ADVANCED DATA EXTRACTION AND PREPARATION VIA TIPSTER (ADEPT)

John Kielty
Ira Sider

Lockheed Martin Corporation

P.O. Box 8048
Philadelphia, PA 19101

kielty,sider@mds.lmco.com

1. BACKGROUND

Shrinking budgets, reduction in personnel and increasing performance expectations are facts of today's Government environment. The intelligence community has seen these trends stress its ability to collect, manage and disseminate information in an efficient, timely manner. The need to respond to rapidly changing statements of operational interest only exacerbates the problem. It is no longer sufficient to monitor certain pre–defined individuals or interests. A proactive collection and indexing service must store virtually all types of information to be prepared for any eventuality.

The Advanced Data Extraction and Preparation via Tipster (ADEPT) Program [Contract Number 95–F147400–000] is a TIPSTER II demonstration project aimed at alleviating problems currently being faced by the Office of Information Resources (OIR). OIR has been chartered to implement enterprise–wide systems to collect, archive, distribute, and manage information obtained from unclassified data sources typically called "Open Sources."

In support of its charter, OIR implemented ROSE, which provides a full range of information management tools that support document input, tagging, archiving, retrieval, and dissemination. The ROSE system is handling an increasing volume of data from disparate sources having widely varying formats. Acquisition of new sources or accommodation of format changes in existing sources can have considerable cost and effort impact on OIR operations, since existing software and procedures often require modification. ADEPT will address the potentially high costs and delays incurred when adapting the current ROSE data preparation module to handle format changes and variations in new or existing sources in an automatic and systematic fashion improving the responsiveness and offering flexibility to OIR's user community.

2. CAPABILITIES

ADEPT was conceived as a vehicle for capabilities to alleviate problems currently being faced by OIR. ADEPT tags documents in a uniform fashion, using Standard Generalized Markup (SGML) according to OIR standards. ADEPT provides a friendly user interface enabling Data Administrators to easily extend the system to tag new document formats and resolve problems with existing document formats.

Data Processing and Extraction: ADEPT processes both well–formed and ill–formed data; accepting raw documents and parsing them to identify source–dependent fields that delineate specific important information. Some of these strings will be normalized. The field names, field values, and their normalized forms are stored as annotations along with the document in a TIPSTER compliant document manager. An SGML tag, defined by OIR, is associated with each annotation. The SGML tags delineate predefined document segments, such as title, publication date, main body text, etc. If ADEPT correctly captures all the fields for a documents format, an SGML–encoded document is transmitted to the ROSE System for information dissemination.

Problem Detection and Diagnosis: ADEPT recognizes problems in the input documents and, offers deep diagnostics and suggestions to the Data Administrator for fixing those problems. Although new sources, format changes and erroneous or ill–behaved data can cause processing errors, ADEPT identifies these problem occurrences, generating diagnostics that describe the nature of the problem, such as where it occurred and why it did not match. From the diagnostics, the Data Administrator can easily determine whether the problem is due to an error (anomaly) in the data or a change in format.

Error Handling and Document Viewing: ADEPT maintains a problem queue and provides GUI windows to aid the Data Administrator with both evaluating the source of problems (data error or new/changed format) and resolving them. The GUI enables a Data Administrator to see the original document, the output SGML template and the fields from which the SGML tags were generated. A Data Administrator can manually change the value of a tag and resubmit resolved document(s) for reprocessing by the system.

System Adaptation: ADEPT enables Data Administrators to manually adapt the system's configuration (mapping templates) to meet new or changed formats. Through a combination of menus, customized panels and, cutting and pasting operations, the Data Administrator can specify the instructions to be used by ADEPT to parse and extract data from incoming documents.

3. SYSTEM ARCHITECTURE

Figure 3–1 illustrates how ADEPT will be inserted into the Rich Open Source Environment Version 2 (ROSE) testbed environment at OIR. After a successfully evaluation, ADEPT may be made operational.

ADEPT will be connected to the ROSE–Feed servers via a 16MB/second Token–Ring Local Area Network (LAN). These servers receive streams of documents from currently five sources/providers: NEXIS, DIALOG, DataTimes, FBIS and Newswire. Refer to Appendix A for a sample document example. After successfully parsing and extracting document required information, ADEPT will transmit a SGML Tagged document over a one–way fiber to the ROSE–Catcher where the information will be archived and disseminated to the OIR user community. Refer to Appendix B for a processed document example.

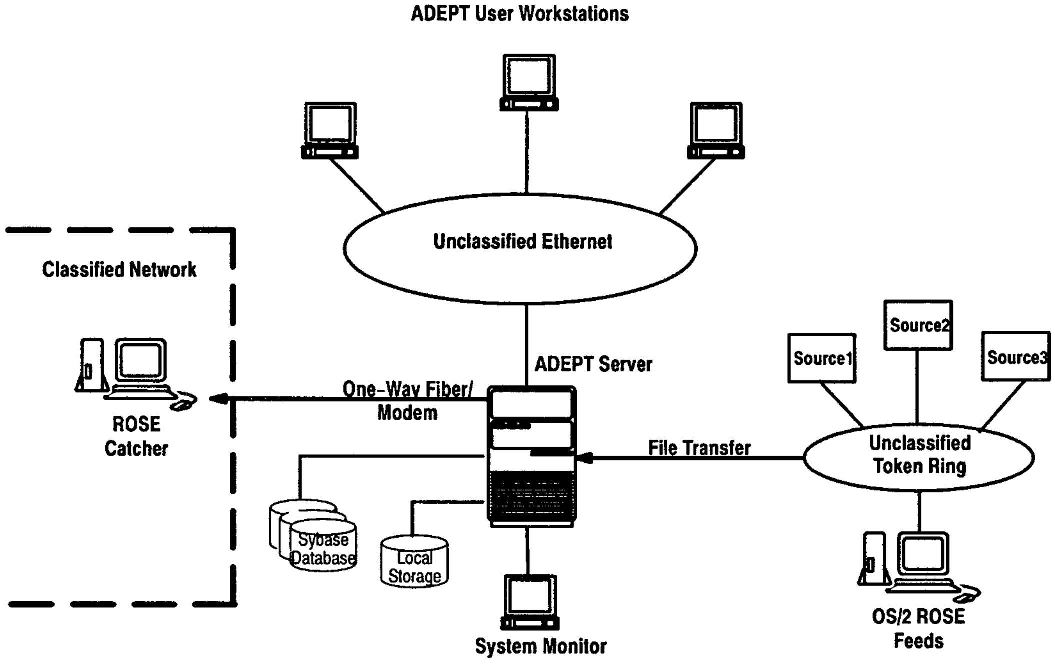


Figure 3–1. ADEPT System Architecture

ADEPT will have the ability process more than one thousand separate sources from the five current OIR providers, at an average of 80 megabytes and a maximum of 150 megabytes per day currently. These figures are estimated to increase by twelve percent per month. Over an average month, ADEPT will operate seven days per week processing and expected 600,000 documents.

Appendix C depicts the SGML tags which will be identified by ADEPT.

4. SYSTEM DESIGN

Figure 4–1 illustrates the design of ADEPT. ADEPT is comprised of eight processes; each performing a specialized task. These processes are: the Document Input (DI), the Document Processor (DP), the Document Management (DM), the Management Information System Manager (MISM), the Problem Queue Manager (PQM), the System Adaptation Manager (SAM), the Administration Manager (AM), and the Output Manager Function (OM).

4.1. Document Input (DI)

The DI process is the interface between ADEPT and the ROSE–Feed servers. Based on the source, a mapping template is selected. The DI identifies and

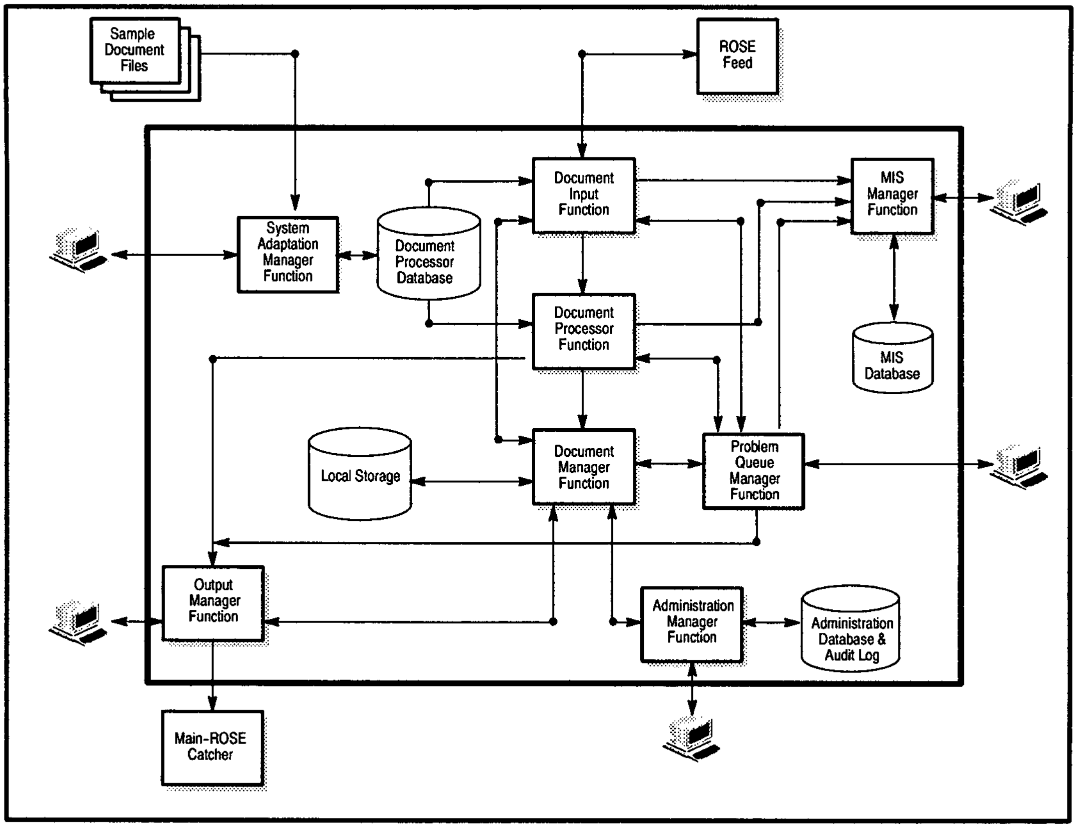

Sample Document Files
ROSE Feed
Document Input Function
MIS Manager Function
System Adaptation Manager Function
Document Processor Database
Document Processor Function
MIS Database
Local Storage
Document Manager Function
Problem Queue Manager Function
Output Manager Function
Administration Manager Function
Administration Database & Audit Log
Main-ROSE Catcher

Figure 4–1. ADEPT Document Parser

separates the ROSE Feed stream into documents. The document and its relevant information is stored in local storage via the DM function calls.

If the mapping template can not be identified, the stream probably came from a source unknown to ADEPT. Unknown sources are sent to the Problem Queue to for user intervention.

4.2. Document Processor (DP)

The DP identifies and extracts all SGML tags defined in the mapping template for the specific source. Each identified field value is validated and normalized (if required) before being stored as annotations with the document via DM function calls. DP creates annotations with the value 'NA' (Not Available) for those non–required SGML tags not present in the document.

If while processing, DP is unable to identify a required SGML tag, validate or normalize its contents, the document is identified as a problem document. DP does not stop processing the document once encountering an error. It completes the document processing; identifying any remaining errors. For each problem SGML tag, DP generates diagnostic information. The diagnostic information contains an error explanation as well as suggested corrective actions. Problem documents are sent to the Problem Queue to await analysis.

4.3. Document Manager (DM)

The DM, the heart of ADEPT, is composed of a set of library routines providing a standard interface between ADEPT and the collections of documents in persistent storage. The DM is TIPSTER compliant and utilizes Open Database Connective (ODBC) to store document and document relevant information in the Sybase System 11 database. ODBC adds an additional layer of flexibility to DM. With ODBC, the Sybase System 11 database can be substituted with any ODBC compliant database on any platform.

4.4. Management Information systems Manager (MISM)

The MISM process manages the quantitative MIS Statistical data used to monitor and evaluate ADEPT. MISM records the document's name, source, date/time stamp, and other relevant information when a:

- Document is received by ADEPT,
- Document is successfully tagged,
- Problem document is identified, and

- Document is transmitted to Main–ROSE Catcher.

Additionally, ADEPT captures similar statistics on problem types and problems associated with each document. The ROSE Data Administrator can perform simple queries and execute quick reports against the collected data.

4.5. Problem Queue Manager (PQM)

The PQM is responsible for managing the problem queue of ADEPT. The problem queue is a visual representation of all problem document information contained in the database. An entry exists for each problem document; it contains the document identifier, source, problem class, status, mapping template identifier, date/time stamp, etc.

At the ROSE Data Administrator's discretion, documents in the problem queue can be sorted and limited by either source, date/time stamp, problem class, mapping template and status.

To investigate/resolve a problem document, the desired document must be selected. For each document selected, the document viewer GUI is invoked. The GUI displays: 1) the original document, 2) the current version of the SGML template for that document, 3) the linkages between the two, 4) diagnostic information associated with the document, and 5) suggestions for fixing the problem tag(s).

The document viewer allows one to modify problem tags based on system supplied corrective actions. If system suggestions are rejected, tag values can be generated from user supplied data. For cases where the original document trigger is garbled due to a transmission error, the user can elect to define a temporary trigger. Notes can created and saved for each document.

After the problems associated with a document are addressed, the document can be resubmitted to the system for reprocessing. PQM functions provide the user the ability to select and resubmit multiple documents.

4.6. System Adaptation Manager (SAM)

The SAM process provides the capability to create, modify, and associate mapping templates with a specific data source. A mapping template contains the directions on how to parse a specific data source. It specifies the SGML tags (i.e., Pubdate), whether the tag is required and any associated field names (triggers within a document) which must be used to extract the SGML tag value as well as corresponding format validation and normalization rules. There is one primary mapping template for each data source received by ADEPT.

Once created, SAM allows the Data Administrator to test their mapping template changes against sample files of documents.

4.7. Administrator Manager (AM)

The AM manages the routine system administration of ADEPT. AM provides login control and user permissions, maintains the system's security and audit log, and enables backups/restores of the system databases.

All user interaction (system adaptations and problem queue manipulation) performed by the user are recorded in the AM's audit log including a record of the change, user identification, and date/time stamp. Both the security and audit logs can be viewed via the AM GUI.

From the AM GUI, the user can authorize others to print, display, search, consolidate, and delete the computer security audit log as well as add, delete or re–enable accounts by changing user permissions.

4.8. Output Manager (OM)

The OM manages the output of successfully tagged documents for ADEPT. The OM's main capabilities include:

- Creation of the SGML tagged version of the document,
- Performing "Special Processing" (when required),
- Providing an interface for passing the tagged document to the Main–ROSE Catcher,
- Providing a GUI which will allow the ROSE Data Administrator to view the original document, the final tagged document and the linkages between the two for any document stored in local storage.

OM retrieves successfully processed documents. For each document, OM walks through the annotations (SGML tags) accessing their associated SGML tag value. The set of SGML tags with their corresponding value constitute the SGML template for that document.

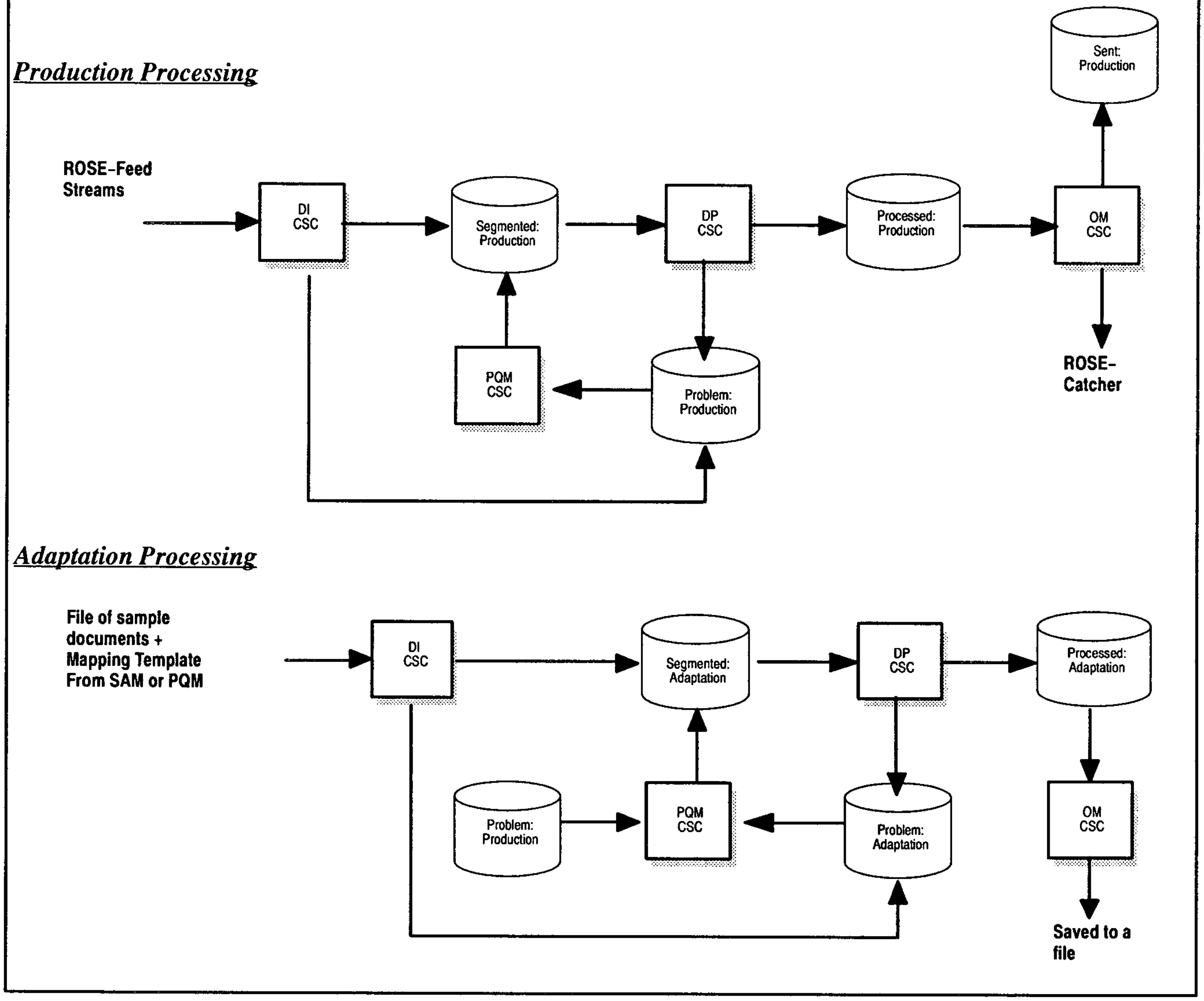
If the document is initially from the ROSE–Feed, OM will send the SGML Template, conforming specific protocol, to the Main–ROSE Catcher. Successfully processed sample documents are saved to a UNIX file for future review.

5. SYSTEM PROCESSING

Information is passed to each process via collections stored within the TIPSTER compliant Document Manager (DM). Collections act as the queues for the processes. A collection contains the information necessary for a process to perform (i.e., documents and document relevant information). The DP, PQM, and OM processes each have a unique collection associated with it. A process begins by accessing the first document in its collection. When completed, the document is moved to another collection for the next process to continue. Since a document moves from collection to collection, each process only depends upon the documents in its collection.

As depicted in Figure 5–1, there are two categories of collections: production and adaptation. ROSE–Feed supplied documents are processed in the production collections. Adaptation testing as well as documents from sample files are processed in the adaptation collections. These two categories of collections will clearly separate adaptation documents from production documents. Documents in the production category will run at a higher priority than those in the adaptation category. Prioritizing enables ADEPT to process both categories of collections concurrently.

Figure 5–1. ADEPT System Processing



6. STATUS

The ADEPT project has completed the System Requirements Review (SRR) as well as the Preliminary Design Review (PDR). A Critical Design Review (CDR) is scheduled for late June 1996; to be followed by a TIPSTER Engineering Review. ADEPT will be installed in OIR's testbed environment in December 1996 where it will undergo a three month evaluation period. After a successful evaluation, OIR will have the option to transition ADEPT to their production environment.

APPENDIX A: SAMPLE RAW DOCUMENT

```
ACCESS # FINP2407547
HEADLINE Bre–X discloses new drill results
      Column:  COMPANY NEWS
      ESTIMATED INFORMATION UNITS: 1.7    Words: 124
DATE    04/06/96
SOURCE  * The Financial Post   (FINP)
      Edition:  Weekly
      Section:  1, News
      Page:    23
      Category: NEWS
      (Copyright The Financial Post)
RE      NME CN
—           Bre–X discloses new drill results           —

Calgary–based Bre–X Minerals Ltd., discoverer of a potentially
huge gold property in Indonesia, disclosed new drill results late
Thursday.

....

ADDED KEYWORDS: GOLD; MINERAL EXPLORATION; INDONESIA

CORPORATE NAME: Bre–X Minerals Ltd. (T/BRH)


             *** Infomart–Online ***

End of Story Reached
```

APPENDIX B: SAMPLE PROCESSED DOCUMENT

```
<DOC>
<SEQ>DTT-96-00115385</SEQ>
<DATE>19960408</DATE>
<TITLE>Bre-X discloses new drill results</TITLE>
<AUTHOR>NA</AUTHOR>
<PUBNAME>The Financial Post</PUBNAME>
<DOCAT>NA</DOCAT>
<DOCTYPE>NA</DOCTYPE>
<FILENO>FINP2407547</FILENO>
<PUBDATE>19960406</PUBDATE>
<PUBNO>Section 1, News Page 23</PUBNO>
<DLANG>NA</DLANG>
<OLANG>English</OLANG>
<PUBLISH>Financial Post Ltd.</PUBLISH>
<SECURITY>UNCLASSIFIED</SECURITY>
<SOURCE>DATATIME</SOURCE>
<IC>CIA</IC>
<SUMMARY>Calgary-based Bre-X Minerals Ltd., discoverer of a potentially
huge gold property in Indonesia, disclosed new drill results late
Thursday.
</SUMMARY>
<KEYW>NA</KEYW>
<SHEAD>
File: 19960408.tst.src
ACCESS #: FINP2407547
HEADLINE: Bre-X discloses new drill results
Column: COMPANY NEWS
ESTIMATED INFORMATION UNITS: 1.7    Words: 124
DATE: 04/06/96
SOURCE: * The Financial Post  (FINP)
Edition: Weekly
Section: 1, News
Page: 23
Category: NEWS
Copyright: The Financial Post)
RE: NME CN
Title: Bre-X discloses new drill results
</SHEAD>
<BODY>
Calgary-based Bre-X Minerals Ltd., discoverer of a potentially
huge gold property in Indonesia, disclosed new drill results late
Thursday.

....

ADDED KEYWORDS: GOLD; MINERAL EXPLORATION; INDONESIA
CORPORATE NAME: Bre-X Minerals Ltd. (T/BRH)
Copyright: The Financial Post
</BODY>
</DOC>
```

APPENDIX C: SGML TAG LISTING

SGML Tag	*Description*	*Req.*
<DOC>	Signals start of SGML tagged document	*
<SEQ>	Unique ID # of document, internally generated	*
<DATE>	Date document received by ROSE-Feed	*
<TITLE>	Title of document	*
<AUTHOR>	Author/editor of document	
<PUBNAME>	Name of publication (e.g., New York Times)	
<DOCTYPE>	Type of information. (e.g. journal article, news story)	
<DOCAT>	Document category (e.g., full article, abstract)	
<FILENO>	Source-specific information	
<PUBDATE>	Date of publication of the document	*
<PUBNO>	Publication number	
<DLANG>	Language of document as received	
<OLANG>	Original language of document	
<PUBLISH>	Company responsible for publication	
<SECURITY>	Security information & controls (e.g. For Official Use Only)	
<SOURCE>	ROSE-Feed source (e.g. NEXIS, DIALOG)	*
<IC>	Intelligence Community agency through which the document is distributed (CIA)	
<SUMMARY>	Summary of document as provided by source; else extracted from first part of body	*
<KEYW>	Words pertaining to subject/content of document	
<SHEAD>	Additional source information, not otherwise tagged (header and trailer of document)	
<BODY>	Signals the beginning of the document text	*