SIX YEARS OF OPEN SOURCE INFORMATION (OSI)

Lessons learned

by Major(res.) Mats Björc[1] [2]

When I first started my maiden voyage on the digital ocean of information, I travelled on a slow boat in the European information archipelago. In those days, the archipelago consisted of distinct and isolated islands. The islands in the archipelago were connected to the mainland by regular ferries with names like Dialog, Data-Star, and FT-Profile.

I began to explore these islands in order to understand the many and rapid events happening around the globe; the recent fall of the Berlin Wall, Desert Storm, and the fragmentation of the Soviet Union. I wanted to get a more diversified picture of the events than that delivered by CNN and the national papers.

I also realised that the way we access information had to change. We had to define our interests very clearly. Otherwise, we would fall off the ferry and drown in the deep ocean of information.

Now, we all surf the archipelago at almost the speed of light. Information continuously appears, changes and sometimes disappears. The need for focus has become greater and is vital to avoid drowning in the digital waves.

In this paper, I will try to share some of the lessons learned during six years of working with OSI (open source information) and give some suggestions for running a small OSI centre serving either the government or business sector, or both. The focus is on digital sources.

FULL-TEXT RETRIEVAL SYSTEM

This is basic! If you work with OSI, no matter what acronym you use to describe it, you must build your system around a full-text retrieval tool. All digital information is worthless if collected and stored but not accessible. Just compare a room filled with boxes containing 300,000 typed pages with one CD-ROM with an indexed and searchable copy of the same information.

With a digital, interactive system, you give the analyst time to digest and think. In the analogue system you never can be sure that the analyst will find the

---

[1] The author is a senior analyst at the Swedish Armed Forces Headquarters. His main responsibility is in R&D of open source intelligence and other related areas such as information warfare. He writes this paper in a personal capacity and he alone is responsible for the views expressed. The paper, therefore, does not reflect the official policy of the Swedish Armed Forces. He can be contacted at 73064.325@compuserve.com or by mail at Humlevagen 9, 18694 Vallentuna, Sweden

[2] The author wants to thank Mr Alan D. Tompkins for his comments and corrections.

needed information. The linear era of reading is a dying dinosaur. If you want to manage the speed and content of the information super highway (ISH) and the back alleys of the analogue towns, you also must add functions for media conversion. I think you must convert the analogue information you find in libraries into digital form.

ALL INFORMATION IN ONE FORMAT...

It may seem to be a mission impossible, but I think one must try to get the majority of the information used for analysis cross-referenced and searchable in one local document database. To achieve this, you first mine the external sources and collect or download all the information needed from various producers and providers. All of this information must then be stored locally in a single, common format. When new information needs arise, you first search your own, local document database, and, if necessary, again search the external sources more selectively and with pinpoint accuracy. The major benefit of a single-format, local document database is savings in time. In addition, it also gives the analyst the ability to discover duplicate information and to identify and use primary sources.

The "one-format" goal also includes media conversion like the scanning and OCR (optical character recognition) of paper documents, grey literature and books and the digitising of photos and maps and perhaps even 3D replicas of objects of interest.

INFORMATION ABOUT INFORMATION

If you use sources like newspapers, magazines and wire services, it is wise to learn more about the sources. Who owns them? What background do the journalists have? How long has the originator of an article studied the subject.

The same questions also apply to different providers and information brokers. It is vital to have good working relations with the providers in order to get information about new services and sources before they appear. It is only with that kind of early-warning system that you can plan your budget and optimise your collection assets in an organisation. In my experience, that kind of information also gives you a strategic advantage over your competitors who lack good relationships with providers.

Store your information about information in a database in order to be able to continually evaluate your sources and providers.

This *meta* information can reveal attempts at disinformation. In a study of a wire service, I discovered that one and the same journalist was an "eyewitness" at three different locations, over 600 km apart, in a period of one hour!

INDEPENDENT INFORMATION AGENTS

Independent information agents are tools that I have come to appreciate in my daily work. All true digital information often contains invisible information headers or bits of information describing the information that also give the user

the ability to automate the collection process in several different ways. The information agents will scan the incoming flow of information in a service and will select for review only those documents which meet a user-specified set of criteria.

The service providers have various approaches to using information agents. I will provide three examples of how I use these agents.

1. **Alert**- Dialog has a function called Alert that allows the user to define any word or subject or combination in almost any of the databases in Dialog. You can define the frequency for the agent's scan; daily, weekly, monthly or even every second day. You also define the way you want the collected information delivered; by e-mail, fax or even by *snail-mail* (regular postal service). I prefer e-mail because it arrives in digital form. Once received, it can be stored in a format of your choice and disseminated instantly to the analyst. As a very useful option, you can order your agents to deliver to different and/or multiple addresses.

2. **RBB stored archive search**- Reuters Business Briefing provides an option to store searches. This is perfect for routine and repeated searches, and those where you want to make *ad hoc* adjustments on-line. RBB has a very user-friendly GUI (graphical user interface) and makes adjustments to your search quick and easy. The drawback is that you have to manually connect every time you want to do your search. Reuters has other remote services that search automatically and deliver tailor-made newsletters by e-mail, like Dialog,. The major disadvantage with those systems is that you can loose control of your searches and they lack the interactive feel.

3. **Journalist**- This is a tool I use for monitoring my areas of interest in the field of information warfare, OSI, etc. on Compuserve. Given your interest profile, it can be set up to deliver text, photos, etc. from sources like Reuters, AP, UPI, the Washington Post, PC Week and many more. The result is delivered as a newspaper, on-line, with your own choice of type fonts and layout. This is a great application for following current information. It can be programmed to make your newspaper contain "The Latest Information on...." every X hours. Of course, Journalist, must be carefully set-up and there are still some functions that could be improved. But it saves time and can actually work as a budget early warning system.

A lesson learned in using agents is that you must be very precise and have very good knowledge of the specific subject's target vocabulary. You also must tune the system continuously. Identify the subjects that you can exclude. It is not always amusing to get the local results in Serbian soccer if you have an agent that is supposed to scan the ISH for the results of a negotiation process!

OPERATION CAPSEC (CAPTURE AND SECURE)

A great deal of important information is present on the net for only a short period of time. When you have identified a specific piece of information, you must capture it and store it in your own system. For as long as you continue to

want that information, you continue to monitor the source. If you use a local indexing system, you also get the benefit of having the information searchable and useable

LRRP-TACTICS

Every day there are between 400 and 1000 new WWW (World Wide Web) pages announced on the Internet. Every week, there are many new services announced on Knight-Ridder, CompuServe, the Microsoft Network (MSN), America OnLine (AOL), and by other providers. Some of these sites can be useful and valuable for your organisation. The traditional analyst does not have the time to explore all these new additions. I think one must assign some sort of LRRP that scans the forefront of the Information Superhighway and reports the location, content and development of valuable sites and services. The members of an LRRP can be *cybrarians*, librarians, or analysts. They must have very good knowledge of the organisation's basic needs in combination with a sixth sense or feel of what will become important. This scouting function serves as a sort of Help Desk for the analyst and the information miners.

PRIMARY SOURCES

In the analogue days of OSI, the analyst read documents in a hierarchical manner. Often, the source that gave the first report of an event guided the analyst to other primary and secondary sources. There must be changes in the way in which an analyst "read" the information. The analogue or hierarchical method represents a passive relationship to information. In contrast, the digital, non-hierarchical serach&read method is worthless if the analyst is passive and try to read the digital documents the traditional analogue way.

Digital information demands powerful, full text retrieval tools. Without them, collected information is inaccessible and useless. Success comes only by activity and, given the right tools, the intelligence output will be better both in quality and quantity. Working actively with the information - the results of your searches continuously gives you hints and clues to other important sources. The secondary reveals the primary... The thing is to identify primary sources and use secondary sources as an aid to finding the primary sources. Don't take for granted that the first, most current bit of information is accurate, but use it as a guide to other sources and prepare your organisation for "incoming!"

RECYCLE YOUR INFORMATION!

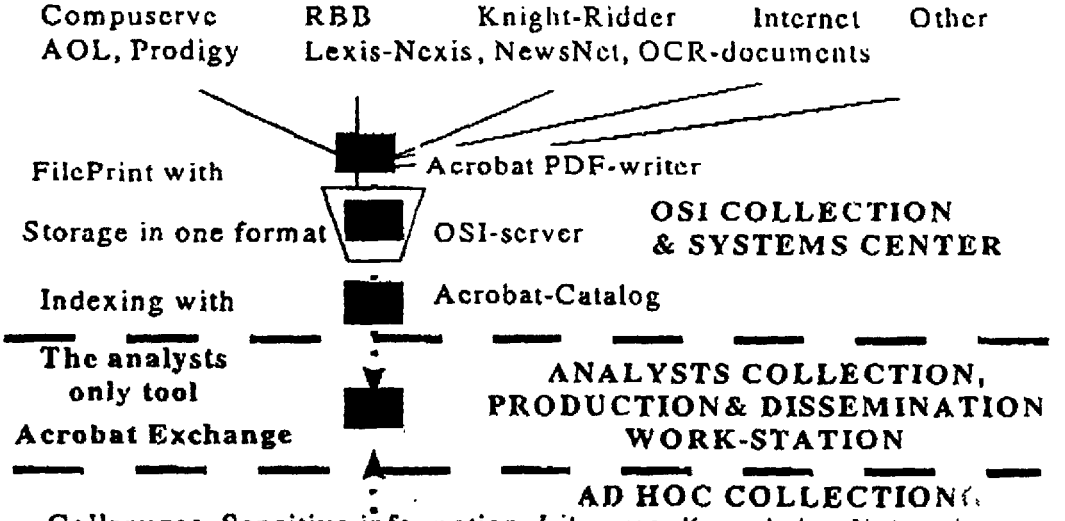
If you store the externally collected information in a local document database, you save money and increase effectiveness. If your organisation has relatively stable information requirements, you will soon discover that you often do not have to go on-line to external sources to meet them. In the analogue days of OSI, the reference library provided solid background material. That hasn't changed, but the library today can be digital and tailor-made to your organisation's needs.

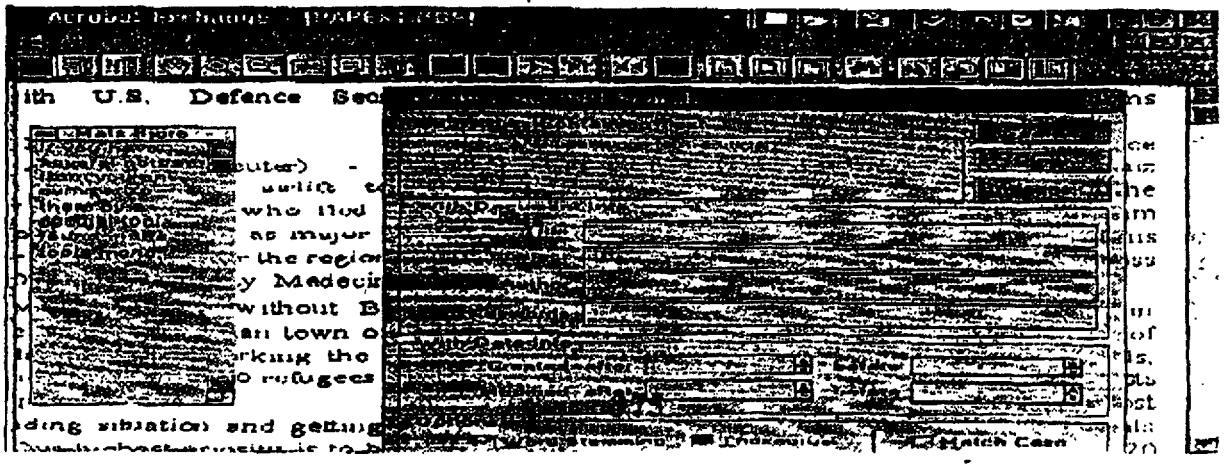
THE ANALYSTS ONLY TOOL...

One of the most important lessons learned is that the end user or analyst must have software that is simple to learn and contains a minimum of "wizard-features". If you are a *cyber soldier* or an information scout you are used to the rapidly changing software battleground. However, for the analyst, the need is not survival on the front line but rather a calm and stable environment for analysing the information and producing intelligence. If you change the operating environment for the analyst too often, part of their days will be devoted to *taming the tool*. Do not waste their time with unnecessary options and new versions of software until there is a real demand for new options.

One of the best tools I've discovered is the Adobe Acrobat system. It is platform- independent software which incorporates the Verity Topic search engine. Acrobat stores documents in their original format. It also includes useful utility programs for adding and reading comments on stored documents. Acrobat also allows hypertext linking within and between documents. Acrobat is a cheap and versatile alternative for any organisation or company with a limited software budget. The retail price for Acrobat is about $1300 for a license for ten users and one administrator.

Figure 1 describes the overall function of Acrobat in the OSI process and figure 2 is a screen shot of the working area of Acrobat Exchange.





SOURCES

If your organisation has a limited budget and only one or a few OSI personnel, you must be very careful in selecting information providers. During the last few years, I have found some very useful providers which, in combination, cover the needs of both government and business analysts..

Reuters Business Briefing

RBB is the service with the most user-friendly Windows-based GUI I've seen so far. RBB contains over 2000 sources including the BBC Monitoring Service, Jane's Defence Weekly, IDR, Reuters, PROMT, major UK papers, the Reuters Photo Archive, etc. The service has excellent business features which can be used to scan your target competitors or monitor the stock-market. RBB also allow you to search text and photos.

Dialog & Data Star

These two services are being merged into Knight-Ridder. The sources range from technical documents, medicine, chemistry, and patents to newspapers and transcripts from press conferences around the world. One of my personal favourites is the Federal News Service transcripts of press conferences in Moscow and in Washington D.C. As an example, I read the hearings about IC21 and follow the speeches in FSB from a single source.

A couple of years ago, when I was a teacher at the Army Intelligence School, one of my friends in the garrison wrote a book about survival in the Arctic wilderness. We did 90% of the basic research through Dialog and Data-Star. It took us one hour to gather all the basic information. When the book was finished, he said that he had saved perhaps 6-8 months of tedious work by that one digital hour. The wide variety of sources in Dialog and Data-Star covered all the needed information; from pharmaceuticals to news about accidents in an arctic climate. The service also has the Alert options described earlier.

CompuServe

CompuServe offers 3000+ sources and access to e-mail, voice-mail, the Internet and a large amount of software ready to be downloaded. With the addition of Journalist, you can use CompuServe as your personal newspaper editor. CompuServe offers local access in almost every nation throughout the globe. This option I find very useful. No matter where you are, you can have your laptop OSI-centre at your fingertips. You can collect, structure, analyse and disseminate with a tool you're familiar with. The voice-mail function have added a new dimension. Now you can distribute recordings with your comments both in voice and text at the same time.

INTERNET

With perhaps 500-1000 new home pages announced on the WWW (World Wide Web), it is almost impossible to know where to begin searching. Here, the use of LRRP tactics is important. Some say that most of the information on the

Internet is useless. They haven't been surfing in the right neighbourhood! If you are part of a large organisation with many information gatherers, it is wise to co-ordinate their bookmarks, i.e. pointers to locations to be searched, and to establish procedures to create the organisation's own source pages every second week or so. If you do so, you don't waste valuable time in the collection process or duplicate effort. And don't forget to use Listserv-the electronic subscription tool.

NEVER FORGET THE SOURCE AND THE SOURCE THAT PROVIDED THE SOURCE....

The quantity of information in the world is doubling every 12-18 months, and it is very easy to loose track of the information you've collected. The best way to store information is to keep the data in its original format, with all references to the source, and, if you have the time, add the name of the provider, the URL (uniform record locator), etc. to the document description stored with the document. When I use the Adobe Acrobat system, I also get the benefit of reading the documents in the producer's or provider's format with all the subtle information that one can derive from the use of different fonts and layout. This is very important if you're analysing documents such as propaganda material.

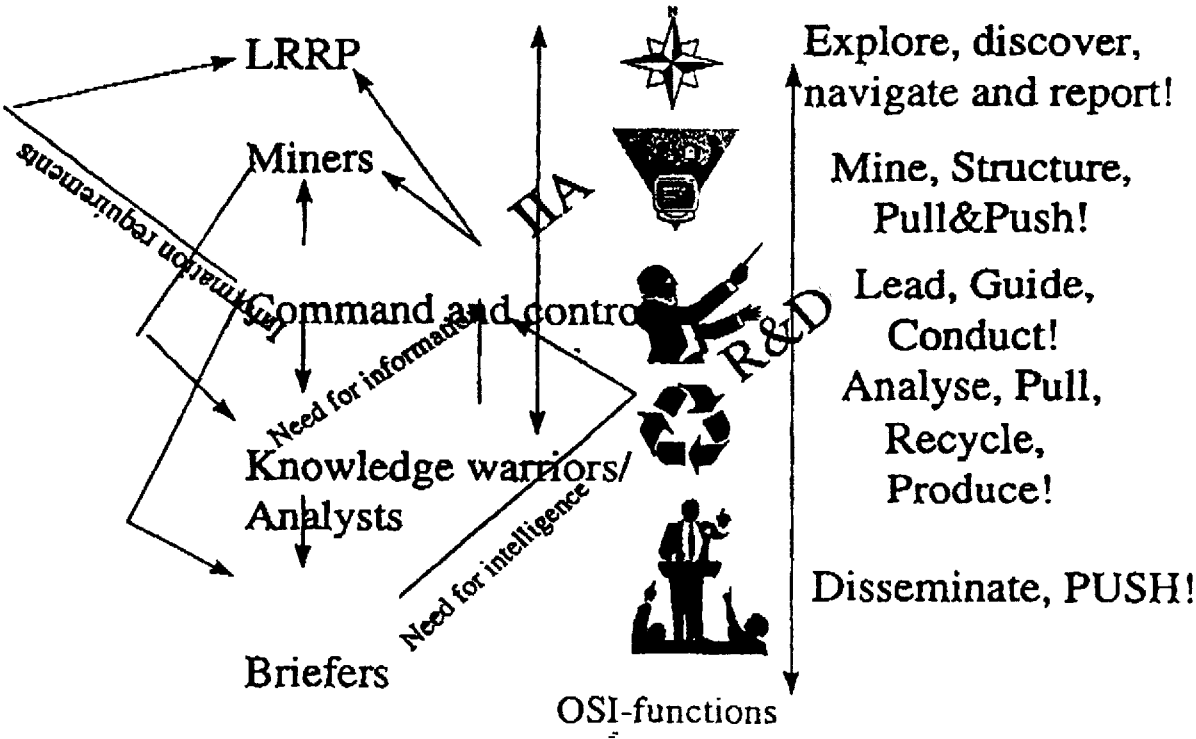
R&D

The rapid pace of development on the ISH demands a constant R&D effort in an OSI-based organisation. Organisations must have constantly evolving and changing systems. These changes may not always be visible to the analyst but the manager of an OSI system must understand and direct them.

As a metaphor, one can take a photo of wild river. The photo is a frozen projection but is very hard to describe in detail. You can paint a picture or write an essay about the river but you can not imagine how the waterfall will look in detail the next second, minute or decade. The only way to comprehend the flow of water is to dive into the water and follow the currents and try to master it during the ride.

I think the same applies to the ISH and OSI. You must actively take part in the process to understand the process;, to have visions and to create new methods and applications. In these areas, specialised R&D institutions are a phenomena of the past. If you have the resources, you must incorporate the R&D functions in the actual OSI process. Your organisation soon will be a *knowledge-organisation* where changes can be very rapid but invisible in the daily tasks. If everyone can influence his bit of the process, the word *changes* soon will loose its meaning

FUNCTIONS IN AN OSI-CENTRE



An OSI-centre can be everything from a laptop and a single account on CompuServe or AOL ,with $1000 budget with one person acting in all OSI-functions described in the text and in the figure, to the $1000000 multiple source organisation with specialists for each function. The only thing to keep in mind is that the challenge is to transform the information to intelligence!

This Article was prepared for OSS95 in Washington DC ,Nov 7-9.

The author can be contacted by E-mail 73064.325@compuserve.com or at the following address : Mats Björe, Humlevägen 9, S-18694 Vallentuna, Sweden.

OSS '95: THE CONFERENCE Proceedings, 1995 Volume II Fourth International Symposium on Global Security & Global Competitiveness: O - Link Page

Previous	Open Source Information Exploitation for Army Force XXI Summary
Next	Dr. Lawrence A. Farwell, Brain Fingerprinting

Return to Electronic Index Page