

IQARIS: A TOOL FOR THE INTELLIGENT QUERYING, ANALYSIS, AND RETRIEVAL FROM INFORMATION SYSTEMS

John R. Hummel and Robert B. Silver¹
Advanced Computer Applications Center
Decision and Information Sciences Division
Argonne National Laboratory
9700 S. Cass Avenue/DIS-900
Argonne, IL 60439-4832

The submitted manuscript has been created by the University of Chicago as Operator of Argonne National Laboratory ("Argonne") under Contract No. W-31-109-ENG-38 with the U.S. Department of Energy. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, non-exclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public and display publicly, by or on behalf of the Government.

KEYWORDS: Information Technology, Information Context, Intelligent Searching

ABSTRACT: *Information glut is one of the primary characteristics of the electronic age. Managing such large volumes of information (e.g., keeping track of the types, where they are, their relationships, who controls them, etc.) can be done efficiently with an intelligent, user-oriented information management system. The purpose of this paper is to describe a concept for managing information resources based on an intelligent information technology system developed by the Argonne National Laboratory for managing digital libraries. The Argonne system, Intelligent Query (IQ), enables users to query digital libraries and view the holdings that match the query from different perspectives.*

1. Introduction

Information glut is one of the primary characteristics of the electronic age. Data can be generated in large volumes in as little time as that required to press the "Enter" key on a keyboard. Although the problem of information glut is usually associated with the print world, it is also a problem in the physical sciences.

Managing such large volumes of information (e.g., keeping track of the types, where they are, their relationships, who controls them, etc.) can be done efficiently with an intelligent, user-oriented information management system. The purpose of this paper is to describe a concept for managing information resources based on an intelligent information technology system developed by the Argonne National Laboratory for managing digital libraries. The Argonne system, Intelligent Query (IQ), enables users to query digital libraries and view the holdings that match the query from different perspectives. The original IQ concept will be further extended in scope to encompass

document and non-document holdings to provide a system for the Intelligent Querying, Analysis, and Retrieval from Information Systems (IQARIS).

2. Overview of the Intelligent Query System

The original IQ system was developed for use with a digital library developed by the U. S. Department of Energy's (DOE's) Environmental Safety and Health (ESH) group. IQ used a commercial search engine that could be enhanced using an electronic thesaurus. The search engine returned the "hits" on the query, based on proprietary algorithms within the search engine, that were ranked by relevance. The query hits could be viewed from different perspectives using a set of visualization tools developed by Argonne. Visualizing the results of a query can provide users with a faster method of assessing if a document is relevant to their needs as compared to the conventional method of just returning the name of the document and then forcing users to manually review each document.

¹Also affiliated with Wayne State University School of Medicine.

One form of visualization is called the “Document View” and shows the distribution of query hits within documents. An example is shown in Figure 1. Being able to see where a query hit occurred within a document can assist a user in assessing the relevance of a document to one’s specific context. For example, a query hit returned from the main body of a document may make the document more relevant than if the hit came from a caption or citation reference.

Another way to visualize the query returns is to use a “Geographic” view that can show on a map the location a document refers to or where it was generated (i.e., published). A third way to visualize the

results, the “Timeline” view, shows when documents were written. Finally, the interrelationships between documents can be an important indicator of the relevance of a document to a user. In the DOE ESH library, a number of documents were generated as a direct result of another document. For example, when an Environmental Assessment (EA) is performed for a given location, a variety of additional documents could also be generated as a result of regulatory requirements, such as a Finding of No Significant Impact (FONSI) or a Record of Decision (ROD). With the IQ “Type View” tool, one can see the interrelationships between documents, as shown in Figure 2.

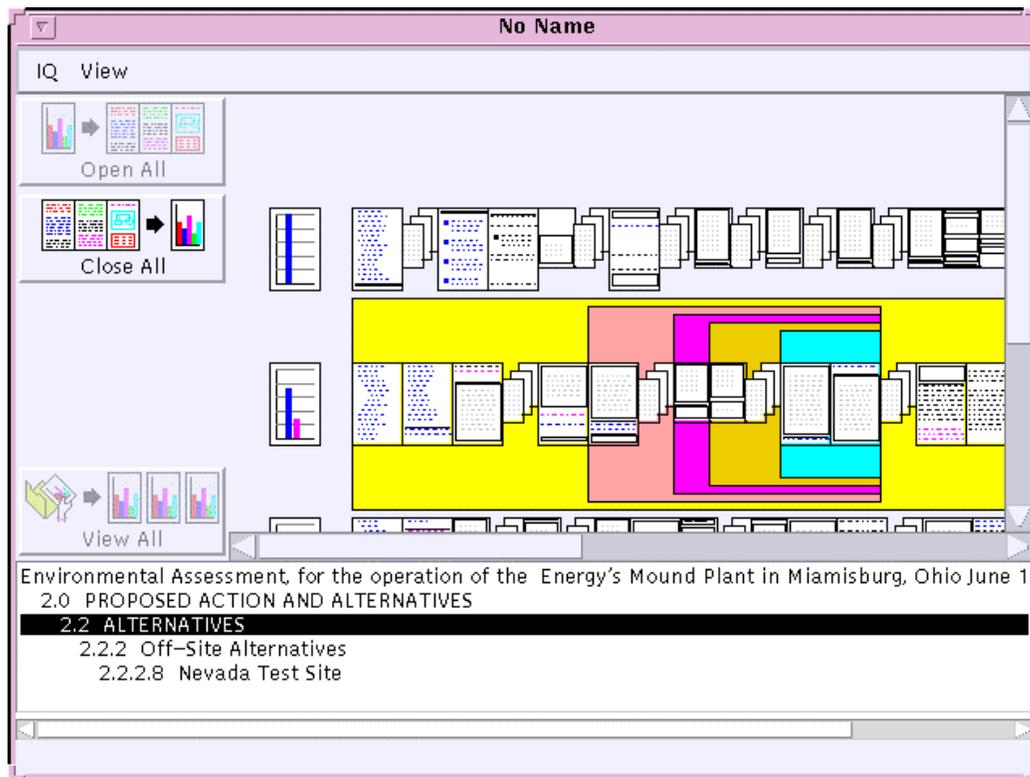


Figure 1. An Example of the IQ “Document View” Showing the Distribution of Query Returns in Documents. A Color-Code Histogram Shows an Overview of the Returns in Each Document. The Page “Images” to the Right of the Histogram Show the Location of the Individual Query Returns Located within the Document. The Scroll Box at the Bottom Details the Specific Sections Where the Query Returns Were Found.

3. Extending the IQ Vision

The original IQ system was developed to support an electronic document library. We are extending the IQ concept to encompass a larger range of information resources that will include documents (in multiple formats), data, images and graphics, and models. To support the extended vision, we will develop additional visualization tools to display and analyze the information resources. We will be demonstrating the benefits of the IQARIS concept using a variety of information resources that have been collected in support of programs involving breast cancer research and those focused on understanding the chemical processes by which cells divide.

3.1 Describing the Information Resources

For any information resource to be of value to the research community, it must be able to be characterized in an easy to

understand manner. This can be accomplished by establishing “metadata” protocols that can be used to describe the information resources. Metadata, or data about the resources in the electronic library, can be thought of as the electronic equivalent of the information found on a conventional “card catalog” entry.

We will use the metadata protocols developed by the Dublin Core Metadata program (<http://dublincore.org>) with IQARIS. The Dublin Core program was developed by an international group of information specialists with the goal of developing a simple content description model for electronic information resources. The Dublin Core metadata elements, which are shown in Table 1, are generally based on commonly understood terms and have a complexity similar to that of a library catalog card.

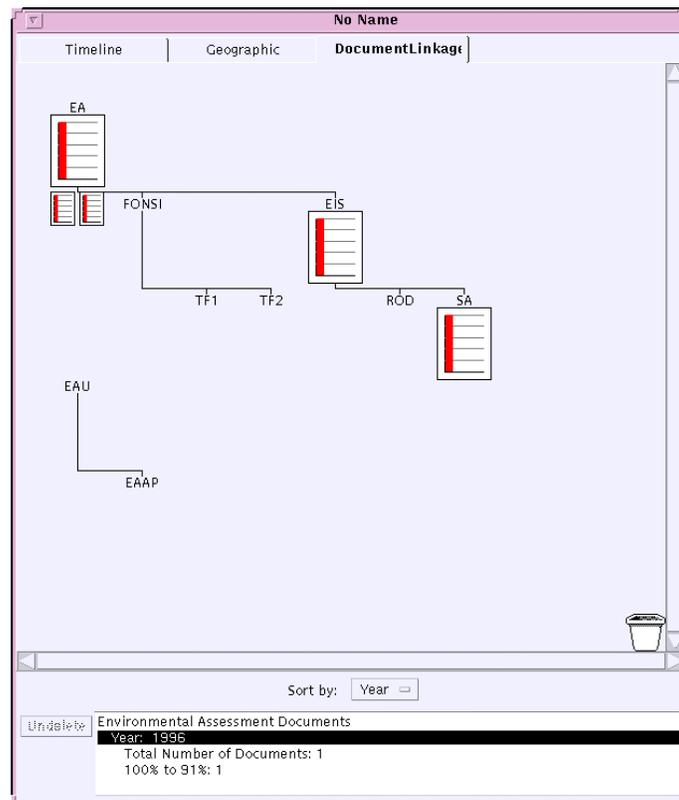


Figure 2. The IQ “TypeView” That Shows the Interrelationships Between Documents Matching a User-Generated Query.

Table 1. A Summary of the Dublin Core Metadata Elements.

METADATA ELEMENT	DESCRIPTION
Title	The name given to the resource by the Creator or Publisher.
Author or Creator	The person or organization primarily responsible for creating the intellectual content of the resources.
Subject and Keywords	The topic of the resource.
Description	A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions of visual or other resources.
Publisher	The entity responsible for making the resource available in its present form, such as a publishing house, university department, or a corporate entity.
Other Contributor	A person or organization not specified as an Author or Creator who has made significant intellectual contributions to the resource but whose contribution is secondary to any person or organization specified in the Author or Creator element.
Date	The date the resource was made available in its present form.
Resource Type	The category of the resource, such as home page, novel, poem, working paper, technical report, etc. (A set of proposed standard types is available.)
Format	The data format of the resource used to identify the software and possible hardware that might be needed to display or operate the resource.
Resource Identifier	A string or number used to uniquely identify the resource. Examples are URLs and ISBNs.
Source	A string or number used to uniquely identify the work from which this resource is derived. For example, a PDF version of a book may have an ISBN for the physical book from which the PDF version was derived.
Language	Language(s) of the intellectual content of the resource.
Relation*	The relationship of this resource to other resources. An example is an individual image from a larger document.
Coverage*	The spatial and/or temporal characteristics of the resource.
Rights Management	A link to a copyright notice, to a rights-management statement, or to a service that would provide information about terms of access to the resource.

*Field is experimental.

The basic Dublin Core metadata elements will be used as the basis for developing the IQARIS metadata protocols and will be enhanced to meet the needs of this application. For example, we would propose to add a new “Location of

Resource” element that would be used to specify the location where the particular resource is found. This element will be required because the information resources will most likely be distributed across different facilities. We will also use the

Dublin Core “Coverage” metadata element to describe any relevant spatial and temporal characteristics of a resource. We will add subelements to this metadata field to describe coverage issues such as spatial extents (2-D or 3-D), spatial resolution and coverage, temporal extents (point, sampled, or continuum), and temporal resolution and coverage.

In addition, we will review metadata protocols in other subject domains to assess if additional metadata elements are required. For example, some of the data that could become a part of the IQARIS holdings are from the geophysical domain and would fall under geophysical metadata standard requirements mandated by the U.S. government.

The Dublin Core metadata elements are especially suitable for electronic and other multi-media resources and can encompass “typical” electronic resources, such as documents, but can also incorporate other electronic resources such as data, images, and computer programs. We will use the “Resource Type” metadata element to differentiate the information resources at the highest level in terms of “types” like computer programs, documents, data, images, or commercial software applications files (e.g., Excel™ spreadsheets or PowerPoint™ files.) We will then use a set of easily recognized icons when displaying the different resource types in the holdings. Figure 3 shows a proposed set of icons that could be used to display the resource types by IQARIS. Each resource type could be subdivided into a set of subcategories if warranted. For example, the “Document” resource type could be subdivided into

additional categories, each with an icon associated.

3.2 Searching and Analyzing the Information Resources

We will provide tools to sort, search, and visualize the information holdings from different viewing perspectives. As with the original IQ system, IQARIS will combine conventional search engine technologies with “smart” visualization tools to analyze and access the information resources.

The primary searching tools will be performed using the metadata for each information resource. For example, the user will be able to sort the resources in terms of factors such as their location, the types they represent, their subject matters, dates created, time periods internally represented, relationships to other documents, etc. Figure 4 shows an example of how the geographical distribution of resource types could be displayed. In this example, the assumed sorting would have been performed on the Resource Type and Location of Resource metadata elements.

We will also provide tools to enable the user to “drill down” into a given view. In the example shown in Figure 4, one could expand the view of the document holdings found at Argonne by selecting the Document icon. The result would be a popup menu that would show the number of documents found at Argonne in each of the different document subtypes registered with Argonne. One could continue to drill down within the subtypes and, at the lowest level, examine the individual card catalog entries of each document.



Figure 3. A Set of Icons That Could Be Used to Describe the Resource Types That Would Be Managed Under the IQARIS Concept.

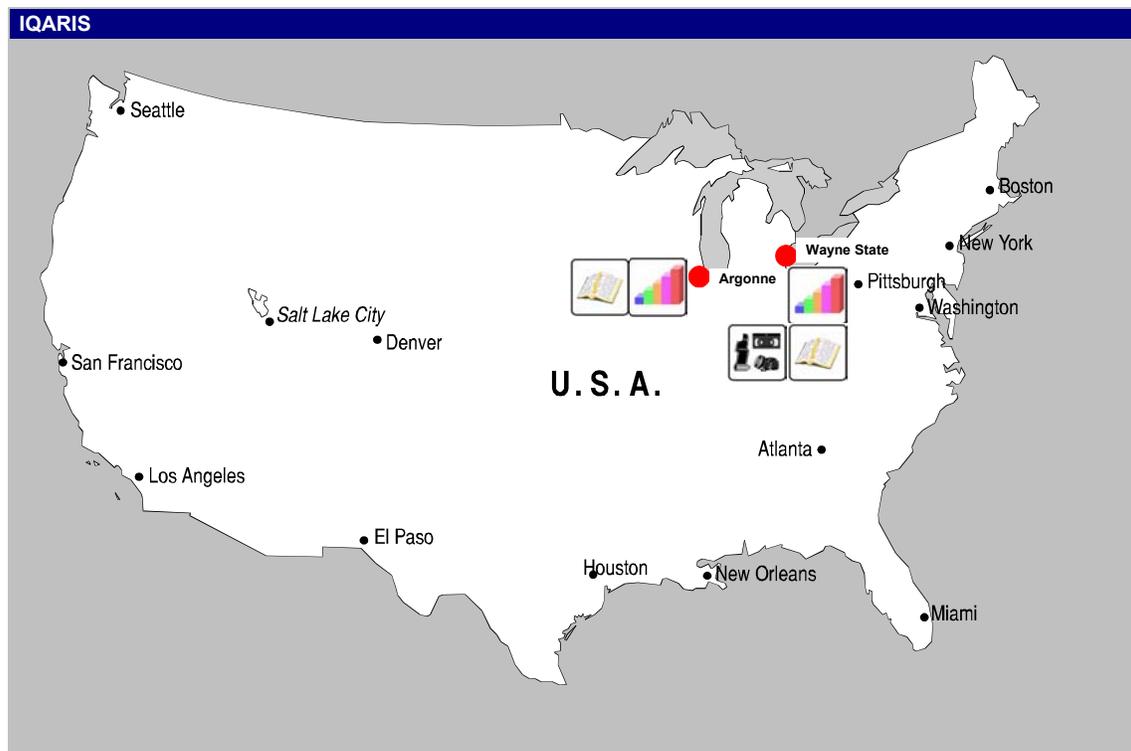


Figure 4. An Example of How the IQARIS Resource Type Icons Could Be Used to Show the Types of Information Resources Held at Different Locations.

When it is appropriate, users will be able to perform additional keyword searches of the information resources themselves. For example, we would propose to enable the viewing of search term “hits” within documents in the same manner as implemented in the original IQ system.

4. Use of IQARIS in Breast Cancer Research Studies

We are proposing to use IQARIS in a study involving the Breast Cancer Center of Excellence at the Wayne State University Medical School. The Bioinformatics needs of the Breast Cancer Center of Excellence will pose exciting challenges for the collaborating investigators. Large amounts of data and metadata will be generated from many different types of observations performed at various locations. The experimental data will come from numerous types of instruments, each with its own unique characteristics and file formats. Individual investigators will perform analyses on their own data using methods that are tuned to their needs. In addition, others will, within the project, need to perform complementary analyses on those same data. The task of organizing and utilizing these data and information sources is, on the surface, daunting. It is for this type

of application that IQARIS is intended and best suited.

IQARIS will coordinate data and information management services and queries from the various researchers in the project. In addition, IQARIS will serve as the search engine for performing context-based searches of data and literature both within (e.g., Cardiff, Hughes, Ross, Womble) and outside (e.g., National Library of Medicine, National Cancer Institute, Centers for Disease Control and Prevention, etc.) the project group. Questions posed through IQARIS will be analyzed by the Thesaurus to determine how to best match the query with available data and information retrieval options. In the process, the Thesaurus will also learn new terms and phrases for each query from the investigators, thus facilitating cross-disciplinary queries and interactions, i.e., as a translator among naive and educated “cell biology,” “MR Imaging,” “Chemistry” and “Bioinformatics” oriented users.

5. Acknowledgments

This work is sponsored by the U.S. Department of Energy under contract W-31-109-ENG-38.