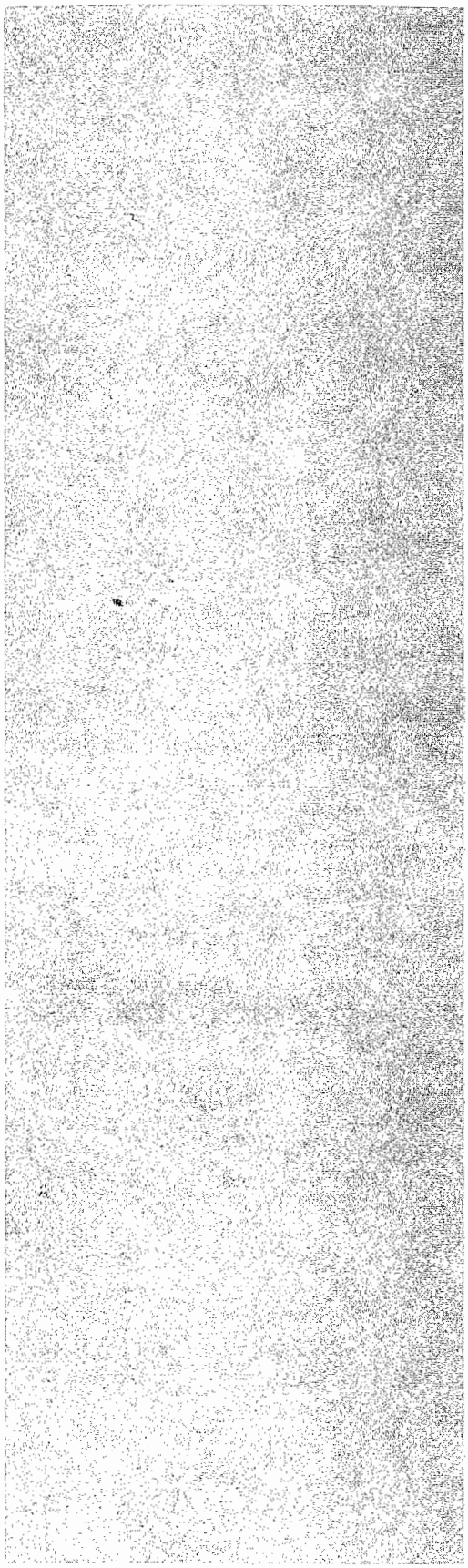
LA-12998-MS



Guidelines for the Implementation of an Open Source Information System

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36.

Edited by Paul W. Henriksen, Group CIC-1
Prepared by Sharon Hurdle, Group NIS-7

This work was supported by the U.S. Department of Energy, Office of Nonproliferation and National Security.

An Affirmative Action/Equal Opportunity Employer

LA-12998-MS

UC-700
Issued: August 1995

Guidelines for the Implementation of an Open Source Information System

Justin Doak
Jo Ann Howell

Los Alamos
NATIONAL LABORATORY
Los Alamos, New Mexico 87545

Guidelines for the Implementation of an Open Source Information System

by

Justin Doak and Jo Ann Howell

Abstract

This work was initially performed for the International Atomic Energy Agency (IAEA) to help with the Open Source Task of the 93+2 Initiative; however, the information should be of interest to anyone working with open sources. We cover all aspects of an open source information system (OSIS) including, for example, identifying relevant sources, understanding copyright issues, and making information available to analysts. We foresee this document as a reference point that implementors of a system could augment for their particular needs.

The primary organization of this document focuses on specific aspects, or components, of an OSIS; we describe each component and often make specific recommendations for its implementation. This document also contains a section discussing the process of collecting open source data and a section containing miscellaneous information. The appendix contains a listing of various providers, producers, and databases that we have come across in our research.

OSIS Components

Any OSIS will consist of many or all of the following components:

- Requirements Document
- Databases
- On-Line Providers
- Copyright Issues
- Internal Documents
- Organization of Data

- Storage
- Document Retrieval Tool
- Value-Added Tools (additional tools to speed-up or increase the accuracy of retrievals or both)
- Automation
- Multimedia
- Security
- Physical Configuration
- Design Document

Requirements Document

This document, based upon feedback from the eventual end-users, can be used to guide the development of the entire system. After identifying future system users, it may be best to initially request input from only a selected few before soliciting input from all users. Interviews should focus on functionality while avoiding excessive "computer lingo." The first cut of the system must be user-friendly, or the users will have a bad first impression and may never use the system.

Mary Meyer of Los Alamos is working on the best way to obtain system requirements from users/analysts. She has written a book, *Eliciting and Analyzing Expert Judgement: A Practical Guide*,[1] on how to elicit expert feedback on a prototype of a system.

Databases

Choosing the databases with the most relevant information is an ongoing process in today's rapidly expanding world of information. Look at "Locating Sources of Information" on page 12 for information on how to locate databases of interest. Once a relevant database has been located, one can probably obtain access directly from the producer of the database or, if that is not possible, the database will probably be available from an on-line provider. There may even be a chance that the database, or some form of it, is available on the Internet.

On-Line Providers

Many dial-up services provide, for a price, a significant amount of open-source information. Most providers have their databases in an identical format so that, even if the databases come from different producers, searches can be performed across multiple databases. Once users have located information they wish to obtain, they can download it, have it e-mailed to them, or print it out locally. Typically, one is charged for connect time, but some services now also charge for information based upon the number of users who will be accessing it. It should be noted that these providers may be considered expensive, with connect time ranging from $60/hr to $360/hr.

Two approaches can be taken when obtaining information from on-line providers. One can go into the provider's system and do all the searching manually. Each provider has its own language for searching, and learning the searching procedure before using a provider could be time-consuming. On the other hand, one can often create an electronic profile with providers that describes the type of data (documents) of interest. The provider finds all the information that matches the profile and delivers it. Creating a profile with a provider can be more expensive than going in and searching; however, there is no need to learn a provider's query language nor is there a need to do daily searches for each provider.

How does one become aware of new sources of relevant information as they become available? If good providers have been chosen, it is likely that they will find relevant new sources as they become available and make them accessible through their service.

In "Appendix A: A Listing of Various Open Source Providers, Producers, and Databases" on page 15, we list several providers. Below, we provide recommendations about which providers to use to obtain a good coverage of available sources.

Recommendations. Organizations with open-source applications may use Dialog, Lexis/Nexis, and Questel/Orbit[2] as their providers. (If access to the Nuclear Nonproliferation Network (NNN) database is required, Compuserve is the only provider that gives access.) Broad profiles should be created with each provider to ensure that no relevant information is missed. (A text search-and-retrieval engine can be used to filter the information after it has been downloaded.) It is probably better to get a few irrelevant documents and most of the relevant ones than it is to miss a large number of relevant documents by trying to filter out unwanted ones. These providers were chosen because they give a good coverage of databases. However, the providers to use may best be determined by noting the usage patterns of analysts. If a particular source or provider isn't being used regularly, it can be removed from the system. Likewise, trial runs can be attempted with providers to determine if they are useful. If documents downloaded from the provider are accessed frequently, it is a useful service.

Copyright Issues

Copyright issues with open source providers and producers must be worked out on a case-by-case basis. In other words, each provider and producer will have certain rules that determine how the information can be redistributed or archived for multi-user access. (Details of several providers and database producers are discussed below.) Dialog is the only current provider that allows redistribution or multi-user access to archived copies; there is an additional cost depending on the number of users with access. All the other providers currently have no policy for redistribution and archiving, allowing only one downloaded copy. (If multi-user access to documents downloaded from the provider is desired, the initial source of the data must be contacted.) For all databases obtained directly from the producer, the producers themselves need to grant the redistribution rights.

The bottom line is that current copyright license practice allows only one copy to be downloaded for personal use with no redistribution. If redistribution is desired, the *source of each database* will need to grant those rights. The only exception to this is Dialog, which grants redistribution rights to the data it compiles at a higher cost.

Current Copyright Status with Particular Providers and Databases

- Dialog[3,4]

  One must pay extra for downloaded information that is to be distributed to more than one user or archived for multi-user access. It costs more to redistribute than to archive the data. The following example was given in the *New York Times* article "Fee Plan to Share On-Line Data."[5] "With a $2 document, for example, it would cost $6 for up to 25 people to have access to the document through the corporate archive. The client would pay $8 for permission to send electronic copies directly to 25

people." For an end-user population of, say, 200 to 300 analysts, the document multiplier is 10 for archiving the data. So, in the above example, it would cost $20 to make the document accessible through an OSIS.

The success or failure of Dialog's Electronic Redistribution and Archiving policy may very well determine if other providers follow suit.

- Lexis/Nexis

  Mead Data Central, owners of the Lexis/Nexis service, currently have no redistribution or multi-user archiving policy. This means that only one copy can be downloaded for a single analyst's use. Redistribution rights will need to be granted by the actual source of the data.

- Questel/Orbit

  When contacted, Roger West of Questel/Orbit (1-800-456-7248) did not know exactly what their current policy was, but he knew that they didn't have a Dialog-like policy of allowing multi-user access at increased cost.

- Emerging Nuclear Suppliers Project (ENSP)

  Chris Fitz of the Monterey Institute reported that, if more than one user is to access the data, a site license needs to be purchased. A site license with six updates per year is $14,000 and with monthly updates is $16,000.

- Foreign Broadcast Information Service (FBIS)

  Two organizations distribute the FBIS data. The first distributes the data as For Official Use Only to government organizations and allows unlimited distribution within each organization. The other distributor sends the data to private entities and will have a Dialog-like policy of providing multi-user access at an extra cost.

- NNN / CompuServe

  When an organization decides to participate in the NNN News Service, it agrees to the following condition, "The participant agrees to respect and abide fully with the requirement that information contained in AP, Reuters, and Washington Post news reports may not be further republished or redistributed." A form must be signed before participation in the network is allowed.[6] This copyright issue will need to be addressed before any NNN information is redistributed to analysts.

  A general document outlines CompuServe's copyright and redistribution policies: "Compuserve Copyright Information."[7] Here is a paragraph from that document discussing use and redistribution of copyrighted material.

  > Any member may download copyrighted material for their own use. Any member may also non-commercially redistribute a copyrighted program with the expressed permission of the owner or authorized person. Permission must be specified in the document, on the Service, or must be obtained directly from the author.

Internal Documents

Perhaps some of the most useful information to put in the hands of analysts is information generated internally. (The IAEA, for example, has thousands of reports and many databases that may be useful to their analysts.) Much of this information may be in hard-copy and would require scanning and Optical Character Recognition (OCR) techniques to make the reports accessible on-line. It may be beneficial to save both the

image and the ASCII text that are generated in the scanning and OCR processes, respectively; one can index the text for later searching, even if it contains many errors from the OCR process, and present the more legible scanned image to an analyst. (The publication, "Information Science Research Institute: 1994 Annual Research Report,"[8] can be used to help choose effective scanning and OCR techniques.)

A more optimal situation would be for the documents to be kept on-line so that scanning is never necessary. To this end, it may benefit an organization to establish a policy that new reports are inserted into the system as they are created and still in electronic form. The technical procedures that are necessary to accomplish this may involve format conversions and other document manipulations.

An example of a system that provides access to internal documents is the Los Alamos Reports OnLine Archives (LAROLA) at the WWW site http://www.c3.lanl.gov:8076/. Here is a short description of LAROLA extracted from that location.

> LAROLA provides 24-hour access to archival copies of a subset of Los Alamos Reports. These articles are fully searchable by author, title, keyword, or natural language queries. Preview images and printable versions are deliverable directly to the user's computer.

Organization of Data

Data should be organized to be of the most benefit to the end-users. For instance, an OSIS for a nonproliferation application may have the data partitioned according to country. As new documents arrive daily into the system, automatic filters can be set up to direct the documents into the most relevant partition (i.e., database).

One may wonder about the importance of ensuring that partitions don't contain multiple copies, or redundant copies, of the same document; it can be frustrating to the user to retrieve multiple copies of the same document. A general rule-of-thumb is to simply inform users of any redundancy inherent in the system as it can be very expensive and time-consuming to insure uniqueness of documents.

Recommendations

- Organize document partitions to be the most beneficial to the eventual end-users.
- Infrequently re-index entire partitions, on the order of once a week, during off-hours.
- Keep partitions to a manageable size, under a few thousand documents.
- Set up the system so that new documents are automatically filtered into appropriate partitions.
- Inform users up-front of the possibility of redundancy in the system.
- As a general rule, break long documents into small pieces and index each piece separately. For instance, a document with multiple chapters should have the chapters indexed separately.

Storage

Design of an open source system should include storage requirements and solutions that meet those requirements. It is estimated that most systems will have storage requirements at least on the order of gigabytes. One possible solution is a network file server containing many hard drives with approximately 1 gigabyte each. The hard drives are

seamless in that data that is logically grouped together can extend across multiple hard drives; the user never needs to know that the data is physically distributed on multiple hard drives. Another solution would use optical storage technology through CD-ROMs and jukeboxes or platters. Although this type of storage is easily expandable to handle larger and larger data requirements, it is currently slower than an array of hard drives.

An image repository database at LANL's main library, using CD-ROM and jukebox optical technology, could be used as a model for optical storage.

Document Retrieval Tool

The primary focus of this section will be on document retrieval tools that can be used to analyze open-source data. Although many other tools exist, the three that are in greatest use at Los Alamos are TOPIC, Wide Area Information Service (WAIS), and Personal Librarian. All of these tools (indeed, all current document retrieval tools) build an inverted word index as the basis for future searches. We describe some of the strengths and weaknesses of each of the tools in the sections below.

TOPIC. This is a commercial tool that can be reasonably expensive, $19,000 for a server and $600 for each client. It performs well even with a large number of documents and it can perform Boolean keyword searches. Perhaps the most appealing feature of TOPIC is the ability to build TOPIC trees (i.e., search queries) that enable users to fine-tune searches for their needs. If, at some point, system requirements change dramatically and specific modifications to TOPIC are necessary, a contract with Verity, the creators of TOPIC, can be purchased to obtain the modifications.

Personal Librarian. Another commercial document retrieval tool, Personal Librarian, is claimed to be very user friendly and allows "fuzzy" searches, that is, the ability to perform partial matches on keywords. Personal Librarian may not be as robust as TOPIC in terms of its ability to handle large numbers of documents. This tool is much less expensive than TOPIC and will run on PC, MAC, and UNIX platforms.

WAIS. This widely used freeware can be obtained from the Internet via anonymous ftp. WAIS is easy to use and install and retrieves relevant documents well. One of the unique features of WAIS is its ability to perform a certain amount of natural language processing. One can give WAIS a document, and WAIS will retrieve all documents that it determines to be similar. The major drawback was the lack of sufficient documentation and technical support because it wasn't a commercial product. It has recently been commercialized, so documentation and technical support should be improved now. We have yet to inspect the commercial version.

Value-Added Tools

In this study, we came across several tools that are either complete systems or components in an open source system. We make no judgments on the utility of any of the products discussed below.

Intelligent Document Detection System. [9] This is a prototype currently being tested at The MITRE Corporation for searching and browsing on-line news sources and MITRE documents.

OSIS. The Community Open Source Program Office has developed an architecture for the intelligence community's new Open Source Information System (OSIS*). The following text is from an informational brochure on OSIS: "OSIS provides comprehensive access to open source information for users throughout the Intelligence Community and the rest of the government. It includes electronic mail (e-mail) and information processing tools, gateways to other government databases, commercial databases, and the INTERNET." OSIS is a distributed system made up of intelligence community agencies who have a node on the system. Each node gives access to a different set of open-source information. Any person who has access to OSIS has access to all nodes on the system.

OSIS may have some parallels to what the developers of an open source system are trying to provide for their own community. In particular, the Open Source Service Agent (OSSA)[10-12] maintained by the National Air Information Center, which is simply a node on OSIS, may provide many ideas on effectively implementing an open source system. Note that Pathfinder, discussed briefly below, is a part of the OSSA system.

PathFinder. This is an analyst's tool to link open-source information pieces. Pathfinder runs on at least the Sun platform and has a front-end to relational databases. Look at Ref. 13 for pointers to PathFinder for more information.

GTE Multiple Open-Source Engineering Solutions.[14,15] They have created a simple user interface to access and capture information. Mosaic and Lotus Notes are integrated so that one can browse the Internet and then save, manipulate, and share relevant data.

Paracel's FDF 3.[16-18] This is a search engine in the hardware of a massively parallel desktop computer. In other words, the product is *hardware* designed specifically for doing free-text searches. It should be noted that the performance of this tool may be significantly faster than software retrieval tools even though *no inverted word index* is created.

Stange Associates' Integrated Text and Geographical System. Stange Associates is working on an integrated text and geographical system that will include maps, photos, and drawings. They want to develop an inexpensive, pc-based product that fulfills IAEA requirements as one of its main objectives. Part of the system will be a text-retrieval engine being developed by another company; Stange Associates hopes the client software for this tool will be in the $50 range but expects it to be more. This tool will be used by Safeguards Operations A, B, & C of the IAEA for safeguards and special inspections. This is a Small Business Innovative Research (SBIR) project and must adhere to the following phases: 1) design (completed), 2) prototype (25% completed), and 3) production (not started). Stange Associates estimate that it will be 1-1/2 to 2 years before the end of phase 3. This product will be obtainable directly from Stange Associates with the following estimated costs: $8-10K for server hardware (exclusive of jukeboxes or other massive storage devices) and client software at $50+ per client.

A contact for Stange Associates is Al Glock, who can be reached at 310-545-9828 or via e-mail at 71541.415@compuserve.com.

*Note the overloaded use of OSIS. We use it in this document to refer to generic open source information systems, but it is used here to refer to a specific system.

Automation

Most stages of the open-source process, including gathering, indexing, partitioning, and filtering data, can be automated. For instance, if subject profiles are established with providers, the providers will perform the necessary searches, perhaps even daily, and deliver all documents that match the profile. Documents can be filtered automatically by using pre-established search queries to select the most appropriate partition for a document. Taking full advantage of the opportunities for automation in an OSIS can keep the daily maintenance of the system to a minimum.

Multimedia

Future OSIS systems may very well integrate maps, photos, in-line images, radio and TV broadcasts, and other forms of multimedia open-source information. These other media forms may increase the utility of OSISs once they are integrated. Multimedia, however, is not the focus of this report; we focus on providing access to free-text documents.

Security

Many organizations with OSISs share a mutual concern that can be paraphrased as follows, "If I allow my computer to be connected to the network to access the OSIS, aren't I opening myself up to attack from the outside world? At the very least, even if there is no outside connection anywhere on the network, my data will be accessible to anyone on our internal network." Given the frequent presence of security flaws in computer systems, one must assume that the above statements are true. However, we feel that the benefits of being connected to a network and accessing outside data sources far outweigh the risks of having a connection that might provide access to malevolent parties.

A solution. A possible solution to this problem, used by several open-source users at Los Alamos, is for one's computer to have two external hard drives; one contains network communication software (e.g., TCP/IP) and is used to retrieve open-source information; the other contains any classified or sensitive data. In this fashion, the system can be disconnected from the network when the classified hard disk is installed by removing the disk containing the network software. To access open-source information, the hard disk with network software is reinstalled, allowing the machine to communicate over the network.

Note that the disk containing the sensitive data must not only be de-installed, but be physically disconnected from the system entirely and stored in a secure safe or area. This will eliminate the possibility of someone logging into the system over the network and installing the disk, or having someone walk into the room containing the system, plug the classified disk in, and then install the disk. Both drives must be external for them to be easily disconnected from the system; a PC frame may facilitate removing and plugging in hard disks. CG Enterprises is one company that makes PC frames and calls them CRUs.

It may be a good idea to establish a standard operating procedure (SOP) to ensure that proper procedures are followed when transitioning from one processing mode to the other. The SOP should describe the steps for disconnecting the system from the network before processing sensitive data and the sanitization or clearing procedures after completing sensitive processing to reconnect to the network.

Physical Configuration

An effective means of configuring the physical aspects of the system needs to be devised. Is the network configuration appropriate for the type of access one is trying to provide? Is the hardware being used for the server and clients sufficient?

Design Document

The OSIS should include a written document describing all aspects of the system. A detailed description of each of the components outlined in this report would be sufficient. *This may in fact be the most important part of the system and deserves considerable attention.*

Collecting Open-Source Data

Depending on the original format of the data (e.g., hard copy or electronic), the collection process can include one or both of the following tasks:

1. Obtaining the information of interest from the open-source provider. (Note that the provider may be an on-line provider, such as Dialog, or the producer of the database.)

   The information may be available via one or more of the following forms: hard copy, electronic media (diskettes, magnetic tapes, compact discs (CDs), for example), the Internet (e.g., ftp archives), and dial-up services. The first two, hard copy and electronic media, require physical delivery of the open-source information, ftp archives require Internet access, and dial-up services require a modem and a phone line. (Of course, no matter what form the data comes in, the supplier may require a fee.)
2. Scanning and then using OCR procedures on any hard copy.

   For hard copy, one must at least scan the document to get an on-line version. Several options exist at this point; one can manually index the scanned document and never convert it to ASCII format; one can convert to ASCII for indexing, but retrieve the scanned document for the analyst; or one can convert to ASCII, run error correction algorithms to correct translation errors in the OCR process, and present this document to the user.

Below we discuss several aspects of collecting open-source information.

Hard Copy

Hard copy is still a common method of receiving open-source information. A typical scenario is that an employee reads an article in a publication that may be of interest to the organization as a whole, and the employee wants to make it accessible to everyone. (Provided, of course, that this type of redistribution right has been obtained.) If the organization uses a text retrieval tool to access a database (or databases) of documents, the document needs to be scanned, OCRed, and indexed before it can be accessed via the analysis/retrieval tool. (Scanners and OCR devices are discussed in "Scanning and OCR" on page 10.)

Electronic Media

Electronic media, including diskettes, magnetic tapes, and CDs, are also ways to distribute open-source data. An example of a database producer that distributes its information in this fashion is the Monterey Institute with their ENSP, International Missile Proliferation Project (IMP), and other proliferation databases. These databases are distributed via diskettes that include software to search and print the data.

Vol. II of the *Gale Directory of Databases* lists databases that are distributed by CD-ROM, diskette, and magnetic tape.

Internet

A large source of open-source data, much of it free, is the Internet. News articles, ftp archives, USENET news groups, conference proceedings, and other forms of information exist in vast quantities. With this vast quantity of data, finding relevant information can be very time-consuming and open-ended. Numerous tools exist to facilitate the task of finding relevant information; Mosaic and the World Wide Web (WWW) are the primary tools and are discussed in the following paragraphs. In addition, there are numerous search tools on the Internet (e.g., LYCOS from Carnegie Mellon at "http://lycos.cs.cmu.edu") that can facilitate the process of finding relevant WWW sites. Using one of these search tools does not guarantee that every relevant site will be found because each search tool only indexes a subset of the information available on the Internet. Even with these sophisticated tools, finding relevant sources can be difficult.

The WWW is a large-scale, networked, hypertext information system started by CERN, the European Laboratory for Particle Physics in Geneva, Switzerland. Hypertext is text that contains links to other texts or to graphics, videos, and sound. Links are words or phrases designated with a color or by underlining or both depending on the browser (e.g., Netscape). Links are selected by clicking on the highlighted word or pressing the return key, again depending on the browser. The word that indicates the link then either changes color or the underlining becomes broken. The same link may be included in multiple documents.

Mosaic is an interface used by the WWW. It is easily installed and can be customized for each individual site or user. Mosaic is integrated with other information interfaces (gopher, ftp, archie, and WAIS), although they are not required. The Mosaic software is free from the University of Illinois (there are now commercial versions, such as Netscape) and is available across several platforms: Sun (X Windows), Macintosh, and IBM (Windows). It can provide access to local files on a stand-alone machine, information on a local area network, or information available internationally. It can access text (free text and formatted text) and graphics (in one of several standard formats used on Suns, Macs, and PCs). Once the text and graphics are in one of these standard formats, they are accessible to all platforms for both viewing and transfer. Information can be transferred through ftp and the TCP/IP protocol between a Sun and a PC. Archived information as well as system documentation can be made available in this manner.

Much of the information contained in this report was obtained from the Internet.

On-Line Providers

See "On-Line Providers" on page 2.

Scanning and OCR

OCR Software. The following is a list of commercial OCR packages:

- OmniPage - CAERE Corporation (handles graphics as well as text)
  Los Gatos, CA
- "Wordscan" - Calera Recognition Systems
  Sunnyvale, CA

- EDT ImageReader - Electronic Document Technology
  Singapore
- ExperVision RTK - ExperVision, Inc
  San Jose, CA
- Recognita Plus DTK - Recognita Corp. of America
  Sunnyvale, CA
- XIS OCR Engine - Xerox Imaging Systems
  Peabody, MA

A comparison is given in the Annual Report of the UNLV Computer Science Department (http://www.isri.unlv.edu under Publications). In summary, ExperVision RTK performed the best overall in the 1994 tests, while Caere OCR and Calera WordScan are both rated very highly and are accurate systems. Word accuracy on a magazine sample for these three systems is greater than 95%.

Scanners. There is a wide range in speed and cost. The following are illustrative.

- Fujitsu (27 pages/minute) - $5,000
- HP IIcx (100 sheets/hour, hand-held) - $200
- Bell&Howell CopiScan II (40 pages/minute) - $17,000
- Sunrise (2000 images/hr, fiche and film) - $108,000

Automation of Data Collection

Each provider will have a means of distributing information to users. For producers of databases who are also providers, automation may take the form of diskettes of updates to the databases mailed to users quarterly. On-line providers, on the other hand, often provide automation in the form of electronic profiles; any documents matching your profile can be automatically distributed to you, perhaps daily. Automation on the Internet could come in the form of mailing lists for newsletters, for example. We foresee more typical data collection on the Internet to be guided by periodic browsing of sources initially determined to be useful.

Miscellaneous

Non-Commercial Software

Many tools may be used in an OSIS that are not commercial software. Some of these tools may be excellent aids to analysts, but one should consider that support for non-commercial software can be difficult to obtain. One should also be aware that, with non-commercial software, part of the cost pays for development.

Open Source Solutions

"Open Source Solutions, Inc. (OSS) was founded in 1992 by Robert David Steele to facilitate and accelerate discussion and understanding—both within governments and in the private sectors of all countries—of those issues that must be resolved if each country and each enterprise is to be competitive in the Age of Information. The company researches the theory and practice of utilizing open-source information to satisfy national security requirements more quickly and cheaply than is possible with classified intelligence capabilities. The company also explores how governments and private sector enterprises can coordinate their open-source collection, processing, and translation efforts to avoid wasting manpower and dollars on duplicative efforts."

This organization has held three conferences on using open sources to solve intelligence problems. It may be worthwhile for the developers of an OSIS to attend these conferences. The 4th OSS Conference will be Nov. 6 - 9, 1995, in Washington, DC. Registration information is available at the WWW site, http://oss.net. "OSS Notices," a newsletter, is put out monthly by OSS. Back issues can be reached at the above WWW site under the heading "Open Source Solutions, Inc. / Complete Back Issues of OSS NOTICES."

Locating Sources of Information

There is no shortage of open-source information; with proper research, open sources will be found that provide useful information. The most comprehensive* directory of open-source information is the *Gale Directory of Databases* containing lists of databases, producers, and providers. It is published twice a year in two separate volumes. Volume I covers on-line databases accessible through dial-up providers, while Volume II describes databases available via electronic media (e.g., diskette). The Gale Directory also has geographic, subject, and master indices. The *Directory of United Nations Databases and Information Services*, another good reference for open source databases, lists the entire collection of United Nations databases.

A suggested procedure for discovering open-source information sources that are relevant to a task is the following:

1. Browse both volumes of the *Gale Directory of Databases* to find potentially useful sources. (This process will be facilitated by subject indices at the back of the volumes.) Before access to the databases is purchased, the database should be inspected to ensure its relevancy.

Do the same as above for the *Directory of United Nations Databases and Information Services*.

Use some of the WWW search tools (e.g., LYCOS) to locate relevant information on the Internet; this will be an ongoing process. In "Appendix A: A Listing of Various Open Source Providers, Producers, and Databases" on page 15, we present some of the information sources we found on the Internet to provide starting points.

Subscribe to one or more of the charge-for-service providers and receive their documents for some evaluation period. It is likely that several of the providers will contain databases that are relevant and, as new relevant databases become available, will add them to their service.

Research Initiatives and Contacts

Xerox Parc. They are doing a considerable amount of research in textual analysis. Examples of the work they do are given below.

- A joint initiative on digital libraries with the National Science Foundation, the Advanced Research Projects Agency, and the National Air and Space Association
- The CLASS Project (with Cornell)

  This project involves scanning and printing brittle, old documents.
- Information Filtering

  Identification of potentially interesting documents is the goal of this research area.

*For those interested in nonproliferation databases, obvious omissions from the directory are the databases put out by the Monterey Institute.

- ACT/VKAT
  These are tools for analysis and data interpretation.
- Multimedia Indexing and Retrieval
  This project focuses on OCRed works and has produced several papers.

Descriptions of current work at Xerox Parc in digital libraries are available through Mosaic by going to the URL http://pubweb.parc.xerox.com/.

The University of Nevada Las Vegas. The following text was obtained from the WWW site "http://www.isri.unlv.edu/." "The Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas (UNLV), was established in 1990 by a grant from the United States Department of Energy (DOE). The overall mission of ISRI is to foster the improvement of automated technologies for document understanding. To help satisfy the needs of the Department of Energy, ISRI has focused its attention on technologies for recognition and retrieval of information from machine-printed documents."

ISRI produces an annual report that analyzes the various scanners and OCR algorithms currently on the market.

Journals

The Association for Computing Machinery, Special Interest Group on Information Retrieval (SIGIR) periodically publishes proceedings and a bulletin.

Conferences

The Conference on Information and Knowledge Management (CIKM) may be an excellent source of information for open-source analysts. The last conference, held in November 1993 in Washington, DC, had sessions on Information Retrieval Systems, Document Processing, Deductive Databases, User Interfaces / Image Databases, Query Processing, Information Engineering, and others. A proceedings from this conference is available. A tutorial was given at CIKM '93 entitled "Introduction to Information Retrieval" by Edward A. Fox of Virginia Tech. Mr. Fox may be reached at the Department of Computer Science and Computing Center, Virginia Tech (VPI & SU), Blacksburg, VA, by e-mail at fox@vt.edu, or phone at (703)231-5113.

Books

Bookstores that carry a good selection of technical books should have several on the Internet, for example, *The Whole Internet: User's Guide and Catalog*, by Ed Krol, published by O'Reilly and Assoc.

References

1. Mary A. Meyer and Jane M. Booker, *Eliciting and Analyzing Expert Judgment, A Practical Guide, Vol. 5 in Knowledge-Based Systems* (Harcourt Brace Jovanovich, London, 1991).
2. Questel/Orbit information can be obtained from the WWW site "http://www.bedrock.com/mall/OQ/oqovw.html.
3. Clinton Wilder, "The Cost of On-Line: Dialog revises fees, others may follow," *Information Week*, April 25, 1994, p. 15.

4. "The DIALOG ERA Service: Electronic Redistribution and Archiving Made Easy."

   Contact Dialog to obtain a copy of this document. See "Dialog" on page 15.

5. Teresa Riordan, "Fee Plan to Share On-Line Data," *The New York Times*, Wednesday, April 6, 1994, p. C1.

6. Nuclear Non-Proliferation Network Agreement Form

   Obtain by contacting the Nuclear Non-Proliferation Network, Carnegie Endowment, 2400 N St., NW, Washington, DC 20037.

7. "CompuServe Copyright Information."

   Obtain this document from CompuServe. Contact and other information is located at http://compuserve.com/index.html.

8. "Information Science Research Institute, 1994 Annual Research Report," Information Science Research Institute, University of Nevada, Las Vegas. E-mail address isri-info@isri.unlv.edu.

9. Stephen A. Glanowski, T. Richard MacMillan, Inderjeet Mani, "Intelligent Document Detection," The Mitre Corporation, Nov. 1994, Center for Integrated Intelligence Systems, 7525 Colshire Drive, McLean, VA 22102-3481, (703)883-6000.

10. "Open Source Service Agent (OSSA), Gateway to the World of Open Sources." This brochure is a description of the tool responsible for much of the functionality of OSIS.

    Obtain copies of References 10, 11, and 12 by sending e-mail to Mr. Bruce Fiene at brf3@naic.wpafb.af.mil.

11. Thomas R. Pedtke, "Intelligence Community Open Source Architecture," Oct. 1, 1994.

12. Bruce Ronald Fiene, "OSSA Access," (see Ref. 10 for address).

13. PathFinder information may be obtained by contacting Presearch Incorporated, 8500 Executive Park Avenue, Fairfax, Va. 22031, Attention: Mr. Kevin Greer, 1-800-922-9259.

14. GTE, "Multiple Open Source Engineering Solutions."

    Obtain copies of References 14 and 15 by contacting GTE Government Systems Corporation, Mr. Steven Z. Stark, 1700 Research Boulevard, Rockville, MD 20850, (301) 294-8688, e-mail: stark@batman.rock.gtegsc.com.

15. "GTE Open Source Applications Demonstration," Open Source Symposium, November 8-10, 1994.

16. "FDF 3, Solutions for Real-Time Text Analysis," brochure from Paracel.

    (See "Paracel's FDF 3" on page 7 for a description of this tool.)

    Obtain References 16, 17, and 18 by contacting Paracel, 80 South Lake Avenue, Suite 650, Pasadena, CA 91101-2616, (818)666-6688, e-mail: information@paracel.com.

17. "FDF 3, Technical Description," brochure from Paracel.

18. "FDF 3, Application Note: Information Dissemination and Categorization," brochure from Paracel.

Appendix A: A Listing of Various Open Source Providers, Producers, and Databases

A provider gives access, usually for a fee, to one or more databases. There is often the opportunity to search across multiple databases and download retrieved information. A well-known provider is Dialog. A producer, on the other hand, is the actual creator of one or more database(s) that are then made accessible to others via one or more providers. For instance, the Monterey Institute is the producer (and provider) of the Emerging Nuclear Suppliers Project database. A producer can give access to its own database(s) and therefore also be a provider. Databases are the products created by producers. The following providers, producers, and databases were uncovered during the research for this report.

Providers

Lexis / Nexis. This provider contains over 2000 full-text sources including BBC (British Broadcasting Corporation) reports, *The New York Times*, and international information. Information can be downloaded but, unfortunately, the format of documents is inconsistent.

European services

- **Data Star,** the European equivalent of Dialog, is a corporate provider.
- **Questel/Orbit** is a member of the France Telecom Group and is an international on-line information company specializing in patent, trademark, scientific, chemical, business, and news information.

Dialog. Provides access to Reuters information and subsumes Data Star. Contact Dialog for a publication describing the various databases they supply. Dialog, 3460 Hillview Avenue, Box 10010, Palo Alto, CA 94303-0993, USA.

Data Times. The primary use for this provider would be for up-to-date Reuters information, although Data Times does provide access to more sources.

STN International (Scientific and Technical Network). This provider is accessible through CompuServe and contains environmental information.

Australian National University. ANU provides access to a wide variety of on-line information through the Coombsquest Social Sciences and Humanities Information Facility. There are over 500 WAIS-indexed, searchable databases.

COOMBSQUEST provides state-of-the-art networked information services. It is developed and maintained by the Coombs Computing Unit to support the research and teaching activities of

- the ANU's Research Schools of Social Sciences and Pacific and Asian Studies,
- other schools and centers of the Australian National University, and
- other Australian and overseas research institutions.

At present, COOMBSQUEST's services include

- archiving social sciences research papers and documents;
- building and publishing specialist social sciences research databases;
- providing information services relevant to the social sciences and humanities research;
- networking world information systems and resources;
- advising on academic computing, information exchange, and e-mail systems; and
- advising on multilingual (esp. Asian languages) text processing.

First! by Individual, Inc. First! is an information service of Individual, Inc. (first@individual.com) whose subscribers receive (by e-mail and perhaps other methods) open press news stories from a customized profile of interest. For instance, the Nonproliferation and International Security (NIS) Division and the library at Los Alamos have profiles. (The NIS profile is more specific to nonproliferation topics than the library profile.) For NIS, there is one profile with approximately 20 people on it for $10,000 a year. One of the more important data sources to which First! provides access is Interfax from Russia.

Here is an example of an e-mail notice received by a user:

```
>                                   Monday, June 20, 1994
>*********************************************************************
>First! First! First! First! First! First! First! First! First! (tm)
>*********************************************************************
>-- Your Smart News Agent --              (C) 1994 by INDIVIDUAL, Inc.
>---------------------------------------------------------------------
>PROFILE TITLE:  Uncustomized Profile
>CUSTOMIZED BY:  Los Alamos National Labs
>=====================================================================
>12,984 Stories Received Today; 16 Relevant To Your Profile:
>---------------------------------------------------------------------
>1. ADD/ KOREA/ WHITE HOUSE DENIES OFFER TO DELAY SANCTIONS VS
NORTH
>2. RUSSIANS/ NO N. KOREAN A-BOMB
>3. U.S. SEEKS JAPAN'S TECHNOLOGY FOR TMD, DAILY SAYS
>4. HATA SAYS JAPAN CAN PRODUCE A NUCLEAR BOMB+
>5. PENTAGON ESTIMATES 50% OF NORTH KOREAN SCUDS COULD GET
THROUGH, AVIATION WEEK REPORTS
>6. ABM MOVEMENT
>7. JAPAN AND EUROPE SIGN NUCLEAR RESEARCH ACCORD
>8. SIEMENS EXASPERATED AT GERMAN NUCLEAR FUEL SPENDING
ABROAD
>9. NORTHEAST SEES RESTARTS AT CONN. N-PLANT
>10. AGENCY ORDERS CHECK ON FUKUSHIMA REACTOR ACCIDENT
>11. DISPOSING OF GERMAN WASTE
>12. HUGHES AWARDED $114 MILLION IN PATENT CASE
>13. AIR FORCE FUNDS MILSTAR, CONSOLIDATES FOLLOW-ON EFFORTS IN
OUT-YEARS
```

>14. SENATE PROPOSES $120 MILLION FOR USAF SPACE SURVEILLANCE STUDY
>15. TAKING THE SHAKES OUT OF STATION
>16. AND IN OTHER NEWS ...

Following this introduction, would be the actual news stories.

ClariNet News Service. ClariNet draws news from a variety of sources. This news is processed and converted into USENET format at ClariNet facilities. It is then sent out via UUCP (the telephone/modem based inter-unix communications facility) and TCP/IP (the computer communications protocol used by many machines, including those on leased line networks like the Internet) to ClariNet customers around the world.

They receive United Press International wire service news directly via satellite, in the same way that newspapers receive it. The wire news comes (more or less) in what is known as the American Newspaper Publishers Association format.

- ClariNet articles have a meaningful headline prepared by a professional journalist. You can scan the headlines quickly to see what you wish to read.
- ClariNet articles are keyworded using the topics the article covers.
- ClariNet articles are not discussions, they are news. There are no follow-ups, although reference chains exist.
- ClariNet articles come with a wide variety of extra headers providing useful classifying information about the article.
- ClariNet articles come fast, and network links are designed to propagate them quickly. They also become stale more quickly, turning into "yesterday's news."
- ClariNet articles on big stories are updated frequently. Each update cancels the previous article and adds a new one with the latest details. You will thus find lots of gaps in ClariNet news groups where canceled articles used to be.
- As a consequence of the above, ClariNet feeds generate hundreds of cancel messages every day.
- ClariNet articles are all copyrighted and may not be distributed without permission. See the licence terms.
- Most ClariNet articles are cross-posted to two to four groups, if their subject matter falls in multiple categories.
- You cannot reply to, or follow up on, ClariNet articles. They are publications, not discussions. Some groups exist for the discussion of ClariNet and articles within it. Most ClariNet groups are marked as "moderated," but you may not submit to them, even by mail.
- Some ClariNet articles make heavy use of underlining as understood by many news-reading programs. (Underlining is done by prefacing a character with an underbar and a backspace.)

The ClariNet articles can be read with Mosaic from the Internet, with subscription.

Producers

Monterey Institute. This organization has several proliferation databases to which it provides access by distribution of diskettes. Probably the most used of these databases

is the Emerging Nuclear Suppliers Project (ENSP) database. However, the Monterey Institute provides several other databases of proliferation interest: the International Missile Proliferation (IMP) Database and five separate databases each focusing on an important aspect of the nuclear programs of the Commonwealth of Independent States (Former Soviet Union).

Databases

FBIS. This database contains foreign sources of information and is owned by the CIA. FBIS creates daily reports that are issued Monday through Friday and cover information published within the previous 48 to 72 hours. (Recall that there are two versions of the FBIS database. See "Foreign Broadcast Information Service (FBIS)" on page 4. The version we discuss here can be distributed to private entities.) Currently, FBIS reports are distributed via CD-ROM. However, FBIS eventually plans to provide Dialog-like access where users can create an electronic profile and do on-line searches. The relevant information will either be faxed or e-mailed on a daily basis. FBIS is in the process of obtaining copyright clearances with their sources that would allow this kind of access; this process is expected to be complete sometime in the summer of '96.

There is an FBIS WWW home page at the following URL (Uniform Resource Locator), http://fwux.fedworld.gov/ntis/fbis.htm. It contains only general information and does not give access to the reports. The phone number to order reports, and the electronic service when available, is (703) 487-4630.

Joint Publications Research Service. This is another service (database) provided through the CIA and contains TV, radio, news, journals, and other sources of information.

National Technical Information Service. This service is provided by the US Dept. of Commerce from Springfield, VA, USA. This source includes results of US government-sponsored research, development, and engineering, plus analyses prepared by federal agencies, their contractors, or grantees. It is available through Dialog and Compuserv.

Office of Scientific and Technical Information. This database is provided by Oak Ridge National Laboratory for the various energy laboratories. Oak Ridge provides access to this database through a modem.

National Energy Database. Petroleum activities are described in this database.

ALERT. ALERT is a weekly electronic service giving information on new technology developments. This database is produced by Technical Insight, Inc. (Contact Kenneth A. Kovaly, President, at 201-568-4744, or P.O. Box 1304, Fort Lee, NJ 07024-9967.)

Library of Congress (LOC). The LOC has an ftp directory accessible through the mosaic URL "ftp://ftp.loc.gov/pub/iug."

The Library of Congress Information System can be reached by telnetting to locis.loc.gov. The LOC catalog can be searched on-line. We did a search on nonproliferation and found 13 items. The catalog is not user friendly at all. The catalog has records of what the library has; the actual documents do not appear to be on-line. (Some may be in the ftp directory.)

The LOC gopher server (LC MARVEL) can be reached by telnetting to marvel.loc.gov (login as marvel) or gophering to marvel.loc.gov port 70. This did not contain much useful information, just a lot of little tidbits about the Library.

The Federal Library and Information Center Committee (FLICC) has a program called FEDLINK (the Federal Library and Information Network) that provides agencies with cost-effective access to a number of information or operation support services.

Carnegie Endowment for International Peace's NNN Database. The Carnegie Endowment for International Peace conducts research in international relations, arms control, foreign policy, and peace negotiations. They publish a journal, a monograph, progress reports, conference proceedings, and maintain a private forum available only through Compuserv, the NNN. It consists of a bulletin board, on-line librarian, daily news reports (Reuters), articles, testimony, press releases from the IAEA and ACDA, for example, and e-mail service. It is considered an extended service from Compuserv. One must have an account, in addition to the basic CompuServe account, specifically for the NNN database. You pay by the connect hour plus a monthly fee, and have to sign up with the Carnegie Endowment. They will send you forms, a unique id # to be used at sign-on, and documentation including a "Guide to the Nuclear Nonproliferation Network." Brian Weinberger, NNN Coordinator, is a contact for more information on the database (202-862-7900)

Reuters. Reuters is an on-line database that contains textual, numerical, and statistical information from more than 2000 international publications. This database provides a valuable information resource to assist with teaching and research. Access to Reuters is available by subscription through the Internet.

Latvian News Service. This news service is free and available through the Internet. It contains the following three services: Baltic Business Weekly, Baltic News Service, and LETA (Latvian Telegraph Agency) News Agency. On Mosaic or gopher its address is gopher://gopher.lvnet.lv/11/NEWS%20of%20Latvia.

The Monterey Institute's ENSP Database. This database contains abstracts of non-proliferation-relevant articles. Updates to this database can be supplied as often as monthly via diskette. Bill Potter at the Institute is a contact who can supply more information (408-647-4154). Much of the information contained in the ENSP database may be available through other sources or providers who supply their information daily.

Los Alamos
NATIONAL LABORATORY
Los Alamos, New Mexico 87545