agency sharing of information. The RFI was designed to support the collection of innovative approaches and research as to the most efficient and cost-effective means to search for, identify, locate, retrieve, and share government information. The RFI contained the following sections:

- Purpose – Goals and objectives of the EEIRS effort
- Background – Information sharing and the Federal Government
- RFI Questionnaire – Survey instrument to gather potential information approaches.

- Instructions to Prospective Respondents – Guidance on format and submission process
- Disclaimers – RFI stipulations and provisions

The RFI was designed to canvass the public and industry segments best suited to identify potential approaches to improved information access, dissemination, and sharing. The EEIRS RFI was released to the public on September 16, 2005. Potential respondents were given approximately 5 weeks to prepare quality responses to the RFI solicitation - responses were due to GSA no later than October 21, 2005.

Scenarios: To help provide the context necessary to elicit quality responses from respondents, the RFI contained the following seven scenarios describing various common information discovery, retrieval, aggregation, and sharing needs.

1. Researching Unexplained Illnesses Among Defense Contractors
2. Performing a Search for an Expert
3. Performing Academic Research
4. Conducting an Information Audit Trail
5. Sharing Law Enforcement Information across Jurisdictional Boundaries
6. Possible Forged Identity
7. Citizen looking for all online government information regarding a unique topic

Respondents were encouraged to use the scenarios to help frame their responses and provide context and examples in response to the questions. In addition, respondents were allowed to develop their own scenarios as needed.

Questions: The EEIRS RFI support team prepared the RFI questions based on input from GSA, OMB and Agency subject matter experts and included the following topic areas:

- General Approach – Recommended vision/approach for enabling the searching discovery, retrieval, and sharing of information.
- Specific Capabilities – Description of the functional and technical capabilities of the proposed approach
- Acquisition Strategy – Procurement strategies and recommendations for the proposed approach
- Implementation – Implementation methodology, estimated timelines and relevant past experience in implementing the proposed approach
- Program Management – Outline of the proposed management methodology including team structure, skill matrix, risk management and performance measurement
- Cost and Benefits – High level cost estimates and benefits for the proposed approach.

The broad intent of the EEIRS RFI was to evaluate the feasibility of search-engine technology versus advance information preparation through developing formal information models such as taxonomies and metadata-tagging to locate, access, retrieve, and share government information in an effective and efficient manner. To facilitate the collection of data and support the analysis process, the RFI questions were designed to address a set of identified functional and technical capabilities for improved information retrieval and sharing.

Questions focused on how well proposed solutions addressed the following information retrieval, dissemination and sharing capabilities:

- *Multiple Jurisdictions/Stakeholders* – This capability represents the ability to search, discover, retrieve, and share information across legal/jurisdictional/organizational boundaries. It is represented in the RFI as questions 3.1.1.a and b.
- *Multiple data sources* – This capability represents the ability to search, discover, retrieve, and share information across databases, websites, repositories, record management systems, etc. It is represented in the RFI as question 3.1.1.c.
- *Multiple data types* – This capability represents the ability to search, discover, retrieve, and share multiple data types (e.g., structured, unstructured, tabular, multi-media, email, geospatial, biometric). It is represented in the RFI as question 3.1.1.d.
- *Data aggregation and Integration* – This capability represents the ability to aggregate information in a value added manner (i.e., combining different data types into useful work products). It is represented in the RFI as question 3.2.1.
- *Notification* – This capability represents the ability to notify users when an information resource has been added to, altered or removed from the shared computing environment. It is represented in the RFI as question 3.2.10.
- *Precision and Recall* – This capability describes approaches for calculating relevance, locating relevant information not containing the original query, indexing the "deep web", etc. It is represented in the RFI as questions 3.2.3, 3.2.4, 3.2.5, and 3.2.6.
- *Data Quality, Authenticity, and Archival* – This capability describes approaches to ascertaining the quality, authenticity, durability, and longevity of an information resource. It is represented in the RFI as questions 3.2.7, 3.2.8, and 3.2.9.
- *Interoperability* – This capability describes standards implemented and/or supported by the proposed approach. It is represented in the RFI as questions 3.1.1.e and 3.2.13.
- *Information Extraction and Visualization* – This capability represents the ability to rapidly discover relevant facts buried in large volumes of data, organizing/presenting/visualizing large data sets in an understandable, user-friendly manner. It is represented in the RFI as questions 3.2.2, 3.2.11, and 3.2.12.
- *Benefits of Advanced Preparation of Information* – This capability describes advantages gained through advance preparation of content for search and retrieval (i.e., metadata tagging). It is represented in the RFI as question 3.1.2.

## 3. THE SAMPLE

**Selecting the initial population:** The EEIRS RFI support team implemented a variety of techniques and leveraged several data sources to compile a comprehensive list of potential RFI recipients from across the United States. Using the sources below, the EEIRS RFI support team targeted representatives from industry, government and academia who are practitioners and subject matter experts in the field of information sharing.

- GSA/OMB and Agency experts
- Existing agency vendor lists
- Market research companies
- Government and Industry Councils
- Subject Matter Experts from EEIRS RFI support team

To ensure all points of view were represented, the support team solicited information not only from the traditional IT vendor community, but also included system integrators as well as management consulting firms, government agencies and academics. The RFI support team reached out to 129 organizations (and in turn a number of associations reached out to their members) including search vendors; metadata advocates and companies using an integrated approach combining both search and metadata technologies.

**Efforts to distribute and advertise the instrument:** The EEIRS RFI support team conducted a variety of outreach activities to generate awareness and interest among potential respondents including:

- Internet: The RFI was posted directly to FedBizOpps (www.fbo.gov), a single government point of entry for Federal government procurement opportunities. Through FedBizOpps, commercial vendors seeking Federal markets for their products and services can search, monitor and retrieve opportunities solicited by the entire Federal contracting community.
- Telemarketing: EEIRS RFI support team telephoned the points of contact identified in the Market Pool List to inform potential respondents of the upcoming release of the RFI solicitation.
- Email Notifications: The EEIRS RFI support team sent RFI notifications and Practitioners Day invitations to the points of contact identified in market pool list.
- Industry Council Announcements: Informed members of the RFI solicitation and encouraged members to respond.
- Practitioners Day: On September 27, 2005, GSA conducted the EEIRS RFI Practitioners Day. The intent of the Practitioners Day was to explain the purpose of the RFI and to provide potential respondents with an opportunity to pose clarifying questions to inform responses to the EEIRS RFI. 86 participants attended representing 52 organizations.

Additional press coverage regarding the RFI was received on the radio and in government trade magazines, including:

- Government trade magazines – Federal Computer News[1] and Government Computer News[2] printed articles regarding GSA's efforts to improve Federal information sharing.
- Radio –WTOP and Federal News Radio made mention of the EEIRS RFI solicitation a few times in their weekly Government Technology Report segment.

**Composition of the survey respondents:** OMB and GSA received 47 responses from industry, government and academia. A breakout of the responses by organization type is shown below. For a complete listing of respondents by organization type see Appendix A.

| Organization Type | Number of RFI Responses |
|---|---|
| Industry | 42 |

[1] http://www.fcw.com/article91035-10-06-05-Web
[2] http://www.gcn.com/vol1_no1/storage/37117-1.html?topic=storage

| | |
|---|---|
| Government | 3 |
| Academia | 2 |
| Total | 47 |

Most of the respondents incorporated both search and metadata - based approaches in answering the 19 capabilities-related questions in the RFI. Therefore, it was difficult to categorize the RFI respondents as purely a search or metadata vendor. As a result, the analysis was performed and rolled up on a per-question basis. This is shown in more detail in the analysis section below.

## 4. ANALYSIS OF RESPONSES

This section provides an analysis of the RFI responses. In performing the analysis, the RFI support team reviewed responses to the questions in sections 3.1 (General Approach) and 3.2 (Specific Capabilities), and attempted to categorize respondent approaches as one of the following three patterns:

- *Search Pattern* – Approaches requiring no advance preparation of content. This includes any approach completely automating the process of indexing, searching, finding, retrieving, and sharing information. Typically, the only level of effort required under this pattern is the procurement of the solution, and its installation and configuration to run in a specific production environment.
- *Integrated Solution Pattern* – Approaches requiring some advance preparation of content. This includes any approach combining search technologies with some human-generated metadata to index, search, find, retrieve, and share information (e.g., a search application enhanced with a human generated controlled vocabulary, such as a taxonomy or thesaurus). Typically, the only advance preparation required under this pattern is the creation of information models (such as controlled vocabularies) to guide information retrieval applications. Manual tagging and/or cataloging of discrete information resources is typically **not** required under this pattern. In this pattern, any advance meta-tagging needed is performed automatically by a software tool.
- *Metadata Pattern* – Approaches requiring significant advance preparation of content. This includes any approach requiring manual tagging and/or cataloging of discrete information resources to enable accurate indexing, searching, finding, retrieving, and sharing information (e.g. a content or records management system requiring content providers to manually categorize documents using a taxonomy or manually tagging documents with specific metadata fields).

### 4.1. Quantitative Analysis

As shown in the summary table below, respondents overwhelmingly favored search-based approaches requiring no advance preparation of content (**56%**) over approaches requiring some advance preparation of content (e.g., search aided by controlled vocabularies) (**30%**) or significant advance preparation of content (e.g., explicit metadata tagging of content) (**14%**):

| Capability | Question | No advance preparation required | Some advance preparation required | Significant advance preparation required | No answer or NA | Total responses |
|---|---|---|---|---|---|---|
| Multiple Jurisdictions/Stakeholders | 3.1.1.a | 6 | 13 | 2 | 26 | 47 |
| | 3.1.1.b | 12 | 8 | 0 | 27 | 47 |
| Multiple data sources | 3.1.1.c | 9 | 12 | 0 | 26 | 47 |
| Multiple data types | 3.1.1.d | 15 | 6 | 1 | 25 | 47 |
| Data aggregation & integration | 3.2.1 | 12 | 11 | 1 | 23 | 47 |
| Notification | 3.2.10 | 14 | 11 | 4 | 18 | 47 |
| Precision & recall | 3.2.3 | 19 | 9 | 3 | 16 | 47 |
| | 3.2.4 | 9 | 13 | 2 | 23 | 47 |
| | 3.2.5 | 17 | 8 | 5 | 17 | 47 |
| | 3.2.6 | 20 | 4 | 4 | 19 | 47 |
| Data quality, authenticity, and archival | 3.2.7 | 14 | 0 | 15 | 18 | 47 |
| | 3.2.8 | 14 | 0 | 10 | 23 | 47 |
| | 3.2.9 | 17 | 0 | 7 | 23 | 47 |
| Interoperability[3] | 3.1.1.e | 2 | 14 | 1 | 30 | 47 |
| | 3.2.13 | 29 | 6 | 2 | 10 | 47 |
| Information extraction and visualization | 3.2.2 | 11 | 5 | 4 | 27 | 47 |
| | 3.2.11 | 11 | 7 | 2 | 27 | 47 |
| | 3.2.12 | 19 | 6 | 0 | 22 | 47 |
| Benefits of advanced preparation of information | 3.1.2 | 18 | 8 | 4 | 17 | 47 |
| Sum of responses | | 268 | 141 | 67 | 417 | 893 |
| **Percentage** | | **30%** | **16%** | **8%** | **47%** | **100%** |
| **Adjusted Percentage (excludes no responses)** | | **56%** | **30%** | **14%** | | |

Table 1: Quantitative Analysis Summary

Appendix B provides a detailed analysis of responses by RFI question. This analysis includes:

- Specific findings i.e., how many respondents presented an approach requiring no, some, or significant advance preparation of content.
- Breakdown of specific approaches, categorizing each response by specific type of approach.

### 4.2. Qualitative Analysis

As a general statement, most respondents were consistent in expressing the opinion that search technologies were the most cost effective means of information discovery, retrieval, and sharing for most data types. One example is the US Department of Energy Office of

[3] The term *interoperability* is used in a manner consistent with the E-Government Act of 2002 (Public Law 107-347)

Scientific and Technical Information (OSTI). OSTI is developer and maintainer of one of the world's best and most comprehensive collections of scientific data and has extensive experience in the use of metadata and manual cataloging methods. In its response, OSTI stated that:

> *"Search technology has progressed far enough so that manual categorization and metadata tagging of textural documents is no longer necessary, and any perceived gain in accessibility does not justify the cost of categorization."*[4]

Many respondents from industry agreed. For example, one large systems integrator pointed out in its response information on the Internet is often poorly meta-tagged, thereby requiring the employment of algorithms and technology designed to understand and analyze content without the aid of additional metadata. This vendor also cites in its experience only marginal gains from advanced preparation of information. Also, one large search company stated emphatically in its response that metadata was not required:

> *"It is not necessary to assign metadata, or a controlled vocabulary, to the corpus of assets in every scenario in order for a highly relevant search to be available. [Our product] can index metadata and filter results based on metadata but does not require metadata to exist in order to provide highly relevant search across more than 220 data formats."*

Then there is the issue of content producer discipline in consistently applying metadata to information resources. As one large business solutions company pointed out:

> *"[We] believe[s] that if knowledge orchestration requires significant manual preparation of content, then it will not be successful. Simply put, experience does not suggest that users will take the time to properly assign meta tags to content. Further, there are significant semantic issues related to having individual users code information using their own interpretation of meaning."*

To paraphrase another respondent describing the pitfalls of taxonomies:

> *Taxonomy forces searchers to guess how taxonomists, the masters of metadata, organized their world, forcing the searcher to follow the arbitrary path of the taxonomist."*

But to be fair, this respondent also stated a mixed approach was appropriate, i.e., neither free text search nor full metadata tagging, to avoid information overload, or as they put it, "*information being hidden in plain sight."*

Some other respondents did, however, advocate a metadata-based approach in certain situations. For example, the US National Archives and Records Administration (NARA), in its response, stated:

> *"For either information or records to be trustworthy, they must have additional information either embedded within the content itself or information associated with the content that can provide some degree of assurance of authenticity, reliability and integrity, now and in the future; NARA believes this can be accomplished via the incorporation of records management Metadata. This is the only way that Government*

[4] Department of Energy Office of Scientific and Technical Information (OSTI) response, page 2

*can be reasonably certain that it is providing itself and others with authoritative information."*[5]

Also, from a "deep web" accessibility standpoint, many respondents, including many search vendors, acknowledged that this capability is difficult to implement without at least some human intervention. One of the more popular approaches involved explicitly mapping in data sources not easily reachable via typical web search engine spiders. As one respondent put it:

*"An organization's structured data (enterprise applications and relational databases) do not usually lend themselves to document-centric approaches. However, they do already have significant context that can be leveraged. This is the metadata implicit in the table and column names, datatypes of element values, business descriptions of the elements, relationships expressed through foreign keys, mappings to a business data dictionary, and other semantic relationships stored separately from the data source or enterprise application. Such context can be made visible, discoverable and query-able using [our] model-based abstraction technology. These "deep web" relational and application resources are thus made visible to search and discovery engines, without the need to invasively impose additional metadata or cataloging on the original information source – the context is already captured by the inherent structure."*

For the most part, respondents advocating or suggesting advance preparation such as metadata tagging was necessary, described largely automated methods of doing so with little need for actual human intervention other than installing, configuring, and running automated tools.

In light of the recent controversy over the pending withdrawal of the GILS (Government Information Locator System) as a federal information processing standard, a significant finding of this study was the lack of support (and, indeed, demand) for the ISO 23950 search interoperability standard, on which the GILS standard is based. Only 6 out of 47 respondents (13%) cited support for this standard in their response. One respondent serving customers beyond the Federal Government, pointed out:

*"To date customer demand for ISO 23950 has been low. Customer requests are an order of magnitude higher for supporting web services (SOAP/WSDL/UDDI)."*

### 4.3. Additional Findings

Beyond a simple popularity measure, it is possible to generalize from the RFI responses certain scenarios when it is appropriate to use the various information retrieval and sharing patterns. These include:

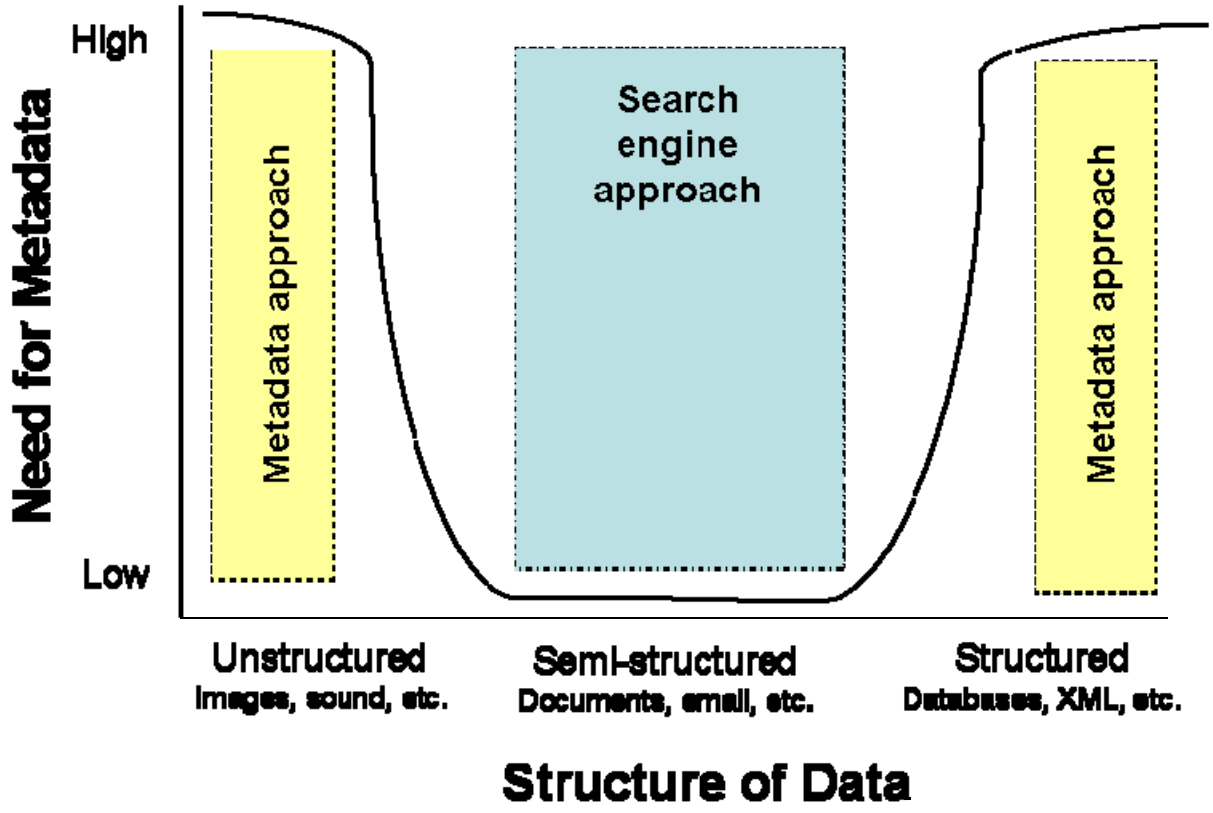
- For unstructured and semi-structured text (e.g., websites, email, etc.), commercial search technology is sufficient. No advance preparation of content required
- Human-generated controlled vocabularies can, in certain circumstances, improve the precision and recall of domain-specific search applications, especially those pertaining to highly technical subject areas.
- For databases and other structured data sources (e.g., geospatial, biometric, etc.), some advance preparation of content is generally required; specifically, mapping

[5] NARA response, page 1

between different data schemas or metadata sets to "join" related fields from different databases.

- Multimedia collections, including sound and video, generally require significant advance preparation of content to enable efficient and effective information retrieval and sharing, typically in the form of metadata tagging, manual categorization, and cataloging. Fully automated indexing and searching of multimedia files is an active area of research, though, so this could change in the next few years.
- Collections containing sensitive information, such as classified information, generally require significant advance preparation of content to ensure authenticity and reliability in the information retrieval and sharing process, the availability of sophisticated search technology notwithstanding.

The Information Sharing Continuum figure, shown below, places current information sharing approaches (i.e., metadata-based approaches vs. search-based approaches) on a continuum from unstructured data (e.g., voice, video, and other multimedia) to structured data (e.g., databases, XML). Search technology currently available cannot readily index unstructured data products (i.e., voice, video, and other multi-media), so the manual creation of indexes, taxonomies, or metadata is necessary. As we move into the semi-structured realm, search technology has matured to the point where search engines can automatically derive meaning from semi-structured data products (e.g., documents, email, hypertext, etc.) without any advance preparation of content. As you move down the continuum, structured data products are given meaning mostly by the metadata (e.g., data models, schemas, etc.) that define them.



## 5. CONCLUSION

To summarize the findings of this study, GSA surveyed a variety of information retrieval experts and practitioners from industry, government, and academia. The respondents favored search-based approaches requiring no advance preparation of content (56%) over metadata–based approaches requiring some advance preparation of content (30%) or significant advance preparation of content (14%). Therefore, one can conclude from this study and other available literature including *The Search* [6] that, with respect to disseminating Federal information to the public-at-large, publishing directly to the Internet all agency information intended for public use and thereby exposing it to freely available or other search functions is the most cost-beneficial way to enable the efficient and effective retrieval and sharing of government information.

At the same time, one can also conclude there are times when advance preparation is more cost-beneficial and even necessary. As an organization moves from a passive or "casual" access model to a more active or "formal" one or from providing web pages and text to geospatial, multi-media or structured databases, the need for advance preparation, including through automated or manual creation of indexes, taxonomies, or metadata tagging, begins to become apparent.

This need however is not revealed by a distinct threshold or "bright line" between no advance preparation and a specific type of preparation. Rather, it is illustrated on a continuum where as complexity and formality increases, so too does the business case (ROI) for advance preparation. In short, one cannot paint with a broad brush when advance preparation must occur. The need is determined by information producers and users on a case-by-case basis. Clearly it is not remotely possible for Federal agencies to engage in comprehensive interaction with all members of the public-at-large. Therefore, again, this study supports as a general principle, direct publication to the Internet is the best way to promote general dissemination and sharing of government information.

[6] Battelle, John. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture.* New York, NY: Portfolio, 2005.

# 6. APPENDIX A: SUMMARY OF RESPONDENTS

This appendix provides list of EEIRS RFI respondents by organizational type (i.e., Industry, Government, or Academia):

## Industry

- Adobe
- Autonomy Inc.
- Barquin International
- Blue Angel Technologies / SAIC
- Boeing
- Broadband Technology Group / Smart Data Strategies / Vertical Horizons One
- CACI Enterprise Solutions Inc. / Endeca
- Cirilab, Inc.
- Computer Sciences Corporation/ Google/ Cherry Road Tech
- Content Analyst
- Corpora Software
- Cyber-Ark Software
- DDMS Technologies
- Deep Web Technologies
- Dymond and Associates LLC
- Fast Search and Transfer
- Google
- Harris Corporation
- HNC Software LLC
- i2, Inc.
- Information Builders
- Inxight Federal Systems Group
- IT.com
- Janya Inc
- Kyos Systems, Inc.
- LexisNexis
- Lockheed Martin
- Mark Logic Corporation
- McDonald Bradley, Inc
- MDY
- MetaCarta
- MetaMatrix
- Microsoft & Vivisimo
- Netzero
- QRC-Macro International
- SAIC
- SchemaLogic, Inc.
- Software & Information Industry Association
- Sun Microsystems

- Synteractive Corporation
- Venecal Global Systems
- WhamTech Inc

### Government

- NARA
- The National Academies
- Department of Energy OSTI

### Academia

- Stanford University
- USC Information Sciences Institute

# 7. Appendix B: Detailed analysis of responses by RFI question

This appendix provides detailed analysis for each of the capabilities listed in Table 1. Specifically, it provides a roll-up of how many respondents presented an approach requiring no, some, or significant advance preparation of content. It also provides a breakdown of specific approaches for each specific RFI question, categorizing each response by specific type of approach.

## 7.1. Multiple Jurisdictions/Stakeholders

This capability represents the ability to search, discover, retrieve, and share information across legal/jurisdictional/organizational boundaries. It is represented in the RFI as questions 3.1.1.a and b.

3.1.1.a: Using the above scenarios as context, describe your overall approach/vision for enabling the searching, discovery, retrieval, and sharing of information across legal/jurisdictional/organizational boundaries.

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring some advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 6
- Requires some advance preparation of content: 13
- Requires significant advance preparation of content: 2
- Not applicable: 26

**Approach Trends**

In addressing this capability, respondents were mixed in their specific approaches. Many advocated simply using web spidering to get at content across organizational boundaries. Others favored a federated query approach, where each organization maintains its own search application and a "meta-search" application aggregates results from those applications. Another popular approach involved integration of distributed systems through the explicit mapping of metadata and data schemas, controlled via a middleware or registry product. A few respondents asserted that explicit management of content was necessary to implement this capability. Specific numbers include:

- Explicit content management: 2
- Web spidering: 6
- Data schema/metadata mapping: 7
- Federated query: 6
- No response/inadequate response: 26

3.1.1.b: Using the above scenarios as context, describe your overall approach/vision for enabling the searching, discovery, retrieval, and sharing of information among stakeholders who are dispersed geographically.

| Analysis |
| --- |
| The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:<br><br>• Requires no advance preparation of content: 12<br>• Requires some advance preparation of content: 8<br>• Requires significant advance preparation of content: 0<br>• Not applicable: 27 |
| **Approach Trends** |
| In implementing this capability, the majority of respondents stated that simply using standards-based access and delivery channels was sufficient to allow geographically dispersed stakeholders to access information. Others advocated web spidering of content into a centralized index, a federated query approach, or a middleware – centric data schema mapping approach. Specific numbers include:<br><br>• Standards-based access/delivery channel (e.g., web browsers and servers): 8<br>• Web spidering: 4<br>• Data schema/metadata mapping: 4<br>• Federated query: 4<br>• No response/inadequate response: 27 |

### 7.2. Multiple data sources

This capability represents the ability to search, discover, retrieve, and share information across databases, websites, repositories, record management systems, etc. It is represented in the RFI as question 3.1.1.c.

3.1.1.c: Using the above scenarios as context, describe your overall approach/vision for enabling the searching, discovery, retrieval, and sharing of information across multiple physical data sources, including databases, websites, repositories, record management systems, and other data assets.

| Analysis |
| --- |
| The majority of practitioners responding to this question advocated an approach requiring some advance preparation of content. Specific numbers include:<br><br>• Requires no advance preparation of content: 9<br>• Requires some advance preparation of content: 12<br>• Requires significant advance preparation of content: 0<br>• Not applicable: 26 |
| **Approach Trends** |

Respondents were split on how to best implement this capability. Many advocated a middleware – centric data schema mapping approach. Others advocated the spidering of content into a centralized index. Specific numbers include:

- Standards-based access/delivery channel (e.g., web browsers and servers): 3
- Web spidering: 6
- Configurable adaptors/agents: 2
- Data schema/metadata mapping: 7
- Federated query: 3
- No response/inadequate response: 26

### 7.3. Multiple data types

This capability represents the ability to search, discover, retrieve, and share multiple data types (e.g., structured, unstructured, tabular, multi-media, email, geospatial, biometric). It is represented in the RFI as question 3.1.1.d.

3.1.1.d: Using the above scenarios as context, describe your overall approach/vision for enabling the searching, discovery, retrieval, and sharing of information comprising many different formats, including documents, email, multimedia (video and sound), geospatial data, structural/tabular data (e.g., fields and records), biometric data (e.g. fingerprints, etc.), and others.

#### Analysis

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 15
- Requires some advance preparation of content: 6
- Requires significant advance preparation of content: 1
- Not applicable: 25

#### Approach Trends

The vast majority of respondents favored an automated indexing approach to implementing this capability (i.e., converting the information resource's content to a neutral text format, then building a word index for efficient searching). Furthermore, many vendors favoring this approach had already implemented products indexing an impressive array of document formats. Specific numbers include:

- Standards-based access/delivery channel (e.g., web browsers and servers): 2
- Explicit content management/metadata tagging: 1
- Automated document indexing: 13
- Configurable adaptors/agents: 2
- Data schema/metadata mapping: 4
- No response/inadequate response: 25

### 7.4. Data aggregation & integration

This capability represents the ability to aggregate information in a value added manner (i.e., combining different data types into useful work products). It is represented in the RFI as question 3.2.1.

3.2.1 Please describe your approach for performing the aggregation and integration of information resources comprising many different formats into value added information products where the value of the end product exceeds the value of the sum of its parts. Data types include documents, email, multimedia (video and sound), geospatial data, structural/tabular data (i.e., fields and records), biometric data and others. Examples of complex, value added knowledge products include:

a. A weather map, combining geospatial data and meteorological data
b. A law enforcement case file linking case notes with related documents and related database records
c. A crime analysis map, combining geospatial data and crime statistics

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 12
- Requires some advance preparation of content: 11
- Requires significant advance preparation of content: 1
- Not applicable: 23

**Approach Trends**

The most popular approach to performing the aggregation described by this question was a federated query approach, where a meta-search service issues requests to other autonomous search services and aggregates the responses into a single view for the user. Other more automated approaches involved the creation of "virtual aggregate records" via indexing and statistical analysis to "connect the dots" between related data harvested from different sources. Specific numbers include:

- Federated Query: 11
- Virtual Aggregate Records: 8
- Visualization: 4
- Container/case files: 1
- No response: 23

### 7.5. Notification

This capability represents the ability to notify users when an information resource has been added to, altered or removed from the shared computing environment. It is represented in the RFI as question 3.2.10.

3.2.10: Please describe your approach to notifying authorized users when an information resource of interest has been added to, altered or removed from the network or other shared computing environment (e.g., based upon one's most recent past search).

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 14
- Requires some advance preparation of content: 11
- Requires significant advance preparation of content: 4
- Not applicable: 18

**Approach Trends**

Many respondents advocated a "saved search" approach, where a user would save a search query. The system would then execute the saved search on a schedule, and notify the user of any new or changed documents via email, pager, or RSS feed. Other approaches include establishing notification rules (e.g., email me when a particular user adds a document); notification by example (e.g., notify me when a document like this document is added to the system); and metadata based notification (e.g., notify me when a document matching this taxonomy node is added to the system). A particularly innovative approach was one using a predictive model, i.e., users would automatically be notified of content that might interest them based on their past activity in the system. Specific numbers include:

- Saved searches: 12
- Rules-based notification: 9
- Notification by example (e.g., user provides document examples): 2
- Metadata based notification: 4
- Predictive model (e.g. train system to notify users based on past activities): 2
- No response/inadequate response: 18

### 7.6. Precision & recall

This capability describes approaches for calculating relevance, locating relevant information not containing the original query, indexing the "deep web", etc. It is represented in the RFI as questions 3.2.3, 3.2.4, 3.2.5, and 3.2.6.

3.2.3: Please describe your approach for executing a search that includes in its results those information resources that are relevant yet do not contain any of the terms in the original query.

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 19
- Requires some advance preparation of content: 9
- Requires significant advance preparation of content: 3
- Not applicable: 16

**Approach Trends**

The vast majority of respondents advocated a query expansion approach using synonyms of the original search terms. Many of these respondents pointed out that this could be fully automated using statistical approaches such as latent semantic indexing (LSI) or similar technique. Others advocated the use of hand-crafted thesauri to perform the query expansion. Specific numbers include:

- Query expansion with synonyms (fully automated): 19
- Query expansion with synonyms (controlled vocabulary): 9
- Document metadata harvesting: 2
- User feedback: 1
- No response/inadequate response: 16

3.2.4: Please describe your approach to providing comprehensive search coverage of all available information resources, and advising users where coverage gaps might exist (e.g., "deep web" or "hidden web").

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring some advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 9
- Requires some advance preparation of content: 13
- Requires significant advance preparation of content: 2
- Not applicable: 23

**Approach Trends**

Many respondents, including many search vendors, acknowledged that this capability is difficult to implement without at least some human intervention. One of the more popular approaches involved explicitly mapping in data sources not easily reachable via typical web search engine spiders. Specific numbers include:

- Fully automated (e.g., web spidering): 9
- Federated query: 4
- Configurable agents or adaptors: 3
- Data schema/metadata mapping: 6

- Custom programming of interfaces: 1
- Partnering/OEM: 3
- Human intervention: 1
- No response/inadequate response: 20

3.2.5: Please describe your approach to providing search query refinement and disambiguation (i.e., recommending alternate queries based on the content of the original query).

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 17
- Requires some advance preparation of content: 8
- Requires significant advance preparation of content: 5
- Not applicable: 17

**Approach Trends**

The vast majority of respondents advocated a query refinement approach using synonyms of the original search terms. Many of these respondents pointed out that this could be fully automated using statistical approaches such as latent semantic indexing (LSI) or similar technique. Others advocated the use of hand-crafted thesauri to perform the query refinement. Specific numbers include:

- Query refinement with synonyms (fully automated): 17
- Query refinement with synonyms (controlled vocabulary): 8
- Document metadata harvesting: 3
- User feedback: 2
- No response/inadequate response: 17

3.2.6: Please describe your approach to calculating relevance when sorting search results. Does your approach use a paid inclusion option and if so are paid inclusion results segregated from typical results?

**Analysis**

The vast majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 20
- Requires some advance preparation of content: 4
- Requires significant advance preparation of content: 4
- Not applicable: 19

**Approach Trends**

Most respondents advocated an approach to calculating relevance involving use of sophisticated text analytics, such as weighting and proximity of terms, and various natural language processing techniques such as parts of speech tagging. Others added popularity analysis algorithms to the mix, such as the Google PageRank algorithm. Other less frequently used approaches took into account any metadata present with the document when determining feedback. Specific numbers include:

- Word frequency only: 3
- Sophisticated text analytics (e.g., weighting, proximity, NLP): 12
- Metadata only: 3
- Word frequency with other metadata: 4
- Link/citation analysis with text analytics: 5
- User feedback: 1
- No response/inadequate response: 19

### 7.7. Data quality, authenticity, and archival

This capability describes approaches to ascertaining the quality, authenticity, durability, and longevity of an information resource. It is represented in the RFI as questions 3.2.7, 3.2.8, and 3.2.9.

3.2.7: Please describe your approach for assisting users in determining the quality or authenticity of information resources.

#### Analysis

Approaches proposed by respondents were split between those requiring no advance preparation of content, and those requiring significant advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 14
- Requires some advance preparation of content: 0
- Requires significant advance preparation of content: 15
- Not applicable: 18

#### Approach Trends

Respondents were divided into two camps on this question, with the search crowd advocating automated approaches such as statistical or link analysis to determine the "pedigree" of an information resource, and the metadata crowd advocating the need for explicit content management to ensure the quality and authenticity of a document. Specific numbers include:

- Explicit authentication (e.g., digital signatures): 3
- User feedback: 3
- Relevance algorithms: 5
- Explicit content management/metadata tagging (e.g., tagged with author's name): 9
- Content popularity (e.g., link/citation analysis, Kleinberg algorithm): 4
- Statistical approach (e.g., forensic writer identification techniques): 5
- No response/inadequate response: 18

3.2.8: Please describe your approach in ascertaining the durability of a given electronic resource (i.e., the likelihood of the resource continuing to be accessible at a given location indefinitely).

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 14
- Requires some advance preparation of content: 0
- Requires significant advance preparation of content: 10
- Not applicable: 23

**Approach Trends**

Caching of content during the indexing process was the most popular approach advocated by respondents to implement this capability, as it provided the ability to present a copy of the information resource to the end user potentially indefinitely, even if the source system was no longer available. Specific numbers include:

- Explicit content management/metadata tagging (e.g., tagged with author's name): 6
- Caching and other automated archiving: 9
- Human intervention: 4
- Predictive model (e.g. the long-term availability of information at a particular data source based on past observations): 5
- No response/inadequate response: 23

3.2.9: Please describe your approach to identifying the likelihood the source information located will continue to be available over the long term (e.g., through archival). Alternatively, is this only possible through advance preparation by the owner/producer and if so is their an automated way to do this or does it demand human intervention?

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 17
- Requires some advance preparation of content: 0
- Requires significant advance preparation of content: 7
- Not applicable: 23

**Approach Trends**

Again, caching was the most popular approach to implementing this capability. Specific

numbers include:

- Explicit content management/metadata tagging (e.g., tagged with author's name): 6
- Caching and other automated archiving: 13
- Human intervention: 1
- Predictive model (e.g. the long-term availability of information at a particular data source based on past observations): 4
- No response/inadequate response: 23

### 7.8. Interoperability

This capability describes standards implemented and/or supported by the proposed approach. It is represented in the RFI as questions 3.1.1.e and 3.2.13.

3.1.1.e: Using the above scenarios as context, describe your overall approach/vision for enabling the searching, discovery, retrieval, and sharing of information leveraging existing capabilities found across the Federal Government

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring some advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 2
- Requires some advance preparation of content: 14
- Requires significant advance preparation of content: 1
- Not applicable: 30

**Approach Trends**

Most respondents elected not to reply to this question. The ones that did seemed to favor a federated or service-oriented approach to leveraging existing Federal government information retrieval capabilities. Specific numbers include:

- Standards-based access/delivery channel (e.g., web browsers and servers): 2
- XML web services: 2
- ISO 23950: 1
- Generic federated search/meta-search: 4
- Data schema/metadata mapping: 5
- Configurable adaptors/agents: 2
- Custom interfaces: 1
- No response/inadequate response: 30

3.2.13: Please describe the interoperability standards implemented/supported by your approach in the following areas. Alternatively, you may also explain why you believe the following are not necessary or cost effective:

- Data Exchange: This category defines the set of standards supporting the interchange of information between multiple systems or applications. Examples include XML, RDF, HTML, PDF, and others.
- Service Transport: This category consists of the protocols and standards defining the format and structure of data and information when accessed either from a directory or exchanged through communications. Examples include HTTP, SOAP, LDAP, WSDL, UDDI, ISO 23950, and others.
- Metadata interoperability: This category defines the set of standards supporting the interchange of metadata between middleware, registries, and modeling/development tools. Examples include XSD, OWL, XTM, UML/XMI, ISO 11179, XSLT, and others.

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 29
- Requires some advance preparation of content: 6
- Requires significant advance preparation of content: 2
- Not applicable: 10

**Approach Trends**

Most respondents advocated use of XML for data interoperability, a standard almost universally supported in databases and content authoring tools. Other respondents advocated a federated query approach to interoperability, using the ISO 23950 search interoperability standard. A few advocated other standards, such as ISO 11179 and RDF. Specific numbers include:

- XML: 29
- ISO 23950: 6
- ISO 11179: 1
- RDF: 1
- No response/inadequate response: 10

### 7.9. Information extraction and visualization

This capability represents the ability to rapidly discover relevant facts buried in large volumes of data, organizing/presenting/visualizing large data sets in an understandable, user-friendly manner. It is represented in the RFI as questions 3.2.2, 3.2.11, and 3.2.12.

3.2.2: Please describe your approach to organizing/presenting/visualizing large data sets in an understandable, user-friendly manner.

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 11
- Requires some advance preparation of content: 5

- Requires significant advance preparation of content: 4
- Not applicable: 27

**Approach Trends**

There were many different approaches proposed to organizing, presenting, and visualizing large data sets including clustering (organizing information resources into topic "clusters"), sorted search result lists, providing concise summaries of information resources via text summarization algorithms, and presenting search results graphically using concept maps, diagrams, and geospatial rendering. Specific numbers include:

- Concept or topic clustering (fully automated): 4
- Concept or topic clustering (controlled vocabulary): 5
- Sorted list (e.g., sorted by relevance): 5
- Text summarization: 2
- Graphical approach (e.g., concept maps, geospatial rendering): 5
- Manual creation of interfaces (e.g., custom programming): 4
- Partnering/OEM: 6
- No response/inadequate response: 16

3.2.11: Please describe your approach to implementing the extraction of specific factual information (e.g. people, organizations, locations, dates, concepts, etc.) from large collections of unstructured resources (e.g., text or multimedia).

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 11
- Requires some advance preparation of content: 7
- Requires significant advance preparation of content: 2
- Not applicable: 27

**Approach Trends**

Most respondents preferred a fully automated approach to "entity" (fact) extraction, using some type of pattern matching (e.g., Regular Expressions, NLP). Others advocated use of entity extraction aided by controlled vocabularies and other metadata. Many respondents did not support this capability, preferring instead to outsource this capability to another vendor. Specific numbers include:

- Fully automated entity extraction: 11
- Entity extraction aided by controlled vocabularies and other metadata: 7
- Manual entity extraction only: 2
- Partnering/OEM: 8
- No response/inadequate response: 19

3.2.12: Please describe your approach for discovering non-obvious yet potentially useful knowledge from large collections of unstructured resources (e.g., text or multimedia).

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. Specific numbers include:

- Requires no advance preparation of content: 19
- Requires some advance preparation of content: 6
- Requires significant advance preparation of content: 0
- Not applicable: 22

**Approach Trends**

Most respondents preferred a fully automated "knowledge discovery" approach using a pattern-matching heuristic such as co-occurrence of concepts and entities, link analysis, clustering, etc. Specific numbers include:

- Partnering/OEM: 3
- Pattern matching (fully automated): 16
- Pattern matching (metadata assisted): 5
- Guided Navigation: 1
- Keyword search: 3
- No response/inadequate response: 19

### 7.10. Benefits of advanced preparation of information

This capability describes advantages gained through advance preparation of content for search and retrieval (i.e., metadata tagging). It is represented in the RFI as question 3.1.2.

3.1.2: To what extent does your approach require explicit work performed in advance to prepare content for retrieval (e.g., metadata tagging, cataloging, etc.)? To what extent does this time investment improve the precision and retrieval of information in a cost effective way? Can you approximate the relative performance increase over no advance preparation?

**Analysis**

The majority of practitioners responding to this question advocated an approach requiring no advance preparation of content. These answers were consistent with responses to other RFI questions pertaining to specific capabilities. Specific numbers include:

- Requires no advance preparation of content: 18
- Requires some advance preparation of content: 8
- Requires significant advance preparation of content: 4
- Not applicable: 17