OSS'03 : BEYOND OSINT: Creating the Global Multi-Cultural Intelligence Web
15-19 September 2003, Washington, D.C.

Intelligence Librarian Tradecraft

Arno H.P. Reuser
mindef2@xs4all.nl

Abstract

A small OSINT support branch like the one of the Dutch Defence Intelligence and Security Service is highly dependent on automatic procedures - and software to support it - in order to process the required amount of digital data in such a way that the data can be easily handled by analysts and is suited for storage for text retrieval packages. This paper shows how the application of existing technologies and careful selection of international developments, such as Dublin Cor Meta Data, Digital Object Identifier, as well as writing one's own software in PERL to automate information management, without adequate assistance of IT, can greatly improve OSINT use and efficiency, leading to what might be called a "content management system do it your own".

1. Introduction

Setting up OSINT support and organising vast amounts of open source information can be quite a challenge in a small Intelligence service with an even smaller OSINT support branch. A demanding environment that puts emphasis on quick information, fast delivery, timely service and dedicated OSINT products, can only be supported by implementing automatic procedures and smart tools, and processes to arrange information in such a way that it can be easily retrieved and converted into products that can be processed quickly by analysts.

2. Background

The most important client within the Dutch Defence Intelligence and Security Service (DISS) is the Analysis and Reporting Division. They demand information on a wide array of topics, mainly (political) news, international relations and security, economy, business and defence affairs. Information needs to be delivered as soon as possible and presented in such a way that it can be processed in a minimum of time and, at the same time, the information should be in such a form that it can to be stored for future retrieval by for instance search engines. OSINT support consists of a mix of librarians and historians.

All DISS personnel have access to the DISS wide Intranet. OSINT personnel all have a second PC connected to the Internet and equipped with extra hardware to handle large amounts of information (ZIP drives, DVD burners, extra HDU space). There is no physical connection between the Intranet and the outside world, such as the Internet. The Internet machines are networked and connected to the Net through an 8 Mb ADSL connection. Extra software is installed specifically for OSINT purposes and for enhanced safety (double virus scanners, firewall, proxies etc.). OSINT personnel have *superuser* RWX rights[1] on their Internet machines in order to respond as fast as possible to changing environmental conditions. OSINT information is available on the Intranet for everybody of the DISS by means of a OSINT Web server. The OSINT website is constantly updated.

The amount of information currently handled by the analysis division is about 2.2 million documents (195Gb) with a monthly growth of 78,500 documents (7Gb). The OSINT electronic library resembles 1.4 million documents (37Gb) with a monthly growth of 75,000 documents (1.3 Gb). OSINT receives about 200 e-mails a day, mainly subscriptions via list servers and major domos.

3. Challenges

It is not difficult to see that a rather small OSINT team that has to maintain a modest library, reply RFIs, support analysis teams and update and maintain an Intranet OSINT website will have a challenge if it is also to handle such extensive amounts of digital information. Two approaches may be identified: the financial approach (quick, cheap and poor): no bibliographic control, simply "dump" everything on the network and let users find out for themselves, maybe with the assistance of some full text search engine, or, the librarian's approach (slow, expensive and rich) : full bibliographic control, description of all information on catalogue cards, adding meaningful index terms and maybe abstracts, thus enhancing the information value, before putting the information on the network.

The first approach is not to be preferred since without any bibliographic control the information would be "lost". Customers would need much time to retrieve information. Information would be hard to find, and hard to use, since a search engine that actually works is still to be invented. The second approach is much better from an information management point of view, however, (expensive) indexers are needed, lots of time required to add bibliographic data, and, last but not least, the amount of information added to the system each day makes it simply impossible to process everything manually.

[1] i.e. the right to read, write documents and to execute programs

The solution may be found somewhere in the middle, i.e. having some software that will - amongst other things - automatically add bibliographic data to information before publishing it, then generate and publish dedicated OSINT products from these data. Unfortunately, such software is not always available off the shelf. More often than not one is confronted with Content Management Software which most of the times is very expensive and very general in nature, such that the user needs to program for months and months to get any results.

Therefore, DISS OSINT decided to develop their own procedures and to write a simple set of programs to support it, based on international developments in the fields of DOI, SFX and Meta Data. These will be described in the remainder of the article.

4. Solution overview

The following approach to solving the problem of managing large amounts of digital data in a meaningful way was chosen:

4.1. Use wherever possible tools and readily available software to automate the transfer of information from one environment to another.

4.2. Design a *Digital Object Identifier* (DOI) system for each document . The DOI is a unique identifier of the document and serves as a replacement of the traditional URI. Since the DISS is a "discrete" environment that will never be connected to the outside world, one can freely experiment with the DOI regardless of international developments in this field.

4.3. Write a *Crawler* and create a database of DOIs and URIs. The *crawler* will periodically search the network for documents, calculate their DOIs and - together with the actual URI - add these to the DOI database.

4.4. Write a *resolver* that will read DOIs and translate those to URIs.

4.5. Write a universal *Parser* that will read documents, identify meaningful data, extract this data, structure the data and generate a "meta data" record.

4.6. Finally, write a (set of) *Publisher*(s) that will read the meta data and produce OSINT products for end-users, or, a program that will generate an up-to-date W3 website based on the meta data, or, whatever.

5. Digital Object Identifier

Simply putting documents on the network and having catalogues and HTML pages linking to them by URI poses some problems. First of all, the documents cannot be moved or replaced, otherwise the link dies. Secondly, deleting documents from the collection will lead to a dead link. Thirdly, linking by URI has a great risk of typing errors. Overall, some management scheme and linking database is needed if one is to use this method.

It would be much easier if there was an intermediate stage. Documents are no longer identified by URI but by an identifier, which is simply a unique number that will identify just this one document and no other, and HTML documents linking to the DOI instead of the URI. Some management is needed to handle DOIs. One could choose the international approach where each publisher assigns its own DOIs and is responsible for its own linking. In smaller environments such as the DISS a fully automatic scheme may be chosen.

6. Crawler and DOI resolver

A *crawler* is needed that will crawl the network looking for new documents. It will assign a DOI to new documents, take the actual URI and put both in a DOI database, or remove entries for documents that no longer exist. The crawler runs periodically, let say once every hour.

A *DOI resolver* is needed that is used to link DOIs to actual URIs. This program is executed from web clients and uses the DOI as a parameter. Once it is called by a client, it will read the parameter, look up the actual DOI in its database and redirect to the document.

Thus, the link

< ... action="http://a.b.c/cgi-bin/Resolver.pl?A20030404121223ffd">

would start the program *Resolver.pl* that in turn would read the DOI *A20030404121223ffd* that was send as the parameter. The program will use its database to look up the real URI and load that document in response to the client's request. The advantages of such a method are obvious. Users do not have to worry about how and where to store documents. The crawler will find them anyway. Furthermore, the risk of typing errors is greatly reduced and documents can easily be removed. Documents may now

freely be moved to other locations on the network. The crawler will find them and update the DOI database accordingly, thus "updating" all HTML pages linking to the document.

Example of a meta data record automatically extracted from a FBIS article.

```
<!-- DC metatags produced by Arno HP Reuser (re-
user@xs4all.nl) -->
<META NAME    = "DC.Coverage.Region"
      CONTENT = "East Asia"  >
<META NAME    = "DC.Coverage.SubRegion"
      CONTENT = "Southeast Asia"  >
<META NAME    = "DC.Coverage.Country"
      CONTENT = "Vietnam"  >
<META NAME    = "DC.Subject"
      CONTENT = "HEALTH"  >
<META NAME    = "DC.Date"
      SCHEME  = "WTN8601"
      CONTENT = "2003-04-07"  >
<META NAME    = "DC.Title"
      CONTENT = "Vietnam Lauded for Efforts to Contain
SARS"  >
<META NAME    = "DC.Source"
      CONTENT = "Hanoi VNA WWW-Text"  >
<META NAME    = "DC.Description"
      CONTENT = "Hanoi Apr. 7 (VNA) - Vietnam lauded
last week by the World Health Organisation for its ef-
forts to contain the spread of SARS has been keeping a
watchful eye on any possible outbreak of the disease."
>
<META NAME    = "DC.Publisher"
      CONTENT = "FBIS PIOS"  >
<META NAME    = "DC.Format"
      CONTENT = "text/html"  >
<META NAME    = "XDC.OrgFileName"
      CONTENT = "sep20030408000063n.html"  >
<META NAME    = "DC.Date.Modified"
      SCHEME  = "WTN8601"
      CONTENT = "2003-08-10T12:11:13"  >
```

7. Parser

Now that digital documents can be safely published and easily linked to, we can turn our attention to bibliographic control. A *parser* was written that will read new documents and identify origin or publisher. The parser will then

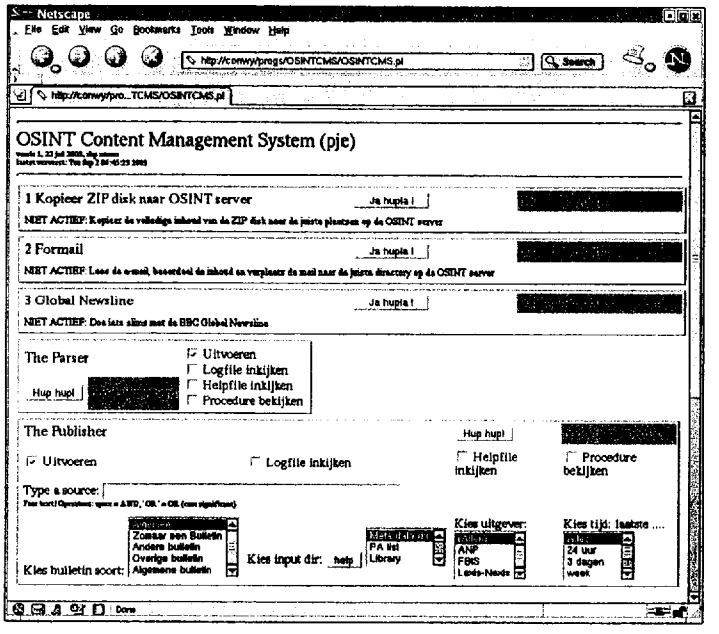
- examine the contents of the documents,
- identify individual articles within the document (if the document is in fact a digest of several articles)
- extract relevant data from each article,
- generate a DOI,
- generate a meta data record,
- publish the meta data record,
- and publish each article until the entire document has been processed, then continuing with the next document.

8. Dublin Core Meta Data

Meta data is the "modern" term for something that librarians and information professionals have been doing since the invention of printing in the late 15th century by Gutenberg[2] : cataloguing. Meta data is a crude derivative of the ISBD cataloguing rules, containing information about the document such as titles, authors, sources, dates, abstracts, and links to other related documents. The meta data is either added to the document, put into a separate database or both. If it is written to a database, the meta data record is linked to the DOI.

There are several international schemes for meta data of which Dublin Core Meta Data (DCMD) seems to be the most promising and useful one for this project. The scheme is designed by librarians to handle digital documents. The parser therefore uses this scheme to generate DCMD records.

Part of the Publisher HTML form used to produce FBIS Daily Bulletins



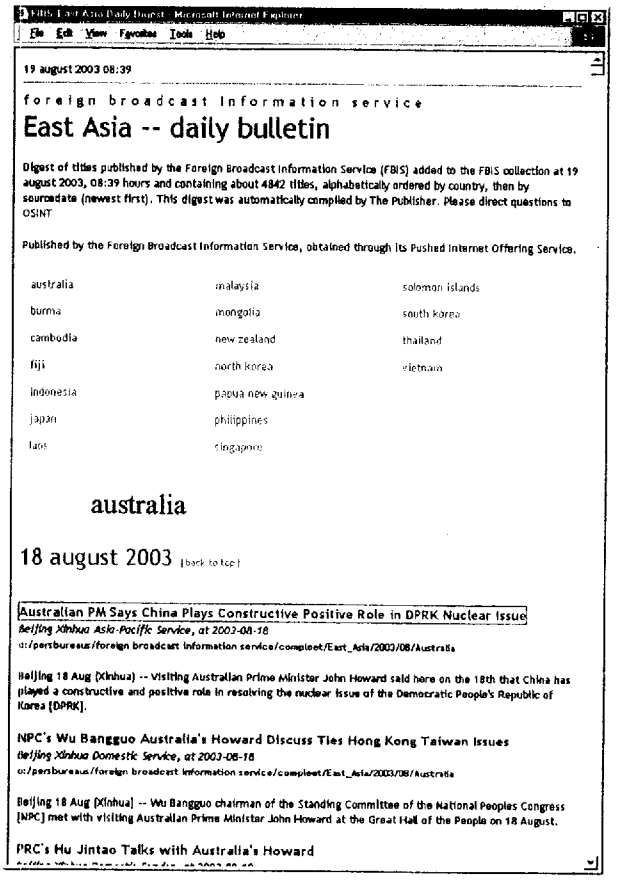
DCMD records are inserted into the documents between meta tags in the top of the document, and, records are written to a separate file, and, they are written to a database (MySQL). The parser itself is a small short program that will only identify the document, identify its type and characteristics and once it has done that, give control to a module that will parse the data from the document. It will also send a set of variables to the module that will decide output, input, length, duration and some other decisions. These are read from a *run control* file that will also give "instructions" on how and where to publish documents. For instance, it is possible to give directives on the directory

[2] Actually bookprinting was invented by Jan Laurenszoon Coster from Haarlem, Holland, but historians decided otherwise ;-)

structure on the network where the documents will be stored, or to limit the length of abstracts, or the length of titles.

The module will parse the document, generate a DCMD record, publish that, create a directory structure if necessary and publish the document. General remarks and especially errors are trapped in a log file. Documents that cannot be recognised or documents that for whatever reason cannot be processed are logged and written to a refusal directory.

Output by the Publisher: FBIS Daily Digest, clickable menu, titles, plus abstracts. One bulletin for each region.



19 august 2003 08:39

foreign broadcast information service

East Asia -- daily bulletin

Digest of titles published by the Foreign Broadcast Information Service (FBIS) added to the FBIS collection at 19 august 2003, 08:39 hours and containing about 4842 titles, alphabetically ordered by country, then by sourcedate (newest first). This digest was automatically compiled by The Publisher. Please direct questions to OSINT

Published by the Foreign Broadcast Information Service, obtained through its Pushed Internet Offering Service.

australia	malaysia	solomon islands
burma	mongolia	south korea
cambodia	new zealand	thailand
fiji	north korea	vietnam
indonesia	papua new guinea	
japan	philippines	
laos	singapore	

australia

18 august 2003 [back to top]

Australian PM Says China Plays Constructive Positive Role in DPRK Nuclear Issue

Beijing Xinhua Asia-Pacific Service, at 2003-08-18

o:/persbureaus/foreign broadcast information service/compleet/East_Asia/2003/08/Australia

Beijing 18 Aug (Xinhua) -- visiting Australian Prime Minister John Howard said here on the 18th that China has played a constructive and positive role in resolving the nuclear issue of the Democratic People's Republic of Korea (DPRK).

NPC's Wu Bangguo Australia's Howard Discuss Ties Hong Kong Taiwan Issues

Beijing Xinhua Domestic Service, at 2003-08-18

o:/persbureaus/foreign broadcast information service/compleet/East_Asia/2003/08/Australia

Beijing 18 Aug (Xinhua) -- Wu Bangguo chairman of the Standing Committee of the National Peoples Congress (NPC) met with visiting Australian Prime Minister John Howard at the Great Hall of the People on 18 August.

PRC's Hu Jintao Talks with Australia's Howard

The meta data generator uses international standards as much as possible. Dates for instance are formatted according to the ISO norm 8660. The great advantage is that search engines such as dtSearch recognise and can handle 8660 dates thus making it possible to search on/for dates instead of just text strings. Sometimes extra fields are needed in the case of for instance country and region names. These extra fields are prefixed with a capital X.

Once the DCMD record is ready, control is handed over to yet another module that will do the rest: create a valid filename for the document, check for and if necessary create a directory structure in the target environment, publish the DC record, publish the article, update the log file and alert whenever an exception was found.

9. Publisher

Finally, some software is needed to actually present information to the customer in such a way that the customer can actually do something useful with it. A *publisher* was written that will read the DCMD file, extract all data, and publish a HTML document such as a current awareness bulletin. The Publisher is directed by a run control file that holds all variables. This enables the product to be used to make almost any bulletin for whatever purpose.

Since filling in variables in run control files is not very user- friendly, the Publisher too, just like the Parser, is controlled and run by a HTML form with variables available in pull down menus, lists and fill-in fields with comments and explanations if necessary. The publisher thus reads variables selected by the operator in the form, but can also be used in full automatic mode by enhancing the HTML form with hidden HTML tags since most of the products will look the same every time they are produced.

The publisher has some extras to enhance usability for the operator. For instance, little red or green flags will indicate when the service was last run, if at all. The operator will see in a glance when to run which service.

10. Overall Information management process.

Most digital information acquired by OSINT is collected by automatic procedures and software. Information can be downloaded automatically by using offline HTML browsers (Teleport Pro), or Copernic Pro, as long as they have the functionality to save, schedule and execute queries. Listservers and major-domos will send information automatically to any e-mail address. Local software is used to interrogate push servers. All this information is collected on the network drives of the standalone Internet PC.

A simple MSDOS batch file is used to copy all downloaded or received data from the network to a ZIP disk. Execution of this file can be automatic by using the NT4 scheduler (AT command). The ZIP disk needs to be moved to the Intranet PC, then, another MSDos batch file is used to move all data from the ZIP disk to the OSINT server.

Finally, Parser and Publisher, as well as some other scripts are run daily or periodically to process the information and publish user-friendly end products. Most of the entire process is thus automated except for the inevitable "air gap" to swap discs or CD's from a black drive to a red drive.

11. Results

Since the meta data records contain structured data ready for use, and since there is now an extractor to read the DCMD records and extract relevant data form them, and since it is an international standard, almost any product can now be made without much effort. All one has to concentrate on is how the bulletin should be formatted. Dates, for instance, are entered according to the international ISO norm, which means that well-known search engines such as dtSearch can recognise them and create indexes from date fields, thus enabling the user to search on real dates instead of date strings.

The OSINT products are preferably HTML files. These used to link to documents somewhere on the network, but now they link to the DOI resolver that will look up the real URL and load the corresponding document. The *crawler* is used during the parsing process, but will also run on its own as a daemon to look for documents that have been moved or removed. If it finds any, it will update the DOI database.

Almost the entire process of digital data processing (from collecting, acquisition, indexing, publishing and dissemination) is automated. Handling 10,000 documents is done in about two minutes (parsing, publishing, generating). Currently, the parser recognises and can process FBIS PIOS documents, BBC Global Newsline, ANP Dutch presswire agency files, LEXIS-NEXIS files, Factiva HTML and plain e-mail files. For each new document type, a new module needs to be written, but since most of the work is contained in general modules that are applicable regardless of document type, all that needs to be done is to add new characteristics to the parser to make it able to identify the new document type, and a parsing module. All the remaining procedural work is already there.

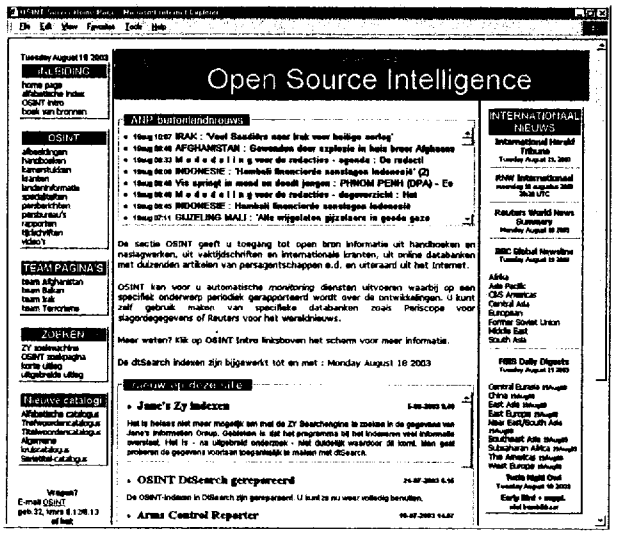
For instance, extracting, parsing, publishing and disseminating FBIS PIOS documents (about 2,500 per run) takes in total about 10 minutes, most of which is lost due to low network band width.

DOI processing is currently under development and is not yet fully implemented. Work is now in progress to write the meta data records to a SQL database and to increase the user-friendliness of the publisher.

12. Future enhancements

Some ideas to further improve the functionality are to translate country and region names into one single language, at the same time also solving semantic problems. A controlled vocabulary - maybe an authority file - would be needed to translate uncontrolled terms into controlled terms. The same principle and procedure can also be used for the listing of source names and publishers.

The end result: the OSINT home page is the central gateway to all DISS OSINT information.



Another idea that came up during the development of the project has to do with documents that do not have extractable country or region names. Maybe it is possible to write a module that will try to determine, based on personal names, names of cities etc., the country or region involved and add those names to extra DCMD fields.

13. Conclusions

One of the biggest gains of the project is its simplicity and universal nature. It can be applied to any type of document, open source, classified, images, etc. All that is basically needed is the meta data procedure and strict rules on how to apply them. It is now possible to present to the end-user true multi-disciplinary products regardless of origin.

This project is not new, nor is it original. All ideas are based on international developments, the progress of which can easily be tracked by using for instance the Internet. The significance of this project

lies in the fact that more or less advanced automation can be done without the acquisition of expensive and cumbersome software and without hiring even more expensive programmers.

A major drawback lies in the fact that it is in fact a derived index, where an assigned index, based on a controlled vocabulary, would be preferred in order to improve retrievability and indexing.

With a little effort, librarians too can learn how to program without the "assistance" of IT. If there is a lesson to be learned, it is that running small OSINT support branches is hardly possible without writing programs and tools yourself.

The major advantage of this project is that within the framework of content management system selection, a lot of experience is already at hand, thus making selection of a suitable CMS easy. This is also a major disadvantage. Too much knowledge of what a CMS can do makes manufacturers' life a little cumbersome.

OSS '03 PROCEEDINGS "BEYOND OSINT: Creating the Global Multi-Cultural Intelligence Web" - Link Page

Previous	OSS '03 Ran Hock The Open-ness of the Internet,
Next	OSS '03 Steve Edwards Open Source Intelligence Gathering Within the UK Police National Intelligence Model (Text),

Return to Electronic Index Page