NSF/NIJ Symposium on Intelligence and Security Informatics
Full Paper Submission

Secure Information Sharing and Information Retrieval Infrastructure with GridIR

Dr. Gregory B. Newby[1]
Arctic Region Supercomputing Center
University of Alaska, Fairbanks

Kevin Gamiel, MCNC
Nassib Nassar, Etymon Systems, Inc.

Abstract

This work describes the emerging standard for information retrieval on computational grids, GridIR. GridIR is based on the work of the Global Grid Forum and offers a security model for end-to-end data channel encryption, system authentication, and logging. GridIR implements a multi-tiered security model at the collection, query and datum level. Unlike monolithic search engines and customized small-scale systems, GridIR provides a standard method for federating data sets with multiple data types. Each of the data sets may be part of a virtual organization in which policy decisions dictate what data items may be shared.

Introduction

Information retrieval (IR) refers to the systems for identifying and presenting documents relevant to human information needs. In the information intensive environments found in national security settings, a variety of IR systems and related systems (e.g., for information display, merger and manipulation) are engaged to address information problems.

Broadly stated, many of the information activities of national security agencies may be conceived of as IR activities, requiring IR systems. However, IR is usually taken to be directed more narrowly at matching statements of information need (a.k.a., *queries*) to documents, including document extracts or synthesized documents. Even in this more specific definition, the National Science Foundation's Data Mining Research program's goal is given in terms consistent with core IR activities:

> Find actionable intelligence from massive data sets and diverse sources, represent it in a way that allows rapid understanding and deliver it to the people who need to know.[2]

Today's average home or office computer user might think that the most sophisticated IR systems of today – and perhaps the only ones really needed – are Web search engines such as Google. More sophisticated information seekers realize there are many thousands of IR systems with varying data sets and search capabilities. These systems

[1] Address effective May 2003. Currently at the School of Information and Library Science, University of North Carolina at Chapel Hill. Email: gbnewby@ils.unc.edu Web: http://ils.unc.edu/gbnewby
[2] Art Becker speaking at a KDD research meeting, September 3 2002.

range from search boxes on Web sites, to commercial databases from Dialog, Lexis/Nexis and others, to *ad hoc* systems for scanning email or finding files on a PC.

A major challenge recognized by the NSF's Data Mining Research program and the broader interagency Knowledge Discovery and Dissemination (KDD) program is how to bring together sets of data and their associated IR systems in order for information seekers to better meet their needs. Such federated datasets are beneficial because they can allow for customized systems for each set of data, but these benefits create challenges:

- Different systems have different types of data (plain text, HTML, numerical data, XML, etc.)
- Different systems have different search capability (for example, some systems might allow Boolean statements AND or OR, others might facilitate searching in particular structured fields or sub-sections of documents)
- Different systems might have different levels of performance – both in terms of time to respond, and accuracy of responses

These challenges have created a rich but largely unaddressed set of research problems for data fusion (particularly, in the context of this paper, for fusion of text). For many data sets, however, these challenges are only the beginning. Information security is a major component for some data sets. As discussed below, security concerns might occur at the level of datum (i.e., a document), user or query source, or an entire set of data or its IR system. A set of additional security challenges for IR emerge when access to data must be restricted in some way:

- How can the data channel be secured?
- How may different search system components authenticate themselves?
- What sorts of access controls may be implemented at the datum, query and data set level?
- How will log files, audit trails, etc. support security?
- How will unauthorized access or other security breaches be identified?
- How might IR systems that share data be tested or modeled, to insure they meet security specifications?

As described below, a new model for information retrieval is under development for IR on computational grids: *GridIR*. With GridIR, many of the basic challenges are met in the standard for grid communication and virtual organizations. Furthermore, the specific challenges related to security are also met. However, significant practical and conceptual issues remain to be addressed through ongoing research and evaluation.

Information Retrieval and GridIR

In information retrieval (IR), computer systems are used to match statements of human information need (a.k.a., queries) to documents. Statements of information need may be anything from a few words (typical of Web search engines) to a structured logical

statement (i.e., SQL) to a profile of the information seeker and her assessments of past documents.

In practice, IR seldom actually presents answers to questions, or even specific information items (such as a table of figures, or a quotation, or a particular passage in a document). Rather, IR systems – notably Web search engines – present a ranked list of document citations in which, it is hoped, the desired information may be found. IR systems may be defined as those computer-based systems that take a number of documents (perhaps in different formats) as input and build data structures so that they may be quickly searched for matches to queries. These IR systems are able to then take queries (perhaps with special formatting or restrictions) as input, then produce a list of documents in ranked order.

We call the set of documents input to an IR system a *collection*. For a given information need, there might be more than one IR system that gives access to a collection of interest. While the days of massive search engines have led optimistic information seekers to think that all the information in the world is a click away via their favorite search engine, even these searchers know that different search engines host different collections. More experienced searchers know that there are thousands of different collections, different data types, different query languages, and so forth.

Rather than seeking to harvest these collections into single monolithic search engines, GridIR seeks to utilize distributed federated collections. There are important advantages over monolithic search engines in the GridIR scenario:

1. Queries will be run against collections with a likelihood of possessing relevant documents, thus eliminating *a priori* those collections unlikely to have documents of interest. For example, someone interested in information about household chemicals might want to bypass e-commerce datasets.
2. Each collection will be customized for that collection's qualities, according to the desires of the collection provider. Customization can include the full range of query processing, document processing and IR techniques such as document and term weights, IR retrieval models (Boolean, vector space, Latent Semantic Indexing), term stemming, stop word lists, etc.
3. Rather than seeking sub-second response times over billions of documents, as monolithic search engines do, GridIR systems may seek such performance over far smaller collections. This enables more complex query processing.
4. GridIR collections, because of their smaller size and locality to a particular collection source (i.e., an organization's Web server), can be updated more quickly, thus eliminating the delay among harvest runs exhibited by search engines. Collectively, we expect the capacity of GridIR to far exceed any search engine.
5. Grid computing offers a notification model in which events (such as the availability of new content) can trigger other events (such as evaluating a query against the new content). This model opens the door to standing queries,

information filtering, and push (rather than pull) approaches to information dissemination.

Grid computing[3] is an important new strategy for information processing. Computing grids offer a variety of services intended to supply the next generation computational platform for a variety of application areas. Fundamentally, grid computing is related to distributed computing and to parallel processing. As with distributed computing, computational grids enable physically separated computers to work together on tasks. As with parallel processing, these tasks may be shared in such a way that each computing element can count on particular resources being available to it. Grid computing extends these models by adding end-to-end security, an interoperability model, and a standards body to extend the scope.

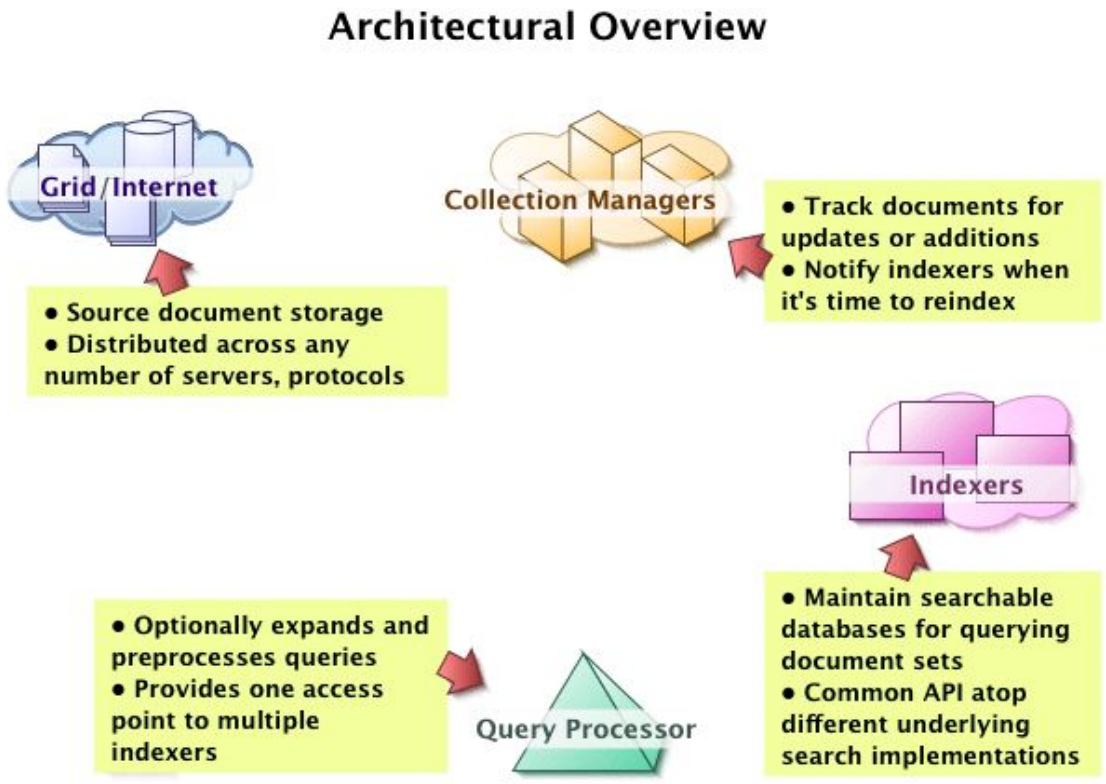
Grid computing is based on a set of emerging standards from the Global Grid Forum (GGF[4]). A fundamental notion of grid computing is the virtual organization. A virtual organization (VO) is a set of systems (or, more literally, specific Grid services) that are able to communicate with each other to accomplish tasks. VOs may be within a particular real organization (such as a federal agency, or a company, or an individual department or division), or it may span multiple organizations. It can exist on an Intranet, the public Internet, or any other network supporting TCP/IP.

Three main components make up GridIR. The components can exist on any computer on the computational grid with sufficient computational resources, software and access permissions to join in a particular virtual organization.

1. *Collection managers* (CMs). Systems that know how to access, stage/store, transform and deliver data items from collections. CMs may have privileged access to data, and may be able to perform complex or computationally intensive transformations. Alternatively, they may be simple Web harvesters or processes that access and deliver local files.
2. *Indexers*. These make up the core traditional IR system. They build searchable collections and process queries to deliver results. Results may be documents, URLs, document surrogates, etc.
3. *Query Processors* (QPs). These can interact with multiple indexers, over time, to gather query responses, merge them, and present them to their human users.

The first and second authors are co-chairs of the GridIR Working Group, which operates under the auspices of the GGF. This working group, ratified in fall 2002, is seeking to develop standards by which any organization with a desire to do so may make their collections searchable via GridIR.

[3] Foster, Ian; Kesselman, Carl; Tuecke, Steven. 2001. "The Anatomy of the Grid Enabling Scalable Virtual Organizations." International J. of Supercomputing Applications.
[4] See http://www.gridforum.org



GridIR in the National Security Context

Executive Order 12333[5], along with other laws and procedures, provides guidelines for information sharing among national agencies making up the intelligence community (IC). There are stringent restrictions on the extent to which IC members may share information. Moreover, due to security classification and other factors (such as need to know), information access within a particular agency may be limited.

There is no panacea to meet all information sharing needs, let alone all information retrieval needs. However, GridIR offers several important components to address information sharing among IC members. Several are derived from the use of Grid-standard technologies:

- End-to-end encryption. Grid computing, using the OGSA (Open Grid Services Architecture) and other standard services, encrypts all communication using ssh.
- System-level authentication. As provided by ssh, each member of the VO is cryptographically authenticated to other members for all sessions.
- Fundamental infrastructure for data exchange, service discovery and logging.

[5] 46 FR 59941, 3 CFR, 1981 Comp., p. 200. Available online at http://www.cia.gov/cia/information/eo12333.html and elsewhere

Other security considerations are enabled through GridIR itself, and are more directly geared towards the needs of the IC.

- Access control lists for collections
- Query processors that must authenticate in order to retrieve data or transmit queries.
- Collection managers that can implement authentication before delivering a datum, or transform the datum as needed. These collection managers may also seek asynchronous policy decisions by humans or external review processes.
- Indexers with customized capabilities for particular sets of documents, particular searchers, or particular information needs as desired.
- Sophisticated methodologies for data fusion from multiple sources.
- Capability for standing queries and information filtering
- Capability for cross-language information retrieval, multiple data types, multiple formats

All of the above features are included in the GridIR requirements document (to be presented at the GGF7 meeting in Tokyo on March 4, 2003 – see http://www.gridir.org). The specifics of architecture are being planned, and several reference implementations are underway. We anticipate functional production-grade GridIR systems and specifications will be available in 2004.

Specific IR Security Issues

In the GridIR environment, information sharing is enabled through the creation of virtual organizations (VOs). These VOs may be ongoing or *ad hoc*, public or private, interorganizational or extraorganizational. In the future envisioned by the GridIR working group, GridIR will enable organizations to make their collections searchable. This could happen in a variety of ways:

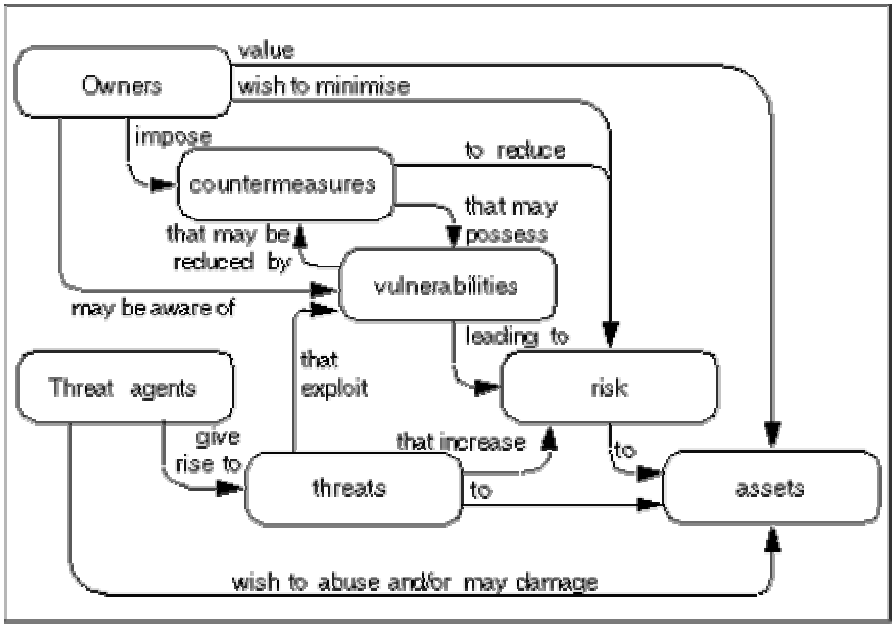
- An organization might use a GridIR collection manager to "publish" its content to indexers. These indexers, which might come from other organizations making up the VO, might build special-purpose sub-collections with particular capabilities.
- An organization might build its own index and make it available to others for searching.
- An organization might use a set of query processors (perhaps letting others create their own query processors) for standing queries and information filtering.
- An organization could use a collection manager as a filter, to authenticate particular information seekers or queries, or to screen or transform outgoing documents.

In all of these scenarios, the goal of GridIR is to facilitate implementation of security policy while providing for greatly enhanced IR capabilities as compared to many contemporary attempts to provide searching across federated data sets. As is typical in security scenarios, the convenience and power offered by GridIR comes at an increased need for diligence to mitigate security risks.

GridIR Security Processes

The Common Criteria[6] (also known as ISO 15408) provides a process to evaluate the security for a particular target system. Under the Common Criteria, there is interaction between a particular system implementation and security policy or related needs. Thus, it is not feasible to assert GridIR is, itself, "secure." Rather, it is the intent of the GridIR standard that organizations will be able to utilize GridIR services to implement systems that meet security needs and policies.

The Security Context from Common Criteria



Grid computing offers a strong security model at some levels, but not the full spectrum needed for GridIR. In turn, GridIR is intended to offer a complete suite of security options for information sharing, but cannot anticipate all threats or organizational contexts. In order to combat this uncertainty, the GridIR development process will integrate ongoing attention to security following the Common Criteria and other "best practices" or guidelines.

Some concerns to be addressed in the GridIR security context include:

- What is the relationship between a security token (i.e., an authenticated Grid service on a remote system) and a human user? For example, high stakes situations might require out-of-band confirmation that tokens come from an authenticated human, not an intruder.
- How can public key encryption and other techniques help to enforce "eyes only" access for data returned by a GridIR service?

[6] National Institute of Standards and Technology. (1999). "Common Criteria for Information Technology Security Evaluation Version 2.1." Gaithersburg, Maryland: NIST. Available online at http://csrc.nist.gov/cc/

- What are scenarios for abuse of GridIR systems from within an organization by authorized users?
- What are scenarios for abuse of GridIR systems from within an organization by unauthorized users?
- What are scenarios for abuse of GridIR systems from outside of an organization by authorized users?
- What are scenarios for abuse of GridIR systems from outside of an organization by unauthorized users?
- For all of the above scenarios, what are the risks, countermeasures and potential vulnerabilities?
- What are appropriate access control levels? Do the existing levels of Confidential, Secret and Top Secret[7] suffice, when combined with restrictions from Executive Order 12333 or other organizational constraints?
- What risks are created by possessing and managing access control lists?
- How can breaches of security be identified? Breaches might include any of the scenarios developed above at the data set, query, user or datum level.
- What methods for testing threat scenarios will be effective at evaluating the robustness of GridIR systems from attack?

For all of the above, the Common Criteria and other standard security practices dictate that GridIR solutions must be carefully evaluated before being implemented in sensitive environments. When compared to other methods of retrieving and exchanging information, we believe GridIR will stand up well. Because of the groundwork provided by the OGSA and the Grid's concept of the virtual organization, some of the fundamental problems of data channel security and system-level authentication are solved.

More importantly, GridIR promises solutions to information problems. The reference systems handle virtually any type of textual data in a variety of IR contexts. GridIR is extensible through the addition of customized collection managers, indexers and query processors, yet offers a standard for these systems to interoperate. Because of core reliance on OGSA and the emerging GridIR toolkit, threats due to poor implementation are lessened – instead, the GridIR toolkit will provide a common basis for more effective information systems and information sharing.

Conclusion

GridIR reference systems, based on production-grade software, are scheduled for release in 2003. The GridIR Requirements Document, to be released under the auspices of the Global Grid Forum, is available in draft form at http://www.gridir.org. Further standards development, slated for the next 24 months, includes Architecture and Specifications documents, reference implementations, an open source GridIR toolkit, and integration

[7] See Executive Order 12958, "Classified National Security Information." Available online at http://www.dss.mil/seclib/eo12958.htm and elsewhere.

with other services and standards for grid computing. The track record of the leadership of the GridIR Working Group is strong and indicates a likelihood of success.

GridIR offers important advances for information retrieval and informatics which have not been discussed here, due to this paper's focus on security considerations. These advances include the development of research testbeds for comparing IR system effectiveness, implementing and evaluating data fusion techniques, providing different user experiences and interfaces for utilizing GridIR, providing sophisticated functionality for collection managers, and more.

One of the primary goals of GridIR is ubiquity. If successful, many organizations that now run their own customized search systems will choose to make these systems accessible as GridIR collection managers, indexers or query processors. While we see no need to discard monolithic search engines entirely (and indeed have great respect for their capabilities), we anticipate an information future where standardized federated information systems work in a grid computing environment to provide results which are then merged and presented to their human user. In this information future, people will have more search capability, organizations will have greater choice in how to make their data accessible, and everything will be based on a secure foundation.

The implications for national security are, potentially, at the highest level. Information access and sharing are among the greatest needs, if not *the* greatest need, for information analysts in the IC. While GridIR cannot bypass the provisions of Executive Orders 12333 and 12958, nor other organizational restraints on sharing information, it can offer a platform on which data sharing systems may be built. As a standards-compliant, flexible, and extensible platform for powerful information retrieval systems, GridIR may be a superior choice to customized systems which, historically, can be expensive to develop and may suffer from security problems brought about by implementation error or interoperability problems.

In national security environments where analysts in the IC are seeking information, GridIR offers this possible scenario:

1. The analyst uses a GridIR query processor to identify potentially useful data sets held by GridIR indexers.
2. The analyst formulates a query. If necessary, the query processor transforms the query for particular indexers.
3. The query is submitted to the indexers. Depending on the needs of the analyst, a query might be persistent, awaiting new documents of potential relevance.
4. The indexer returns references to documents matching the query to the query processor (documents may be of any available type including plain text, marked up text, multimedia, or mixed). References may be information rich (URIs, abstracts or even full text) or information poor (hashes or document numbers).
5. The query processor ranks and displays all returned references based on available information (in many cases, fusion methods will allow accurate ranking of references based on relevance scores, regardless of what indexer provided the scores).

6. The analyst can select documents to view based on whatever information is available via the query processor.
7. Selected documents may be accessible via the public Web, an intranet or other internal source, or an external organization. In all cases, the user may be required to present access credentials to view the document.
8. If necessary, out-of-band communication may occur before a document is delivered (in GridIR, events may occur asynchronously – so, it's realistic to expect that a particular query might yield some documents immediately, while others take time to arrive). This out-of-band communication can include organizational communication, data transformation or redaction, or assembly of customized documents.

GridIR offers important advances for information retrieval and informatics research in the national security context. When combined with data fusion, information visualization, filtering, and other techniques, GridIR provides a secure and standards-based platform for future information retrieval systems.