EUROPEAN COMMISSION
DIRECTORATE GENERAL JRC
JOINT RESEARCH CENTRE
Institute for the Protection and Security of the Citizen
Support to External Security

Europe Media Monitor

- System Description –

Clive Best, Erik van der Goot, Ken Blackler,
Teofilo Garcia, David Horby
Web Intelligence Action
SES Unit, IPSC

T.P. 267
I – 21020 Ispra (VA) Italy
Tel: + 39 0332 785044
Fax: + 39 0332 789185
Email: clive.best@jrc.it

December 2005

EUR 22173 EN

European Commission
Directorate-General Joint Research Centre

EUR 22173 EN
ISSN 1018-5593
Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

Table of Contents

Figures

1 Background

The Europe Media Monitor (EMM), has been developed by JRC on behalf of the Press and Communications Directorate of the European Commission (DG PRESS). DG PRESS drive it's development by specifying their requirements at regular monthly meetings. Since January 2005 this service has been formalised through an Administrative Arrangement between DG PRESS and the JRC. EMM provides the software services which process incoming News Reports from News Agencies, Press Reviews from Capitols and the major web based news services in Europe. EMM began operations in May 2002 and has expanded it's services over the intervening years to become the Commission's primary source of live news related services. The front end web interface is through an automatically generated NewsBrief, which is updated every 10 minutes with the latest top stories. The internal Commission version of the NewsBrief is accessed by up to 20,000 users per day. The external public version of EMM which contains only open sources has found a growing number of customers in Europe and elsewhere. However, this main service is only one of a suite of other services that EMM provides. This document aims to describe these services and document the software architecture of EMM.

2 EMM Services Overview

A summary of the main services provided to PRESS by the EMM software is given below. Each component is then described in detail in later sections.

2.1 News Brief

The EMM News Brief is the main view of the current live news filtered by the EMM software. It is designed for efficiency and multiple accesses from many users. It provides a live snapshot of all the current contents of the Alerts and a summary of the top Breaking News Stories in each language. EMM NewsBrief provides a webv interface to view thye latest edition of EMM Panorama produced by the RNS system and access to all the latest Press Reviews immediately they are published on the Rapid News Service.
Full Details are described later

2.2 Rapid News Service

Since the summer 2004 the Rapid News Service derived from EMM has been used to edit and publish the twice daily EMM Panorama and to alert cabinets and spokespersons of breaking news. This is the main tool used by PRESS staff on duty from early each morning. RNS is currently being updated for the following extra functions

- To allow the definition of Alerts to be modified and defined.
- To allow DGs to customise their own newsletters and edit Alert definitions
- To provide an editorial interface to the Press Cuttings reviews. This editorial interface unites cuttings found in representation press reviews and those entered centrally in Brussels.

Full details are described later.

2.3 EMM Press Review

EMM Review is a web based publishing system used by the representations and delegations to remotely publish the daily press reviews from Capitols. This system was developed during 2004 and is now used by 30 capitols. EMM review is fully in integrated into the other components of EMM. This allows RNS to access any cuttings appended to the reviews in the capitols. It allows the reviews to be accessed directly from the News Brief. It also allows the reviews to be automatically published onto INTRACOMM.

EMM Review publishing is via a login account – one for each capitol. Each capitiol can have two reviewers –one of Printed Press and the other for Audio Visual News. The reviewer logs into the site and enters items to the day's review. Each item can be classified under a controlled list of topics and associated to a controlled (but expandable) list of sources. The review is only visible to the reviewer while it is being edited. Once the reviewer is happy with the content it can be published. The published version is 1) made visible on the site 2) written to XML format for inclusion in the News Brief and INTRACOMM, 3) emailed to the Intracomm publishing system.
The reviews are held in a web interfaced database. This allows users to search fro items across capitols, or to select categories across categories. The result is always in the form of another review which itself can be printed.

EMM Review is described in more detail later.

2.4 EMM Breaking News Service

EMM Breaking News service employs a numerical method for automatically detected breaking news in any language and for determining at any moment which are the top stories in each of the languages. The results are used to update each of the language versions of the NewsBrief. Users can also unsubscribe to email alerts from the Breaking News Service at one of 3 levels. 1) high 2) very high 3) Ultra high. Whenever a news story breaks at the requested level an automatic email or SMS is sent to subscribers.

The EMM Alert system is a real time system for identifying articles on predefined topics. The Breaking News system must identify a sudden increase in articles which mention proper nouns. The system must maintain counts over a two week period of the frequency of capitalised nouns used in each language. Whenever a new noun OR an existing noun at unexpected usage occurs the algorithm must detect it. To avoid false hits the system demands that more than two sources of known quality report the same noun.

2.5 EMM Alert System

EMM news alerts detect and sort articles as they appear in Europe's on-line media. Alerts can be considered rather like "fishing" for articles yet to be published. Each alert definition consists of a list of multilingual keywords (the bait) designed to catch future articles (the fish). When caught, the article is placed into the appropriate alert, which contains up to some maximum number of most recent catches. Alerts are intended to cover a single topic area. They can be permanent or set up especially just to cover a forthcoming event.

Alerts run continuously, scanning and checking all newly published articles against multi-lingual lists of keywords. Each on-line source is checked as frequently as every 15 minutes. The alert scan on a new article is done in a fraction of a second.

2.6 EMM Archive

The EMM 'news archive' provides a dynamic searchable interface to a large repository of stored articles. All articles detected by 'Media Monitoring' are stored for later reference. Using selected keywords you can search as far back as required for articles on a particular topic or topics. Only news sites specified by DG PRESS are monitored providing a tailored news information service. In order to retroactively search for news of interest all articles from the monitored URLs are stored. The text of the stored articles is indexed meaning that it is possible to pick up news items from the past.

2.7 EMM WAP Service

The EMM WAP service allows access to the current content of EMM alerts and Breaking news. It is accessible from mobile phones and works best with modern 2.5G GPRS phones. When combined with WAP Push alert messages

Mobile Phones : Mobile phones have advanced from being 'first generation' analogue phones, to 'second generation' digital phones and with the current development of 3-G broad-band networks will soon have the high-speed, always on data connections. Most users in Europe currently have 'second generation' digital phones meaning that as well as voice calls, they can send and receive data. Although current bandwidths are small, limited data services such as text messaging and WAP can be implemented.

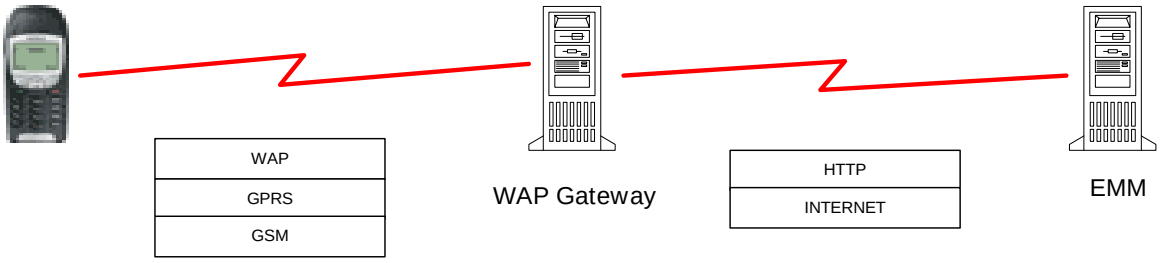

Figure 1 The Mobile Internet

EMM WAP service is accessible through http://press.jrc.it/wab/home

2.8 EMM Regional News Service

The EMM regional news service has been developed for DG RELEX under an agreement in the Rapid Reaction Mechanism. It is based on the existing alert system for worldwide countries, but has been fine tuned to meet RELEX needs regarding groups and issues in those countries. Countries of the world are divided into "regions". The latest news articles are filtered from the RELEX country alerts and presented together on a single page, together with a monthly time plot of articles for each of the countries.

2.9 EMM News Explorer

EMM News Explorer keeps a historical record of the major stories for each day in 6 languages – English, French, Italian, Spanish, Dutch, and German. Every night a clustering analysis is performed on all the previous days news. The clustering is based on identifying keywords in articles and then matching the overlapping keywords. This

groups articles on the same subject into clusters. The software identifies person names, organisation names and place names in the texts of each cluster. Each cluster then represents a main story of the day in each language. The size of the cluster measures the magnitude of the story. Identified place names are then cross-checked against a multilingual gazetteer to geocode the main stories. The clustering analysis uses past work on multi-lingual thesauri to try to identify stories in one language against stories in another language. The same story can also be tracked in time by identifying clusters from the previous day with overlapping keywords.

Finally all person and organisation names are kept in a long term database. Entities mentioned in the same article as other entities become linked. The analysis has been running every day since November 2003 and some 200,000 persons have been identified. A fuzzy matching is made across languages to try to identify the aliases of the same person and checks are made in Wikipaedia to find aliases mentioned there.

The results are stored in each day in XML files for display through the public website. This can be seen at http://press.jrc.it/NewsExplorer. A visualisation of the links between persons is shown in Figure 2.

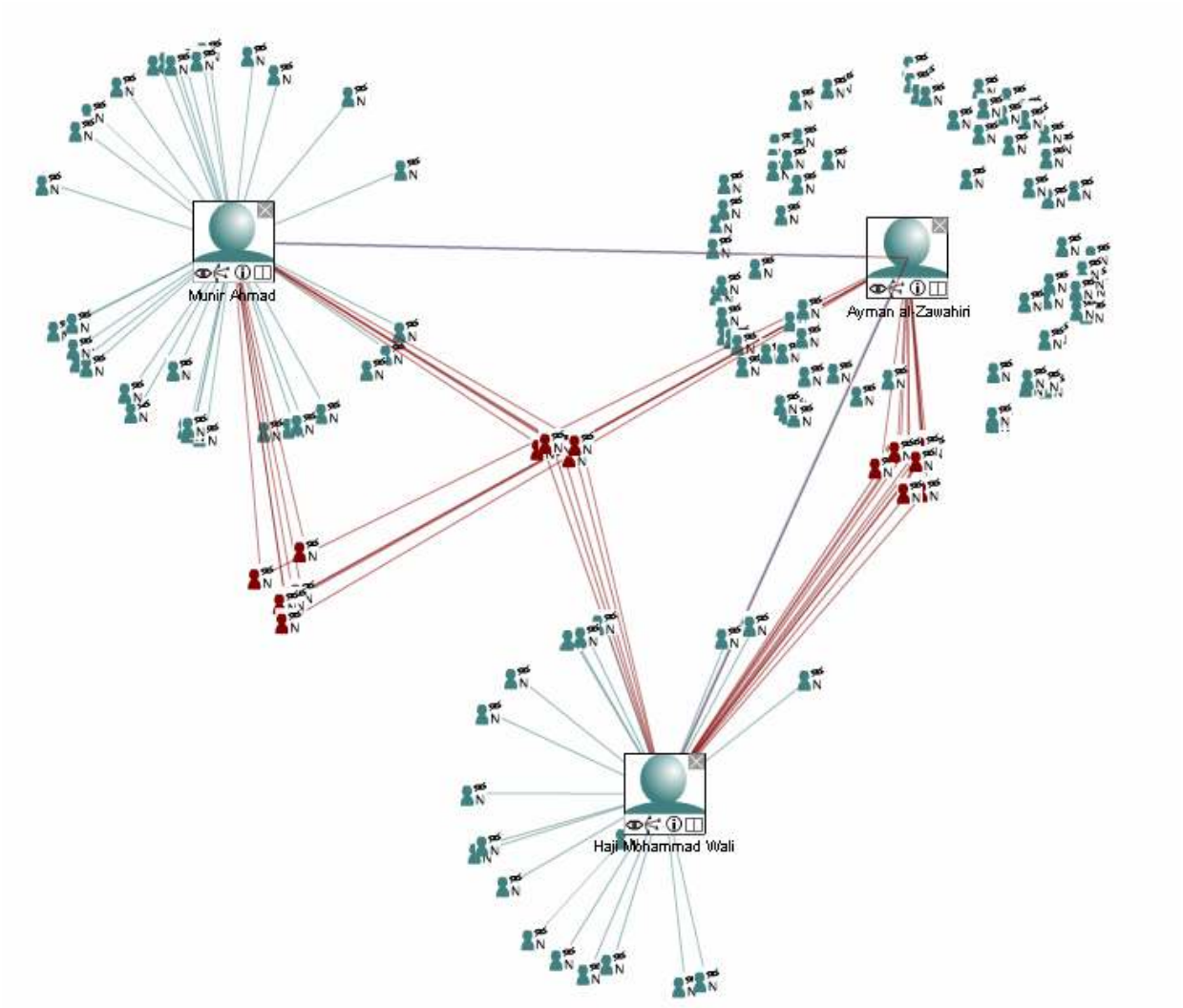

Figure 2: NewsExplorer- Visualisation of persons linked through News Reports

3 EMM Operations

The various components of EMM are running on a set of rack mounted servers hosted in the JRC. These computers are installed in a secure server room at the JRC. Public access is via the Internet access of the JRC at 32 Mbit/sec capacity through UUNet, and Commission access is via a dedicated lines within the Commission firewall. The figure shows the physical machines running the alert, Breaking News and web servers.

Figure 3: Dell Rack mounted servers running the EMM public and internal web servers and alert systems.

The system has proved highly reliable with an up time of over 99%. During 2005 there has been two hardware failures, but these have not caused operations to be closed. It had been intended to duplicate the hardware with a failover system which could pick up operations in the case of a catastrophic failure of the main systems. This would be hosted in another building for security reasons. Unfortunately the Commission framework contract for servers has been blocked all year and it has proved impossible to purchase the hardware. Some emergency spare components have been purchased and it is hoped to complete the original plan for a failsafe system by the end of the year.

Public access to http://press.jrc.it is growing extremely fast. There has been no active marketing of the site, but it is clear from the graphs in Figure 3 that it is proving very

popular. Figure 4 shows the daily accesses to the Commission EMM site. This highlights the week day nature of the service and shows about 20,000 different sessions per day.

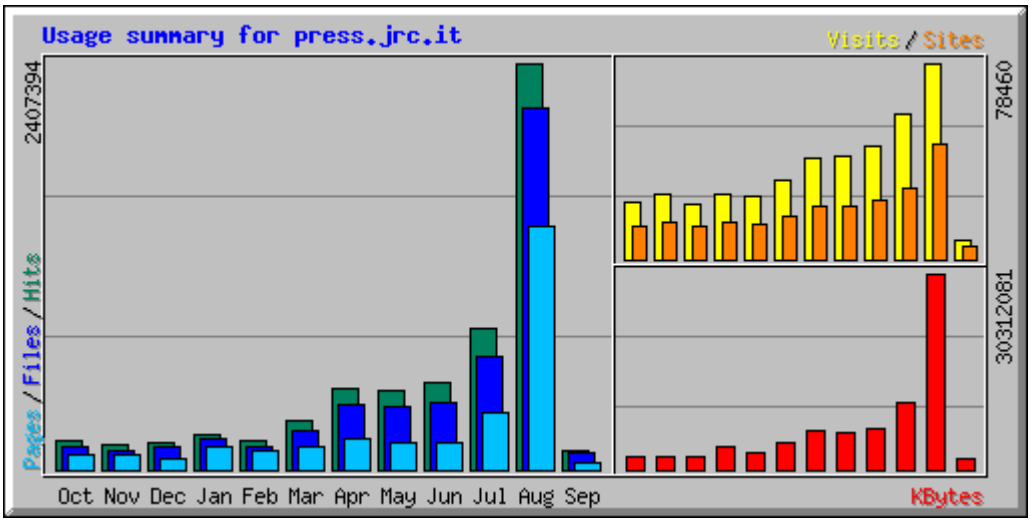

Figure 4: Rapid increase in usage of the public version of EMM. In august 2005 there were about 2.5 Million hits on the site.

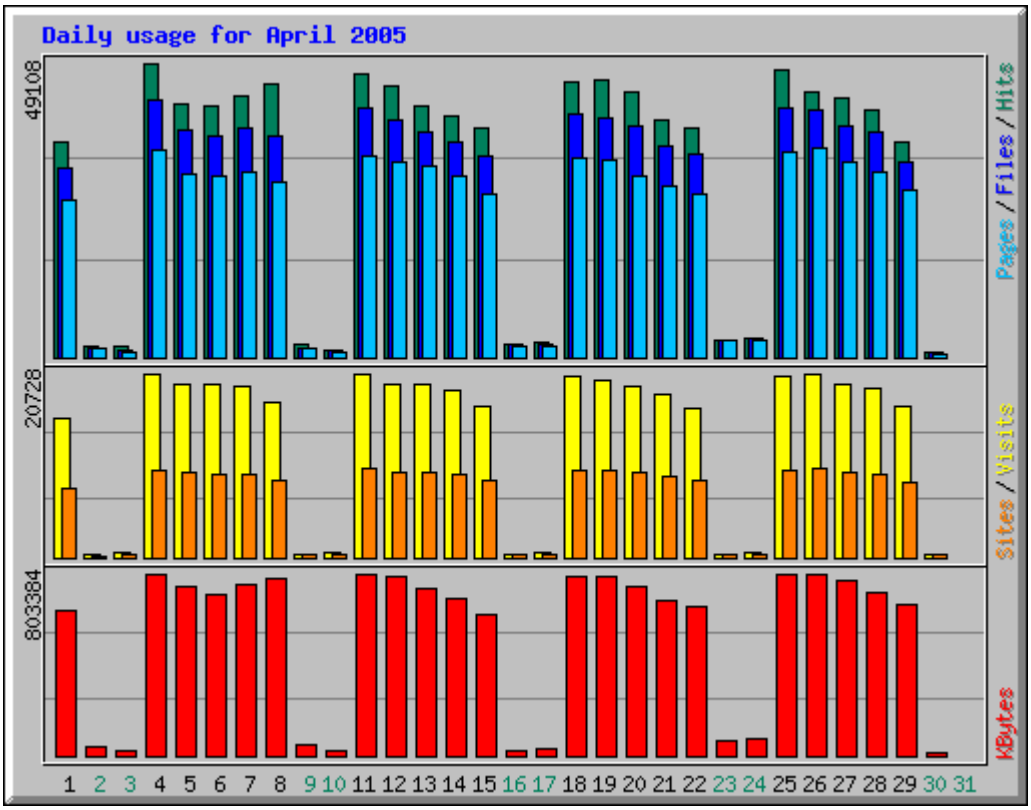

Figure 5: Accesses to the Commission's version of EMM. There is a strong weekly pattern with about 20,000 visits every week day.

4 EMM System Overview

EMM is implemented as a suite of WEB applications. Each application communicates through an HTTP Post mechanism to inform dependent applications of results and to initiate further action. The Web Applications are Java Servlets running on the Open Source Java container and Web server Tomcat. The applications are spread over 10 rack mounted servers running on Windows 2003 Server. The EMM Archive is running on Oracle 10i running on RedHat Linux V7. The diagram gives an overview of all the systems and how they interact.

EMM functions in two security zones –Commission and Public. The first is within the Commission's firewall and processes content restricted only to Commission staff. In practice this is mainly the News Agency wires which arrive round the clock either over dedicated communication lines to the Berlaymont Building in Brussels or via email and secure web site. The Informatics Unit in DG PRESS manage the reception and handling of the major News Agency feeds in Europe. These typically are Reuters, AFP, Associated Press, ANSA and other national agencies. Other News Agencies are handled at Ispra through an Email and sometimes secure web interface. These are mostly agencies from new member states. The Commission web site is http://emm.jrc.cec.eu.int

The second zone gives access to EMM derived content on the Internet. This contains just publicly available free of charge web content. EMM is an aggregation and syndication system. It does not produce content of any kind. It's power is in arranging content from multiple sources in different languages into topics. For copyright reasons the display of this content must obey legally accepted rules. It must always refer to the original source, and cannot display more than about 10% of the article as a short description. The user is always referred onto the original source to read the article in full. This explains the use of RSS for EMM's syndication and display. The publicly accessible web site is http://press.jrc.it This site gives a single overview of the current content. The associated service News Explorer gives a historical overview of top stories, and identifies entities and their relationships. These will be described later.

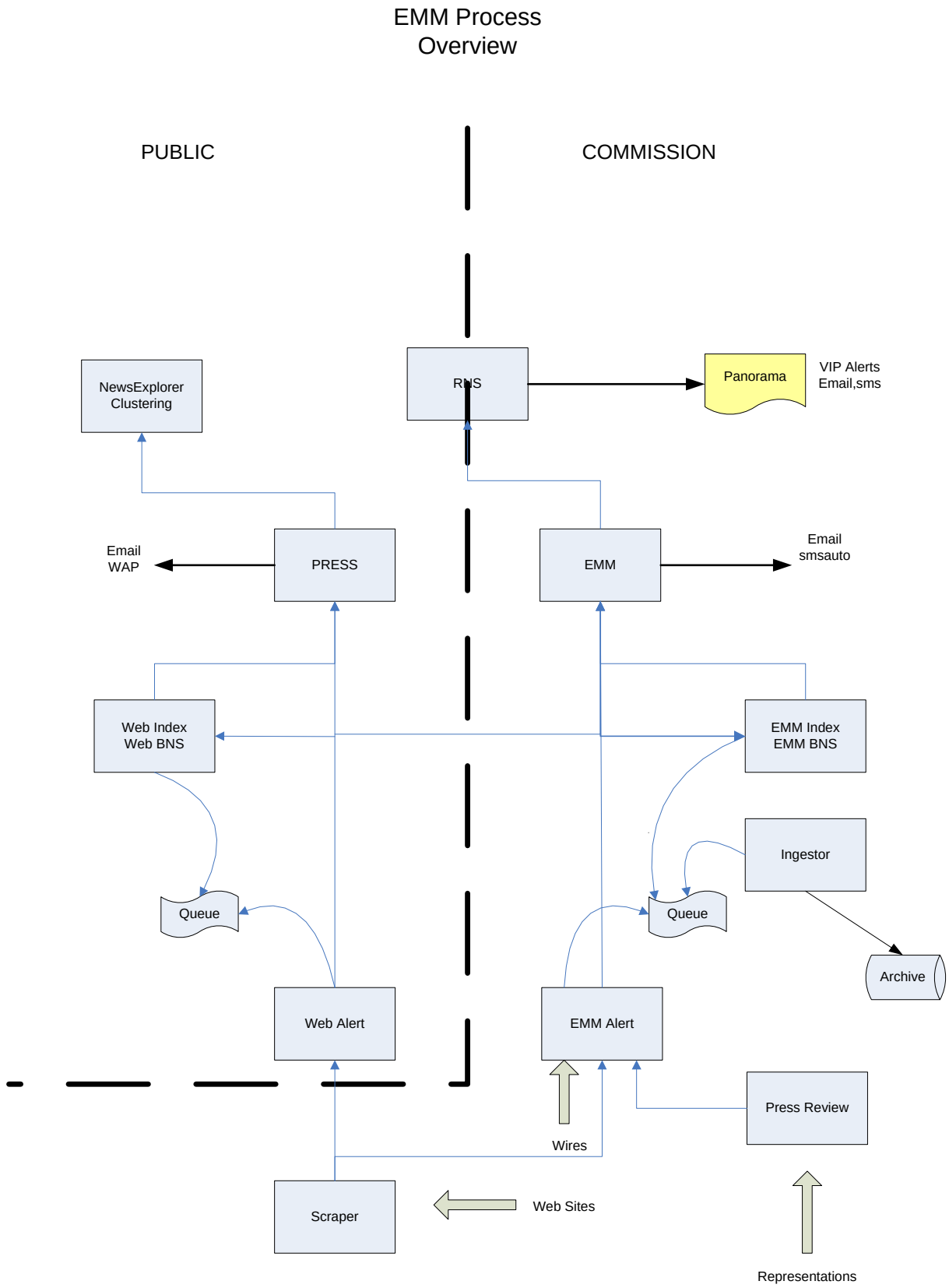

Figure 6: Schematic Overview of EMM Applications.

EMM processes incoming news articles in a modular process. Some 800 web sites are scanned up to every 15 minutes by Web-scraper. This system is driven by two configuration files. One defines the sites to monitor and then for each site a recipe file tells scraper how to analsye the site. The current headline content of each site is cached in RSS files one per site.

The Alert System is called after each scrape and checks whether any new headlines have been detected since the last scrape. If so it follows the URL and extracts the pure text from the web page and filters the content through it's 650 Alert definitions, then passes the content to the Queue for onward processing. The Breaking News System processes the content to identify sudden appearance of new topics, the indexer maintains a free text search index of recent articles. The results are updated continually in RSS and BNS XML files, which update the user Web interface – NewsBrief every 10 minutes. The details of these processes are covered in the next section.

5 EMM Software Overview

The EMM software is built on a JAVA/XSLT framework using Apache/Tomcat and public domain software. A platform independent architecture is described which allows for a scalable expansion. The long-term archive is available only to authorised users. Lastly yet crucially, EMM fully supports multiple languages across all software components. All alerts, search criteria and web interfaces use UTF-8 character encoding and therefore support all world languages.

5.1 Really Simple Syndication

Really Simple Site Summary (RSS) [2] files, based on XML, provide a method of syndicating and aggregating Web content. Originally invented by Netscape [3] for channel content their uptake has concentrated in the generation of headline news services. EMM's internally uses RSS 2.0 as the container for content for most services. All EMM alert files are available as RSS2.0 feeds over the web. EMM core News syndication services are based on converting or "scraping" web pages from news sites into RSS format. Increasingly the source site provides the RSS feed directly, but overwhelmingly it is still the case that the feed is generated by the scraping software (EMM scraper). An example RSS 2.0 file is given below. This also demonstrates how EMM processes character encoding by converting always to UTF-8.

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0" xmlns:html="http://www.w3.org/1999/xhtml">
  <channel>
    <link>http://www.afp.com/arabic/home/</link>
    <title>daralhayat</title>
    <description>General mainstream arabic newspaper</description>
    <item>
      <title>العرب السنة يريدون اسقاط مسودة الدستور في الاستفتاء</title>
      <link>http://www.afp.com/arabic/news/stories/050829110033.0l8trffw.html</link>
      <description> 29/08/2005 11h00 انهى العراقيون الاحد اعداد
        الصيغة النهائية لمسودة الدستور بالرغم من الاعتراضات
        الشديدة لممثلي العرب السنة الذين لم يخفوا رغبتهم بالعمل
```

```
        على اسقاطها في الاستفتاء الشعبي العام المقرر في الخامس
        عشر من تشرين الاول/اكتوبر المقبل.
      </description>
    </item>
    <item>
      <title>الجعفري : لا مكان لحزب البعث في العراق الجديد</title>
      <link>http://www.afp.com/arabic/news/stories/050829112419.2jhfseig.html</link>
      <description> 29/08/2005 11h24 اعلن رئيس الوزراء العراقي
        ابراهيم الجعفري الاثنين انه &amp;quot;لا مكان&amp;quot; لحزب
        البعث في العراق الجديد داعيا البعثيين الى فتح صفحة جديدة
        مع الحكومة العراقية.
      </description>
    </item>
    <item>
      <title>ارتياح لدى بوش بعد اقرار مسودة الدستور العراقي لكن
        المخاوف مستمرة</title>
      <link>http://www.afp.com/arabic/news/stories/050829063150.wyjfonva.html</link>
      <description> 29/08/2005 06h31 رحب الرئيس الاميركي جورج بوش
        الاحد باقرار مسودة الدستور العراقي بصيغتها النهائية
        مقللا من اهمية الانتقادات التي لا تزال لدى العرب السنة
        لكن سفيره في بغداد ادب اكثر تشكيكا حيث حذر من ان رفض هذه
        القيلة لمسودة النص خلال استفتاء 15 تشرين الاول/اكتوبر
        سيخلق &amp;quot;مشكلة&amp;quot;.
      </description>
    </item>
  </channel>
</rss>
```

The various services of EMM may add extra metadata tags to the basic RSS 2.0 format. This is the case for the Alert system which documents how an article (item) satisfied the trigger condition and which other alerts (if any) were triggered.

5.2 EMM Scraper

Scraper is directed to defined pages in Media sites via an Channel Directory (extension of OCS Open Content Syndication) directory file. Scraper itself consists of a 3-step process:

1. Clean the HTML by removing non-standard tags and unnecessary tags, JavaScript etc. This step is performed using a proprietary HTML scanner/parser.
2. Convert HTML to XHTML . This involves closing all opened tags automatically and ensuring a robust DOM representation of the text page. Here the formatting of the page is not important, and table levels and Javascript user interface options are removed
3. Transform XHTML to RSS using a style sheet (XSLT [11]) for each site. The Java JAXP transformation factory is used.

When a new web page to be scraped is added to the Channel directory, a new style sheet has to be written for this page. This part of the process requires human intervention (and ingenuity) and the style sheets in fact represent the distilled intelligence of the human reader.

XML Spy is the tool that is used to develop new stylesheets. The methodology is to base the new style sheet on an existing one and edit the stylesheet to pick out the patterns of the headlines on the XHTML version of the web site. Normally some skill is needed to identify the page (pages) where the transformation will be applied. This is often the entry page to the site, but in some cases it is best to select the one avoiding excessive use of interface enhancements. Some sites are divided into several different sections eg. Economy, World News, Local News etc. In these cases each section can be scraped individually.

Experience has shown that the best policy in defining stylesheets is to identify a pattern in the headline URL. Often a class attribute will allow a pattern match just on those links which point to published articles. An example stylesheet which converts an XHTML page into RSS is given below.

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:html="http://www.w3.org/1999/xhtml"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
    <xsl:output encoding="UTF-8" indent="yes" method="xml" version="1.0"/>
    <!--The main RSS template-->
    <!--Variable declarations-->
    <xsl:variable name="baseURI">http://www.eupolitix.com</xsl:variable>
    <!--The templates-->
    <xsl:template match="/">
        <xsl:apply-templates mode="feed"/>
    </xsl:template>
    <xsl:template match="html:html" mode="feed">
        <rss version="2.0">
            <channel>
                <link>http://www.eupolitix.com/EN/</link>
                <title>EUpolitix</title>
                <description>European Political News</description>
                <xsl:apply-templates mode="item" select="//html:a[starts-
with(@href,'/EN/News/')][text()!=''][contains(@href,'.htm')][not(contains(tex
t(),'more'))]"/>
            </channel>
        </rss>
    </xsl:template>
    <!--All the items are generated HERE-->
    <xsl:template match="html:a" mode="item">
        <xsl:variable name="link">
            <xsl:value-of select="@href"/>
        </xsl:variable>
        <item>
            <title>
                <xsl:value-of select="normalize-space(.)"/>
            </title>
            <link>
                <xsl:value-of select="concat($baseURI, $link)"/>
            </link>
            <description>
                <xsl:value-of select="following::text()[1]"/>
            </description>
        </item>
    </xsl:template>
    <xsl:template match="html:* | html:td"/>
</xsl:stylesheet>
```

In this example all headlines on the HTML page with links containing '/EN/news' are taken except those where there is no text or those where the text is '.more'. This is because other clickable links contain either a Picture or the …more link which we wish to avoid. Also shown is the use of a BaseURI variable to handle the absolute address of the link , rather than the relative address.

Scraper is implemented as a multithreaded Java servlet that automatically schedules the scraping process for the various URLs. The custom written HTML parser used by Scraper reduces the original HTML to an absolute minimum, thereby exposing the underlying site organisation. This facilitates the production of the style sheets and makes the transform process capable of surviving most site layout changes. Scraper is fully multilingual converting all incoming encodings to UTF-8.

This conversion to UTF-8 is essential for EMM to handle multilingual sources often with different character encodings. By converting everything to UTF-8, which is the native character handling of JAVA the software is able to treat Arabic and English in the same way.

Scraper has another important function which is to detect newly published articles to be passed onto the Alert system. It does this by keeping a Cache of all the RSS feeds for each site. After a scrape has been done it can compare the new result with the cache and pass.

The scraper is split into three versions to handle the different sources of news. Web scraper is the main system which monitors most of the public web sites. EMM scraper is reserved for Commission only sources. TVscraper is a special system to separate off some sources from audiovisual suites. The reason for this separation is because these are later used in automatic SMS alerts to spokespersons.

A schematic of the scraper systems is shown below.

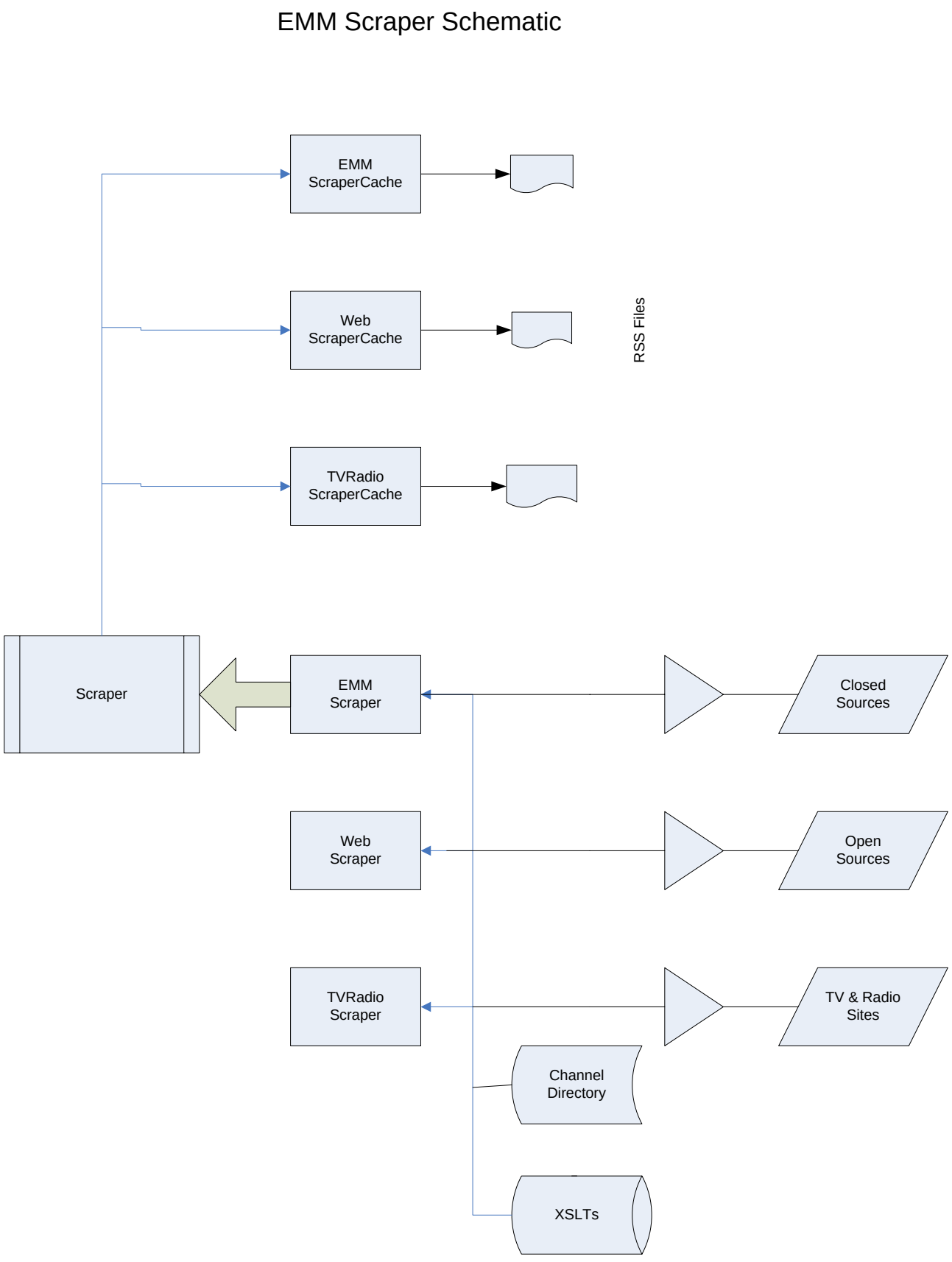

Figure 7: Schematic overview of EMM's scraper applications

5.3 EMM Alert System

EMM's Alert system is it's most unique feature. It drives most of the content and adds value to news monitored. The overall objective is to process as rapidly as possible each discovered article and decide which subjects (Alert definitions) are mentioned. If an alert criteria is satisfied the article is appended to a result RSS file, one for each alert definition. If a user has subscribed to an immediate email alert for that topic then a processor is called to send the article item by email. If a special alert called SMSauto has been satisfied and certain timing criteria are satisfied then an automatic SMS message is sent to a small number of persons.

A number of technical challenges have been overcome in implementing the alert system. Firstly, the real textual content form web pages needs to be extracted from raw HTML. It is no good triggering alert son adverts or on sidebar menu items which have nothing to do with the content of the web page. Secondly the alert system must be extremely fast to keep up with incoming articles and to alert interested person s as required.

A simplistic approach to determine if an article should trigger an alert would involve simply looking for the occurrence of words from a predetermined list of words in the text of the article. Every time the word is found, the weight associated with the word is then added to a total score. If after the list of words has been processed the total score exceeds some preset threshold, the article triggers the alert. This is a straightforward approach that could be implemented using any text search engine.

The main drawback of this method is the fact that the article text has to be scanned for every word in the list. When looking for one or two words this is not a problem, but when the number of words grows this would lead to unacceptable processing times.

To overcome this problem, an algorithm has been developed whereby the whole of the article text is scanned only once. The algorithm matches all possible patterns in parallel as the text is being scanned character by character. The algorithm is based on a parallel state machine (i.e. a number of state machines executing in parallel) where the machines to be loaded into the processor are loaded from a hashtable, indexed by the first character of the word in the text to be scanned.

The algorithm also loads all state machines relevant to the patterns starting with the two wildcard characters '%' and '_' and keeps track of so called 'multi word patterns'.

The algorithm keeps track of the total score of the article on a 'per alert' basis. This means that a pattern detected in the text can contribute a different weight to different alert. Furthermore the exact article text and the number of occurrences is tracked for each alert.

There are two alert systems as to separate the content form the Commission service and the public service, each services the two types of scraper. The alert system receives an RSS item from each scraper via an HTTP post. The web app retrieves the HTMl from the provided URL and processes it to extract the textual content. This is a non-trivial step which involves identifying patterns of textual content within the web

page and uses a heuristic approach. The text is then filtered in parallel against about 10,000 keywords arranged across 600 alert definitions.

5.3.1 Software Implementation

The alert system is a unique service run independently of a database and is a pure JAVA/XML application, based on Scraper and AlertMonitor. AlertMonitor uses the same multithreaded design and job-scheduling approach as Ingestor. It also uses the same custom HTML parser. The alert monitoring is based on a new on-the-fly text-processing algorithm (see 6.6). It includes a new e-mail processing module. The Alert system will detect articles on given subjects within minutes of publication. By sending immediate e-mail and keeping web summaries the Alert system can notify interested users of these new articles immediately. The presentation is through JAVA/XSLT transformation of the XML encoded alert information. Apart from the alert information also the alert definitions and email subscriptions are kept in XML formats. An Alert definition consists of any number of keywords and associated weights. An alert triggers if the sum of matched keyword weights in the article text exceeds a threshold. Negative weights can be used to suppress unwanted contributions. The information generated by the system includes the total score, and the exact text that matched the patterns defined for this alert and the number of occurrences of this text.

The text-processing algorithm uses a state machine to process in memory keyword lists against incoming text. It currently processes a single article of a few thousand words against 1250 pre-defined keyword patterns in 20 milliseconds (see 6.6).
By avoiding the use of a database and by exploiting the multi-threaded advantage of Java, the alert system is lightening fast and reliable. The results are held in RSS files that are continuously updated. This allows fast access by many users over the Web.
The NewsBrief accesses these RSS files to display the results of all Alerts through a single interface. Another major advantage is the ability to feed third party web sites with content specific news. This is done by transforming the XML feeds into Java Script content embedded in external web pages.

The alert system also feeds other subsystems via EMM Queue. EMM Queue holds caches of postings from the alert system for distribution among the next stages in the processing In particular the Ingestor system receives the raw text content for ingestion into the EMM archive. The EMM archive is an Oracle database which holds articles kept for research purposes only. No access is permitted.

5.3.2 Alert Definitions

There are two ways to define an alert. Both methods can be used for the same alert. Some concepts are best expressed using the keyword and weight method. This method can guarantee a 'perfect match'. Other concepts are better expressed using the combination of keyword lists.

5.3.2.1Keywords and weights

The first and easiest one, is to use a list of keywords, with associated weights. To receive as wide a coverage as possible, keywords in different languages are recommended. Keywords are expressed by patterns which can be simply the keyword

itself, or a keyword description using 'wildcard' characters. The effectiveness of the alert in this case depends on the choice of words. Few truly characteristic terms give better results then many more general words, even if these words frequently appear in the text of relevant articles. A simple example of an alert definition using keywords and weights is given below. The values of the maximum number of articles, the threshold for each alert and the weights of the patterns can be chosen by you, but they have to be integer values.

```
alert=EuropeanParliament
maxArticles=50

words, threshold=20
european+parliament                    20
parl_ment%+euro%                       20
euro%+parlament%                       20
europa+parlamentet                     25
europaparlamentet                      25
```

This example contains keyword patterns in several languages. The system will accept terms in any language, and an alert may have any number of keywords.

The system keeps track of the total weight by summing the weight of the individual patterns, and when the weight exceeds the threshold the system considers the article to be part of the alert. Weights can be negative and this can be used successfully to exclude certain articles.

Patterns used for the alert system can be specified using two wildcard characters, _ (underscore) and % (percent).

The _ character denotes one single character, but not a blank
e.g. p_t would match pot, put, pat, prt etc. (to be precise, the _ character matches exactly one of: a-z, 0-9, "'", "_" or "-")

The % character denotes zero or more characters
so comm% would match comm, commission, commisioner, common etc.

Using these two characters it can be very easy to build common patterns for multiple languages because they can substitute accented characters.

The + sign can be used to build multi term strings,
e.g. romano+prodi will match romano(whitespace)prodi

A pattern definition should normally be in lowercase, but can contain uppercase characters. In that case the pattern will only match text that has an uppercase character in the same position,
e.g. Euro would match EURO, EUro, EuRo etc. but NOT euro or eURO because the first 'E' in the text has to be uppercase.

Forcing uppercase can be used for acronyms that would otherwise cause problems.

5.3.2.2 Using combinations

The second way is to define one or more combinations of lists of words. A combination consists of one or more 'or' lists of patterns, and zero or one 'not' list of patterns. The combination scores as valid for the alert if and only if a pattern is found from each and every 'or' list and none of the 'not' list patterns occurs in the text.
In effect, each list expresses a certain concept and an article is considered for the alert if every concept is found in the article but rejected if the 'not' concept is found.
A simple example of an alert definition using lists is given below :

```
alert=IrishReferendum
maxArticles=50
        combination
                or=
                        ireland
                        irish
                        iers
                        ierland
                        irland%

                or=
                        referendum
                        volksabstimmung
```

An alert can be defined using more than one combination. A combination consists of one or more 'or' lists of patterns, and zero or one 'not' list of patterns.

5.3.3 XML Format

EMM alerts are defined in XML files. One XML file groups the alert for a given DG or for a given set of applications. The XML format is illustrated in the following example

```
<alert id="Iraq-EU">
  <maxArticles>50</maxArticles>
  <contact>Clive.Best@jrc.it</contact>
       <definition>
              <mustContain>
                     <combination>
                            <or>
                                   <pattern>Iraq</pattern>
                                   <pattern>Irak</pattern>
                                   <pattern>Ιράκ</pattern>
                                   <pattern>Iraque</pattern>
                            </or>
                            <or>
                                   <pattern>european+commission</pattern>
                                   <pattern>commission+européenne</pattern>
```

```
                    <pattern>europäische+kommission</pattern>
                    <pattern>europa-kommissionen</pattern>
                    <pattern>ευρωπαϊκή+επιτροπή</pattern>
                    <pattern>europese+commissie</pattern>
                    <pattern>comissão+europeia</pattern>
                    <pattern>euroopan+komissio</pattern>
                    <pattern>comisión+europea</pattern>
                    <pattern>eu-kommissionen</pattern>
                     <pattern>commissione+europea</pattern>
                    <pattern>europeiska+kommissionen</pattern>
                    <pattern>Evropská+komise</pattern>
                    <pattern>Komisja+Europejska</pattern>
                    <pattern>Euroopa+Komisjon</pattern>
                     <pattern>Eiropas+Komisija</pattern>
                    <pattern>Europos+Komisija</pattern>
                    <pattern>Európai+Bizottság</pattern>
                    <pattern>Kummissjoni+Ewropea</pattern>
                    <pattern>Európska+komisia</pattern>
                    <pattern>Evropska+komisija</pattern>
                    <pattern>EU</pattern>
                    <pattern>UE</pattern>
                </or>
            </combination>
        </mustContain>
    </definition>
</alert>
```

This alert will identify news articles which refer to Iraq and to the European Union within the same text and will work in all 15 main languages of the EU. The second example shows the format for a weighted alert definition. Here there is just one pattern. This can be increased to contain multiple patterns each with a different weight, including negative weights. The alert is triggered when the sum of the weights of all words found is greater than the threshold value.

```
<alert id="MichaelMann">
    <maxArticles>50</maxArticles>
    <description>Michael Mann</description>
    <definition>
        <words threshold="50">
            <word>
                <pattern>michael+mann</pattern>
                <weight>50</weight>
            </word>
        </words>
    </definition>
</alert>
```

An Overview of the Alert System is shown in figure 8.

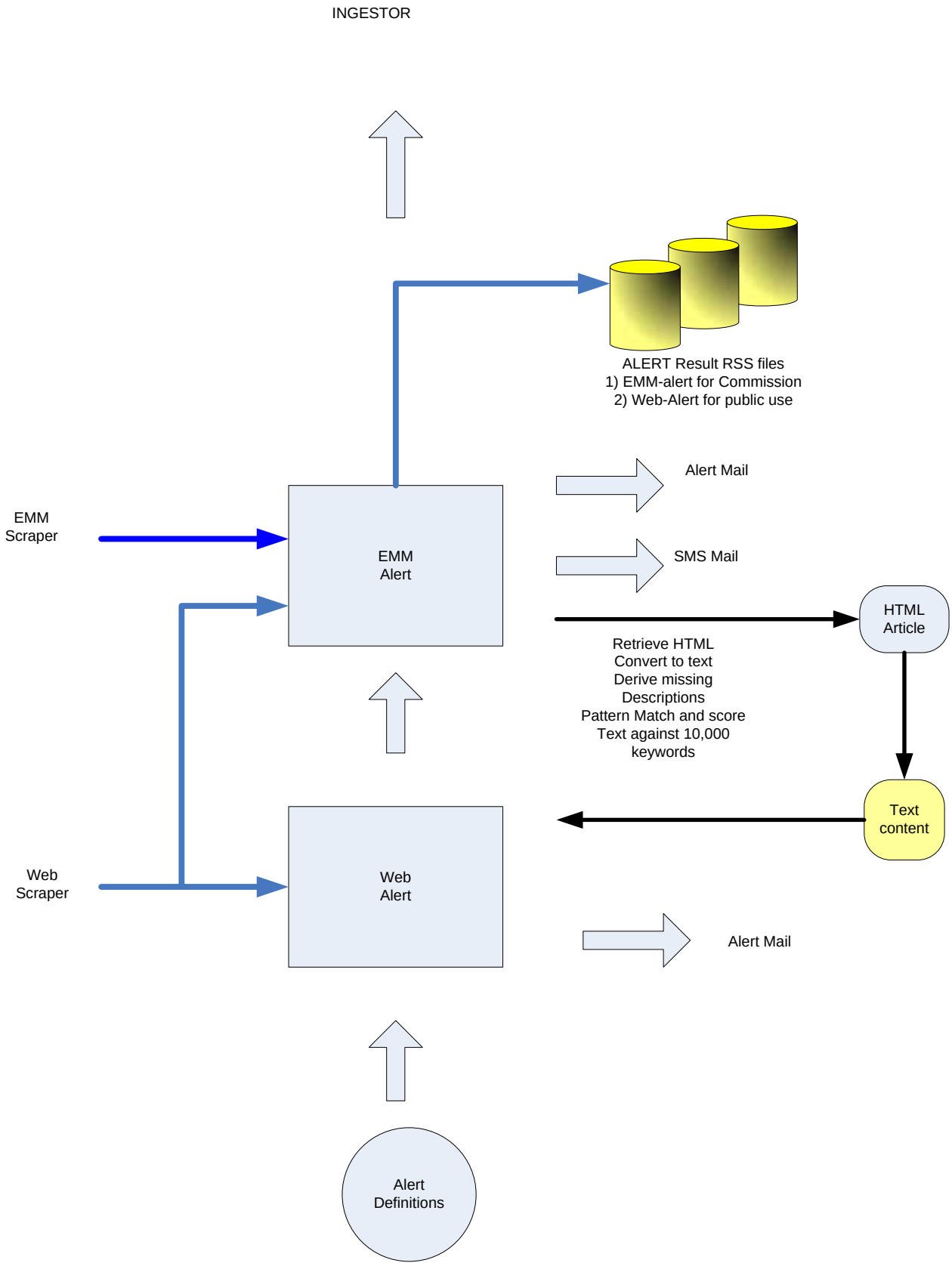

Figure 8: Schematic Overview of EMM Alert Application

5.3.4 EMM Alert Statistics

The alert system keeps hourly statistics of the number of articles detected for each Alert in the system. This is stored in XML files accessible on the web server. One file is stored every day containing the hourly values and taken together form long term time series of event statistics. As major events occur – so their development is recorded in the statistics files.

There is a special alert type called a “theme”. These are stored in the alert definition file themes.xml. These are intended to be general alert definitions which apply globally and typically relate to “Conflict” “Food Security” and the like. Another set of alerts have been defined which cover articles mentioning any of 220 different countries. There is one alert per country. The statistics system records also the combination statistics of the overlap between “themes” and “countries”. If a single article triggers a given country alert eg. Irak and a given theme alert eg Conflict, then the combination statistics are incremented by one.

This combination measures the relative reporting of a theme and a country, and has proved useful for trend analyses. An example of these statistics are illustrated below.

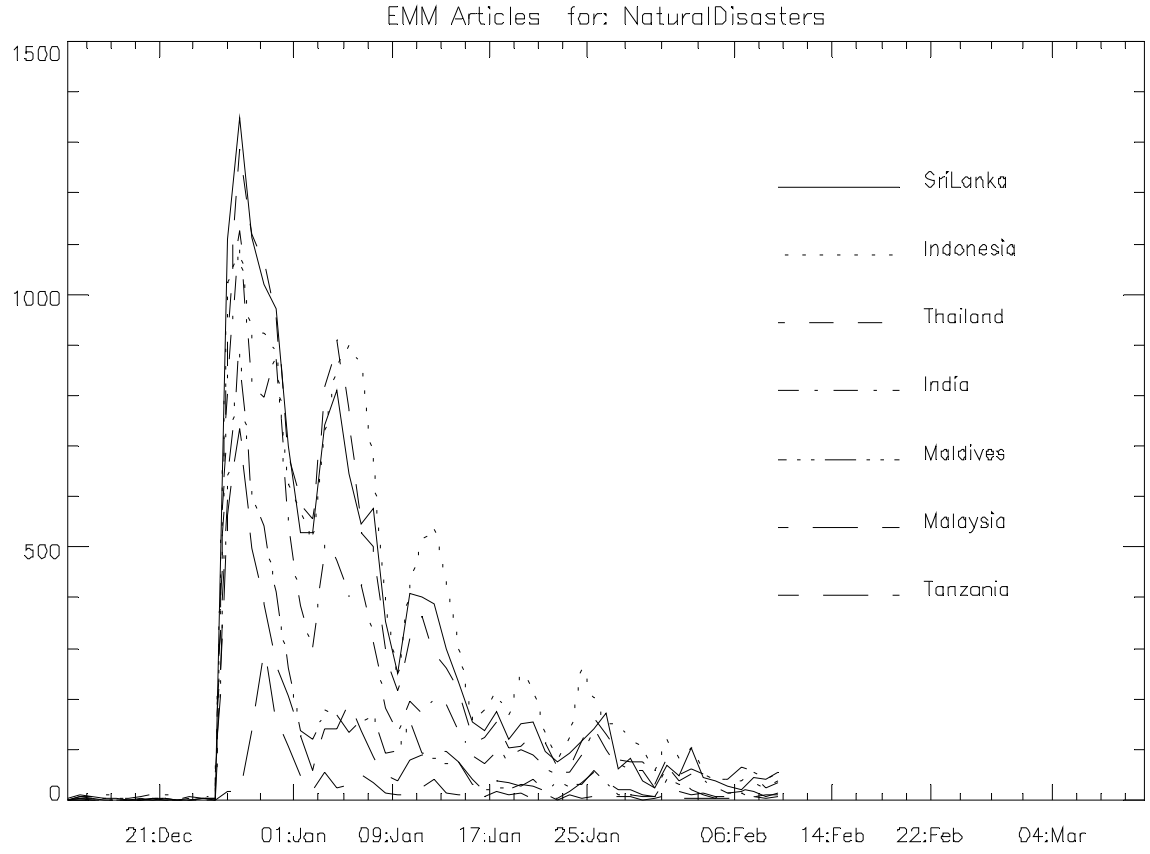

Figure 9: Natural Disasters statistics for the Tsunami Disaster.

5.4 EMM Queue

EMM Queue acts as a buffer between the Alert system and the other processes in the chain. In this way the Alert system is protected against any delays or faults in the following processes. At peak times the database is a bottle kneck and this queue

ensures that the alert system can proceed at full capacity. There are two parallel queues. The first EMM Queue holds the articles destined for the Commission internal system while Web Queue holds the articles destined for the public system press.jrc.it.

A specialised Filter on the results of an Alert has been introduced. This allows to define extra conditions on the content of a given alert. For example some clients insist only on articles from a defined set of sources, or a set of languages. The Filter definitions have been derived from the HTTPget Alert API filterAlertXML. Basically constraints on any of the XML tags produced by the alert system (RSS + extensions) can be defined. An example of an AlertFilter is given below. Here we filter articles from the Bangladesh alert only from the sources defined AND from which articles didn't trigger the Sport alert (i.e. Remove sports stories).

```
<profile name="Bangladesh-Spot">
  <maxItems>100</maxItems>
  <combination>
  <or>
  <category>Bangladesh</category>
</or>
<not>
  <category>sports</category>
        </not>
        </combination>
  <source>bbc</source>
  <source>skynews</source>
  <source>LeMonde</source>
  <source>independent</source>
  <source>ft</source>
  <source>telegraph</source>
  <source>itv</source>
</profile>
```

A schematic overview of the EMM Queue system is given in the figure

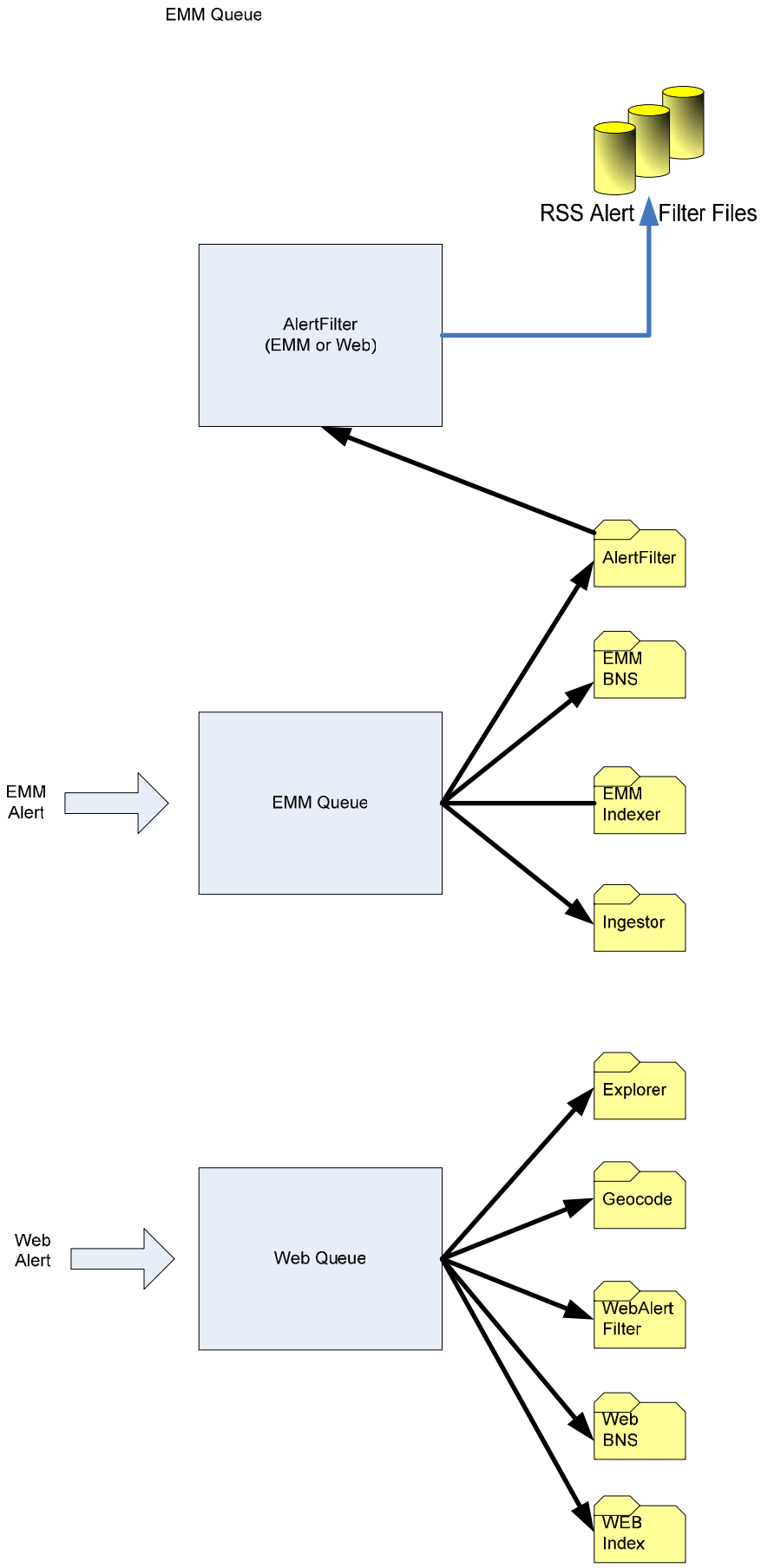

Figure 10: A schematic overview of the Queuing Systems

5.5 EMM Ingestor

Ingestor archives EMM articles to the EMM Archive Oracle database. Ingestor receives input from the EMM Alert system via the Queue, and keeps a cached queue of incoming posts. These contain the extracted texts + html and flag which alerts trigger. Ingestor is a Java Servlet. It scans the RSS file and queries the archiver to check whether any of the headlines is new (i.e. not already indexed in the archive). For articles with an identified image it visits the article URL and retrieves in-line images and stores them in the archive, updates the <IMG> links accordingly and stores the HTML in the archive. It extracts just the text from the HTML and writes that to the archive for indexing. A simple heuristic provides a description of the article, which is used for presentation purposes.

Ingestor is a multithreaded Java servlet that implements a proper job queue. The job scheduling is done on a round-robin basis but the scheduler will automatically restart a thread when a job is hung. Ingestor receives the textual contents of the web pages from the Alert system via the EMM Queue. Ingestor downloads any identified images. The schematic overview of the Ingestor is given in the figures

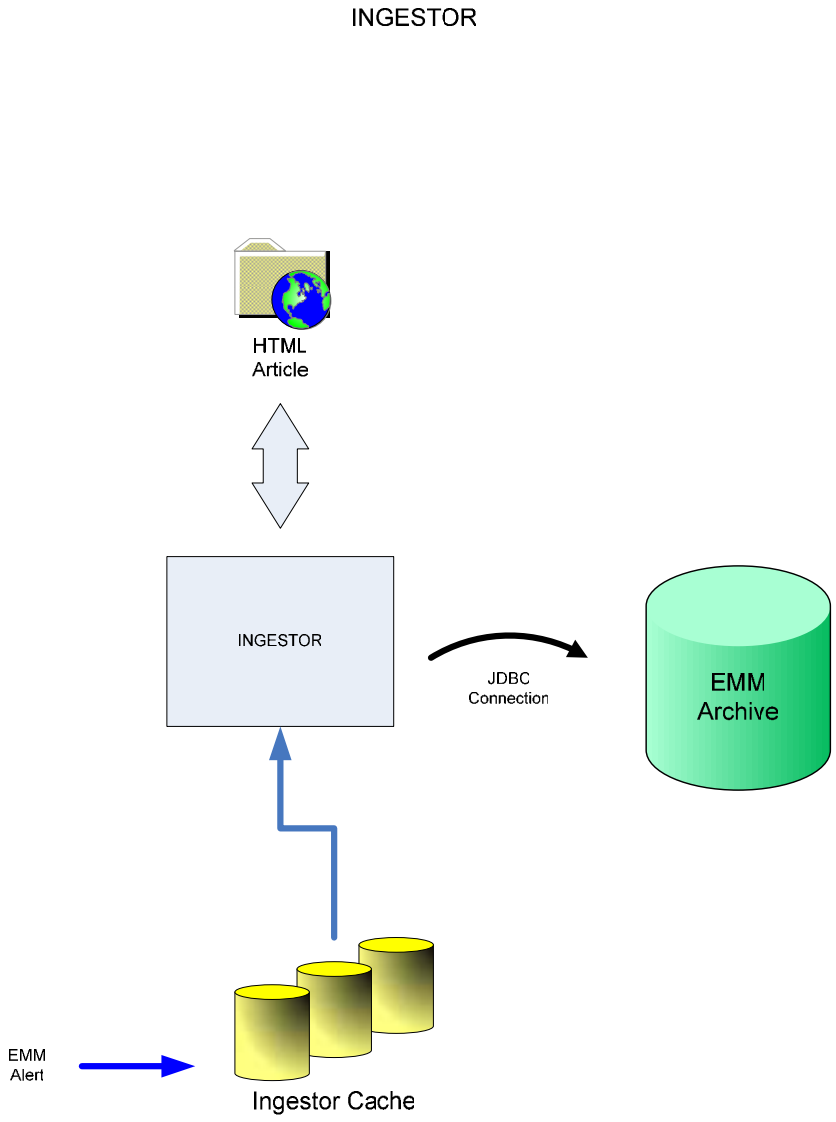

Figure 11: Schematic overview of EMM Ingestor

5.6 EMM Archive Databases

EMM Archive is database implemented on an Oracle 9i installation with a JRC designed database structure. A simplified database schema for the Article database is shown if Figure 12. The main table is the Article table which is linked via a GUID (unique identifier) to the content (text extracted from HTML), Binary (HTML, PDF etc) any images associated with the article, and the RSS elements for the article. Originally the GUID was an article index incremented by one for each new article, but now it is formed by an MD5 hash of the content text.

The system must handle currently the insertion of 25000 articles (including images) per day from ingestor. In early days there were problems caused by the simultaneous indexing and retrieval of the database. For this reason the database was divided into a short term archive of just the last weeks content and a long term archive of all content except for today.

Ingestor keeps a pool of JDBC connections and inserts articles into the STA database. The simplified database design of both LTA and STA is shown in Figure 12. Oracle indexes articles in the STA at regular intervals to eventually allow researchers to search for and extract the articles. The archive is a closed system accessible only to authorised persons. There is no access except for research purposes. The STA acts as a buffer and contains the last week of data. Every evening the days content is written to the LTA and a free text index run. This ensures a rapid response time for searches executed on the LTA during normal hours. The LTA currently contains over 10 million articles dating back to May 2002.

The database can be accessed via a controlled web interface. This web application is called EMM-archive. This is implemented as JSP pages and a Java servlet running on Tomcat. The access to the database is through a JDBC connection. This user interface to the archive is described later.

Two other databases are run to support services of EMM. The State of the World server provides world news reports from every country in the world. Long term trends for countries are held in the SOTW database. The EMM clustering analysis is supported by a database containing a gazetteer of place names and an entity store of extracted persons and organisations. This database is used off line for the data processing for the News Explorer and is described later. A Schematic overview of the databases in use is shown in Figure 13.

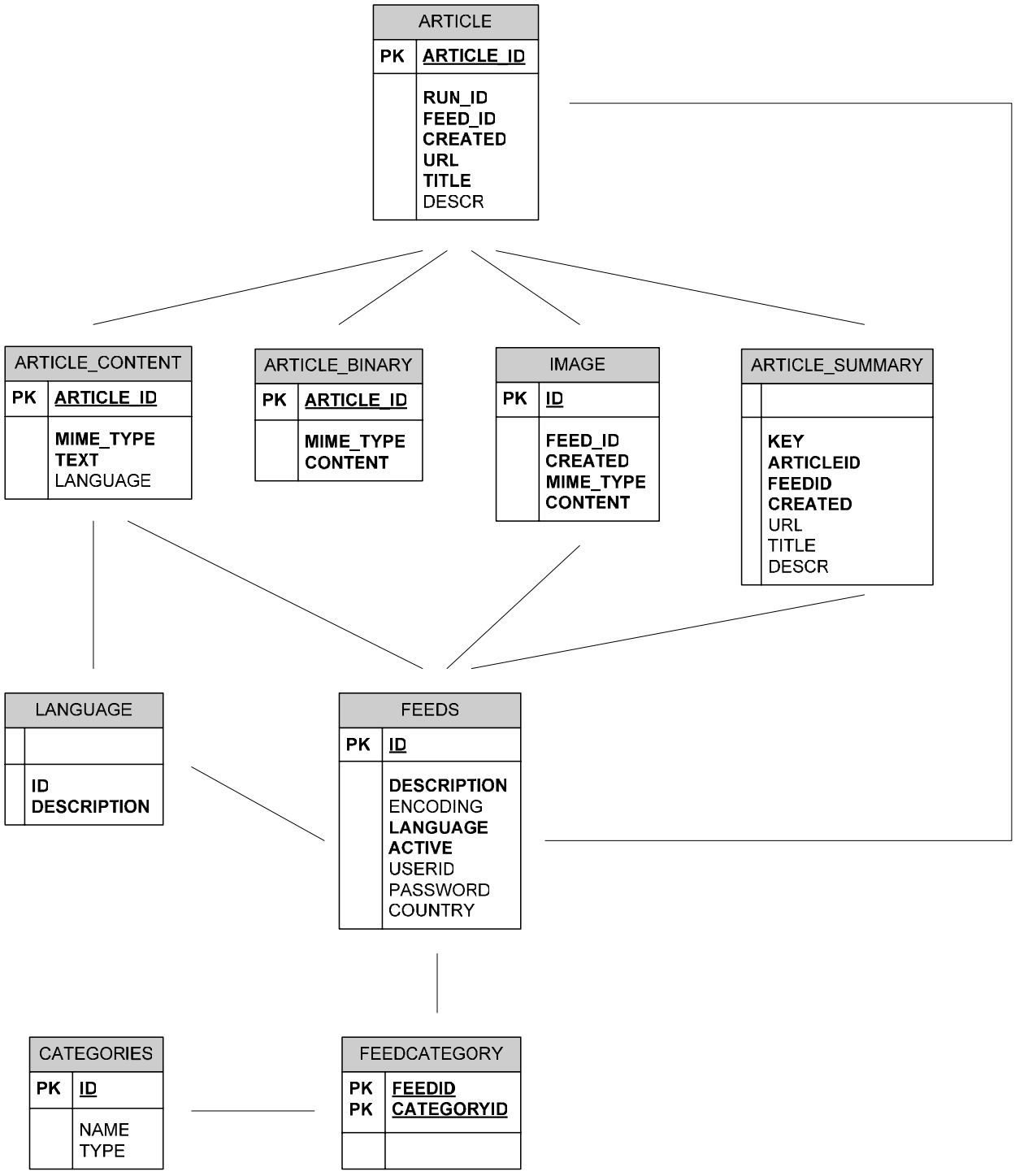

Figure 12: Simplified Database Structure EMM Archive

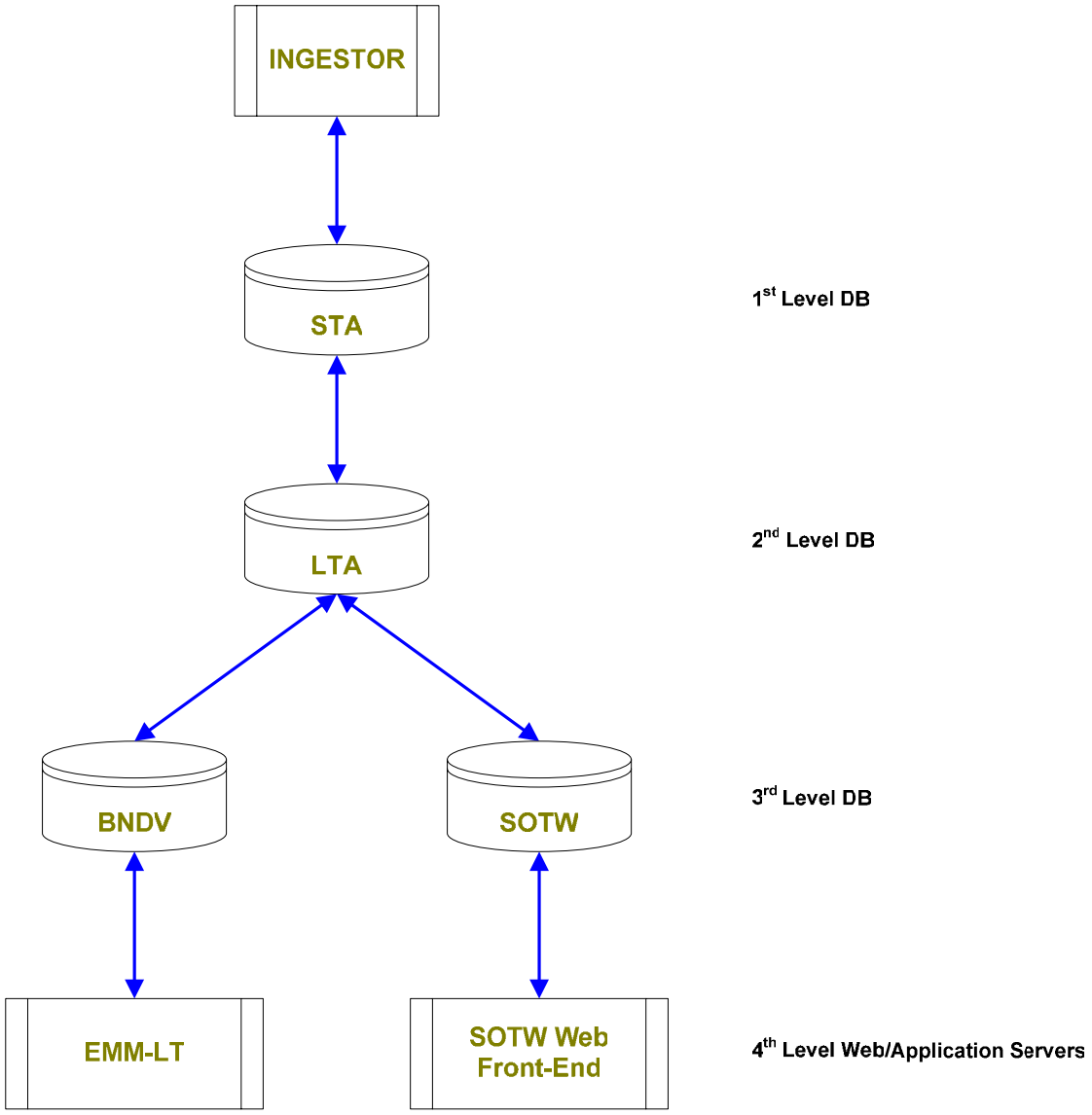

Figure 13: Schematic of databases used by EMM and associated systems

5.7 EMM Archive Search Interface

The 'news archive' provides a dynamic searchable interface to a large repository of stored articles. All articles checked by 'Media Monitoring' are stored for later reference. Using selected keywords you can search as far back as required for articles on a particular topic or topics. Only news sites specified by our customers are monitored providing a tailored news information service. In order to retroactively search for news

of interest all articles from the monitored URLs are stored. The text of the stored articles is indexed meaning that it is possible to pick up news items that would be missed by other search engines.

The interface allows the tailoring of searches to include certain words while excluding others, as well as matching character patterns (using wildcards). A filter can also be applied to restrict the articles searched to only those which are relevant to the 'European Union'.

Unlike most news on the Internet the news on 'Media Monitor' is not 'transient'. The articles stored and are always accessible. This avoids the problem of 'broken links' and also means that you can research the historical reporting of a topic. Even though it is an archive it is updated every ten minutes so a list of retrieved articles will be very much up-to-date.

Figure 5: Access to long term archive

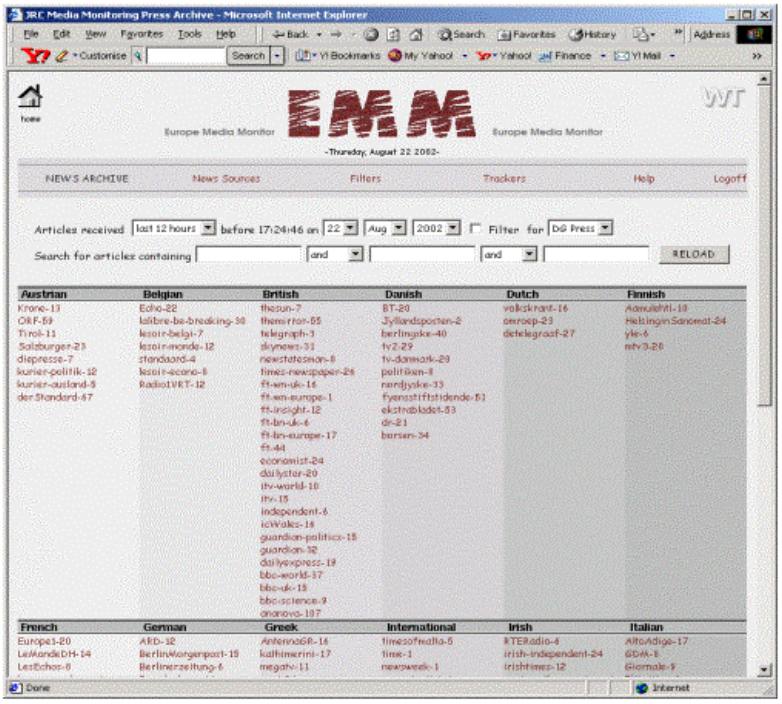

Figure 14: Long term archive

Figure 14 shows the default page of the archive listing the numbers of articles from individual sources ingested over the last 12 hours. This time period can be changed and the baseline time (now) changed to another date. The search interface then allows boolean combinations of words to be searched for by selecting Reload. The result is then the set of found sources with numbers of articles. These are listed as shown in Figure 613, by clicking on the source name.

Figure 6: Search results for "Prodi" from one source

Figure 15: Search results

5.7.1 News Trackers

News trackers follow a given storyline over an extended period of time. They keep an automatic record of articles referring to a particular subject. They can also be tracked backwards in time, meaning that if you wish to analyse a subject after the event the tracker can generate it's history from analysing past articles. Once created new articles are tracked automatically. Figure 7 shows a News Tracker listing chronological orders of articles by day. A tracker can be configured to handle either a single language or all languages. Likewise the length of time in days to track backwards can be configured.

Figure 7: News tracker for "Romano Prodi"

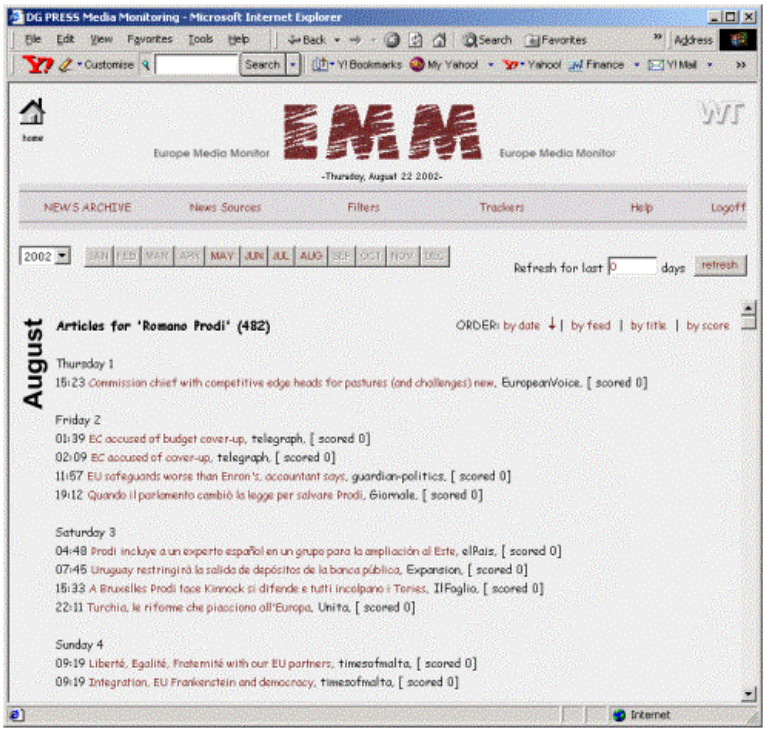

Figure 16: News Tracker for Romano Prodi

5.8 EMM Indexer

All articles are indexed using the open source Lucene Java Indexer. This is to allow a free text search and an advanced interface for users over the Web. There are two separate indexers running one for the Commission users which includes the newswires and subscription sites and one for the public interface which indexes just the web based articles.

EMM indexer picks up incoming articles from the queue and indexes them once every 10 minutes. The searchable index is then updated with the new articles. The index is run on the text content itself, and the feed, the language, and the country of origin. This allows the advanced search to select each of these index fields and then perform a free text search.

5.9 EMM Breaking News

The objective of the EMM Breaking News system is to detect unexpected stories not covered by the alert system. The goal is to identify the big stories in each language and to alert users when large breaking news stories occurs at any time of the day. The Breaking News detection system also drives the EMM NewsBrief content by providing the live top stories of the moment.

Before implementing the system into the EMM system a small case study was undertaken to assess the feasibility of the objectives. This was done by selecting a past event and using historical data stored by 'EMM' in an archive (the articles processed by EMM are stored for analytical purposes) to see if an event would have been automatically detected. The EMM system stores all the data it receives and therefore it was possible to recreate the data stream that would have been received during a breaking news story. The main objectives of this study were:-

- To show that it is possible to automatically identify a ' breaking news' story.
- The breaking news event should be a topic that is not commonly in the news. and so would not have been picked up by simple 'keyword' searches,
- The system should be able to identify the breaking news story in a timely fashion.

The data selected for the analysis covered the period from 30th September 2002 to 20th October 2002 inclusive, the period of the 'Bali bomb blast'. Approximately 232,564 articles were published over this period and stored on the EMM system.
To identify news events a simple java program was written that parsed the title and description of each article and checked for capitalised words. Capitalised words were selected because in most cases the main subject of the article is capitalised, this also means that the process tends to pick up the main actors and locations of the new story whilst ignoring non-capitalised words that are usually of lower significance. In general, the process of identifying capitalised words works for most languages, although not all. For the German language all nouns are capitalised which tends to produce a lot of data, however, by using an extensive list of stopwords (see 5.9.4.1) this problem can be overcome.

One of the first problems we had to tackle was how do we define what breaking news is? There are a number of problems involved in identifying a 'breaking new story'. A simple method is to analyse how a human would identify a breaking news story, for instance, while looking at a news site on the Internet. The methods used by a human are listed below: -

1. *Highlighting and Capitalisation*: text that is in larger print: on the first scan of a news site (or newspaper) the user will initially identify text in larger print as well a capitalised words. Based upon these (s)he will make a judgment on what the article is about.
2. *Unusual topics*: topics that are not normally in the news can be identified as breaking news. If a user hasn't read about the topic in the last couple of weeks then it is likely that this is a new news topic instead of a article relating to a earlier news event.
3. *Coverage*: if there were a lot of articles on the topic this would also constitute a breaking news story, the more widespread the coverage the more important the story.

5.9.1 Daily Variations

Analysis of the dataset has shown a number of periodic fluctuations in the publication of news stories. The first and most obvious is a daily variation. If we look at the number of articles published over time we can see that in general there is a peak around 7 p.m. (CET) and a dip around 5 a.m.(CET) see Figure . Not all of our news sources lie in the Central European Time zone, but the majority do, indicating that many journalists publish specifically for the news at the end of the day. Other peaks have also been noticed at around 10 a.m. and just after lunch. This fluctuation is illustrated in Figure , where the number of articles per hour has been plotted over the period 29/09/2002 to 20/10/2002. If we take a moving average using the previous two hours a daily cyclical trend can be seen. The formula used to calculate the moving average is defined in Equation 1: Moving Average.

Equation 1: Moving Average

$$A = \frac{\sum_{t=-T}^{0} n}{T}$$,where

- *T* is the number of prior periods to include in the moving average
- *n* = number of occurrences over the last hour
- *t* is the time now
- A is the average number of occurrences per hour

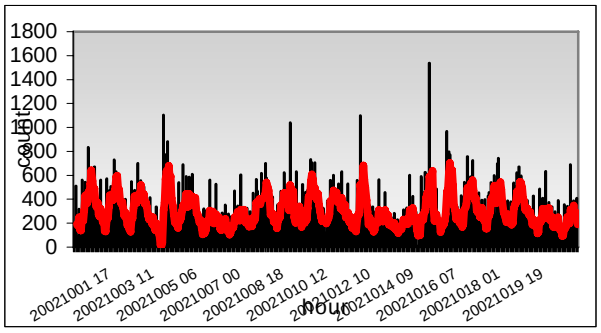

Figure 17 : Articles per hour with 2 hour moving average

5.9.2 Weekly Variations

If we change the moving average sample period to 48 hours another cycle reveals itself. We can see that there is a weekly variation on the number of articles published. In general we tend to find that during each day of the week there are roughly the same numbers of articles published but at the weekend the number drops off, with more articles being produced on a Saturday than on a Sunday.

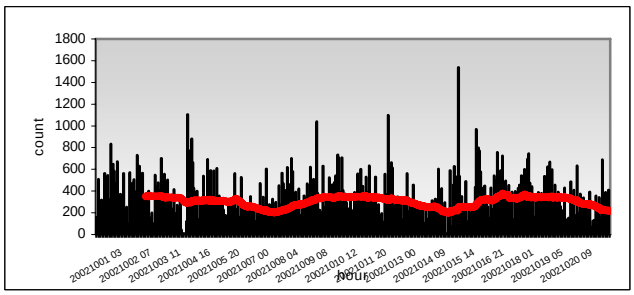

Figure 18:Articles per hour with 48 hour moving average
Figure 19: Weekly Averages (19/08/2002 to 30/09/2002) shows in more detail the publication frequency of articles over a week (starting from Monday).

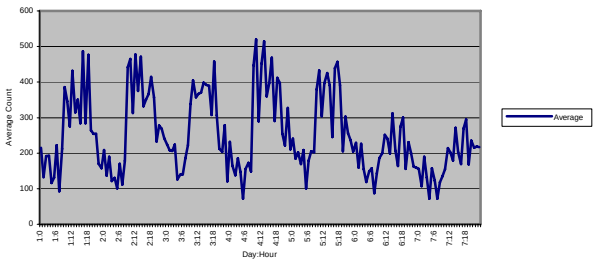

Figure 19: Weekly Averages (19/08/2002 to 30/09/2002)

These fluctuations were not a seen as a major problem in identifying breaking news stories. During the night few stories are published. However, should a big story occur then it is likely that reporters would be summoned to work in order to cover the story. It would be possible to normalise the data against the number of articles expected at that hour of the day, but that would likely lead to the augmentation of minor stories occurring during the night against those published during the day.

5.9.3 News 'Noise'

A major factor obscuring a breaking news story is news articles that appear on a regular basis. For instance during the Iraq Crisis there were so many articles being published per day on Iraq that it swamped other new stories. If we look at the weeks prior and including the 'Bali Bombing we see that there are regular articles being published on 'Bush', 'Washington' and 'Iraq'. These are effectively noise, hiding events such as the first reports of the Bali Bombing and Jimmy Carter getting the Nobel Peace Prize.

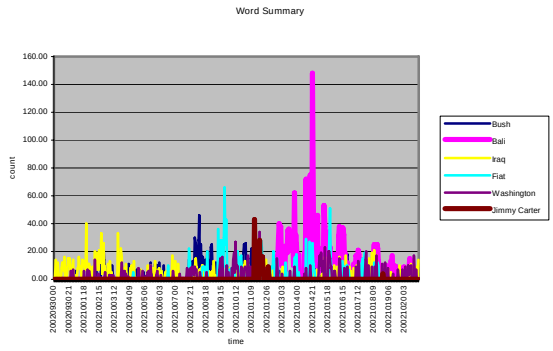

Figure 20: Word Count for period 30/09/2002 to 20/10/2002

The Bali incident was so large that it is not hidden by the news noise, in fact it has suppressed the publication of other articles (note how few articles are published on 'Bush', 'Iraq' and 'Washington' during the period when the Bali bomb blast occurred). If we wish to suppress this 'noise' we can use the simple technique of normalising the frequency based upon the incidence of words over a preceding time period. This normalized value will suppress repeated stories. For instance, we can see that the publication of articles on 'Iraq' is reasonably constant over the selected time period. However, there are no articles published on 'Bali' until the 12th October. We can use the formula defined in Equation 2 to generate a better representation of breaking news. This formula also takes into account coverage that a topic receives or 'topic spread'; Another problem identified affecting the identification of breaking news stories was that often a word would be repeatedly capitalised in all articles from the same source. A news source would prefix all stories with a location or author's name. This difficulty was fixed by adding a multiplying factor into the equation defined in Equation 2. The multiplying factor is the number of news sources (i.e. separate news wires and web sites) divided by a coefficient. So if a topic is widely reported it achieves a greater score.

Equation 2: News Scoring Algorithm

$$S = \frac{n_t}{\overline{n}_t} \beta \frac{n_F}{N_F}$$

where

S = score of topic

n_t = *number of occurrences of topic in last hour*

$\overline{n}_t$ = *average number of occurrences of topic per hour*

β = *multiplying coefficient*

n_F = *number of feeds (sources) from which the topics are derived*

N_F = feed scaling coefficient

Initially we used a rolling average $\bar{n}_t$ taken over the previous week (168 hours) and a multiplication coefficient of *β=6* (this number was reached purely through trial and error to produce the best result). The value of N_F was set at 3 as a breaking news story should usually be present in at least three different news sources. The results of this equation applied to the period covering the Bali bomb are illustrated in Figure .

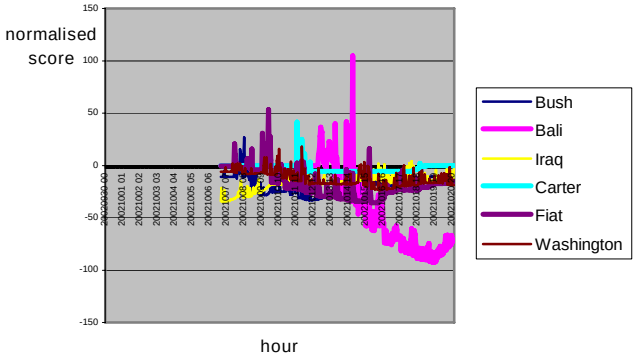

Figure 21: Normalised Word Frequency for period 30/09/2002 to 20/10/2002

This equation has effectively suppressed the 'Iraq' news stories. Most of the 'Bush' stories have been suppressed and it is much easier to distinguish the 'breaking news'. The 'Bali' bomb story is now identifiable as an event at 19:00 hours on 12th October 2002, whereas before the filtering the event it was only identifiable at 21:00 hours, two hours later. These two hours could be significant when notifying users of news events.

5.9.4 TOPIC FILTERING AND GROUPING

The basic data generated by extracting a list of words needs some 'cleaning' before we can use if for analysis. This process can be broken down into three distinct steps:-

5.9.4.1Stopword Removal

There are a number of words that are capitalised, which we do not want in the analysis. For instance in the English language the days of the weeks and months are capitalised but are unlikely to be of relevance to a news story. For this reason, a list of 'stopwords' is kept (on a per language basis). All incoming words being added to the system are checked in case they are stopwords. If they are then they are removed.

5.9.4.2Permanent Word Associations

There are many words that tend to have a permanent word association. For instance 'Saddam', 'Sadam' and 'Saddam Hussein' all refer to the same person. For this reason a list of these word groupings is kept and incoming words are processed to see if they

belong to any permanent word groups. These groupings need to maintained by hand and groupings may change over time. If a new person called Saddam appears in the news the association with ‘Saddam Hussein’ will need to be removed. Once a word association has been defined these words are then grouped together under a single topic heading, summing the two values and thereby giving an increased count for that topic.

5.9.4.3Transient Word Associations

Similar to the “Permanent Word Associations mentioned above there are often words which have close associations (e.g. ‘George’ and ‘Bush’ or ‘Shepherds’ and ‘Bush’). We can automatically create these associations from the articles analysed. This association is only stored on a short-term basis. For this reason the concept of ‘rolling topics’ was used. Rolling topics are defined associations between words, which are transient. As with the ‘permanent word associations’ defined above they are a way of grouping sets of words before performing an analysis, but these groupings are constantly changing. This grouping is done by seeing how often the words appear together in the same article. So if ‘George and ‘Bush’ appear together in 90% of all articles that contain one or other over the last few hours then it is likely that an article containing only ‘Bush’ would be related those that contain both. So we can group that article under the ‘George Bush’ topic. This association is only kept for a short period while it is statistically significant based upon recent articles. If a story about ‘Shepherds’ and ‘Bush’ appears the next day then the ‘George Bush’ association will have expired and a ‘Shepherds’ and ‘Bush’ association will be created. This technique does have the shortcoming that it could group together two distinct but contemporary stories about ‘George Bush’ and ‘Shepherds Bush’, but the chances of two significant stories with this association occurring is relatively rare. This filtering and grouping of words is performed immediately after the article’s title and description have been parsed and produces much better results by avoiding the duplication of stories under different headings.

5.9.5 INTEGRATION INTO EMM

The processing needed to discover breaking news stories can be quite intensive and needs to be done on a real-time basis. For this reason it was decided to set up a separate system from the ‘EMM’ servers so that there would be no impact on performance.

The ‘Breaking News’ system is ‘fed’ information from the EMM system (via HTTP), which passes information on each new article as it arrives. The title and description of the article are then parsed for capitalised words. These words are added to a database and processed to remove ‘stopwords’ and to group words that are about the same ‘topic’ (see 5.9.4).

The ‘breaking news’ system is now operational and fully incorporated into the ‘Europe Media Monitor’ and is used to feed the front page of the News Brief and for breaking news alerts. For example, during the arrest of Saddam Hussein, the level of ‘breaking news’ level reached a very high level as illustrated in Figure .

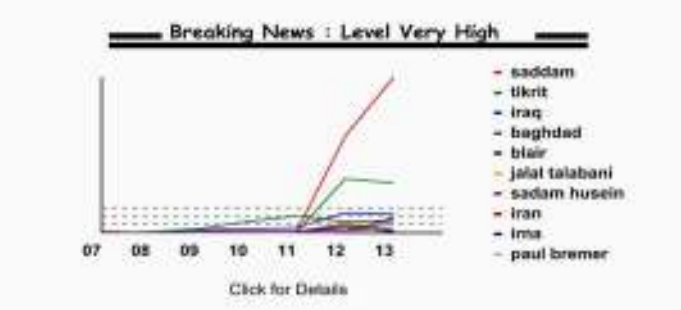

Figure 22 : Saddam Hussein Captured

This automatically triggered e-mails and SMS messages to subscribers of the alert system.

5.10 EMM email and SMS service

These systems handle the sending of emails and SMS messages to subscribers. Users who subscribe to email alerts through the web sites are added to an XML configuration file. There are two email systems running – one for the public site and one for the Commission site. The software uses the Java Mail interface to a standard SMTP server. In this case the JRC's mail gateway is used to send emails.

When an article triggers an Alert the alert system, a check is made to see if any persons are subscribed to an immediate email., In this case the RSS item is sent the the email system and sent automatically. The other method of subscription is through a daily summary email. In this case every morning at 7 am . the Email system sends the found articles over the last 24 hours as an HTML email containing the content of the alert RSS file.

An automatic SMS service is in operation to keep spokespersons and selected key assistants informed of EU related news reports out of hours and at weekends. This service follows a timed roster definition defined on the RNS system and described there. The alert that triggers the SMS messages is called SMSauto and the definition of the keywords is defined through the RNS alert editor interface and managed centrally in Brussels. This SMS service is outsourced under contract to a telecom provider who provides a discounted price per SMS. Subscribers and Rosters are maintained in XML files .

A manual SMS service is operated within working hours and is controlled from RNS. In this case the SMS messages are sent directly through a GSM modem operated at JRC Ispra. The cost of this more limited number is bourn by the JRC service phone contract.

The figure below shows the output of an automatic monitoring system which shows how many emails and SMS have been sent over the last 7 days.

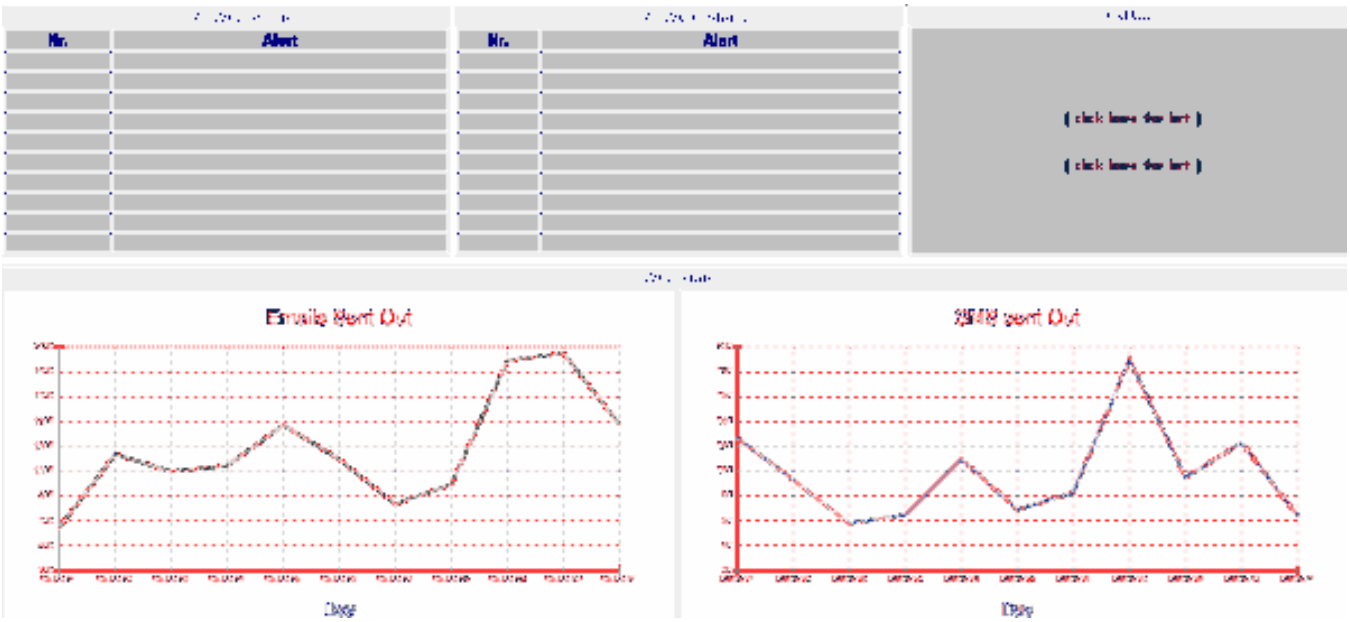

Figure 23: Monitoring automatic SMS and Email Alerts

5.11 EMM News Brief

The News Brief is the presentation layer of EMM for web users. It presents an ever changing content reflecting the current content of the alerts and the breaking news stories. The design is intended to be that of a newspaper. All the content of Alerts are accessible as RSS files, and all the Breaking News stories in each language are accessed as XML files.

The software is based on XSLT transformations of the XMNL content into an HTML front end. The alerts are collected into groups usually for each DG and I nto themes and country regions. The definition of which alerts are grouped into which categories is defined in an XML file held on each of the alert systems. This file is called subscriber.xml and the output of the NewsBrief is modified by editing this file. Alerts which do not appear in this file are not presented on the NewsBrief. An example section of this file is given below.

```
<subscriberInfo>
        <subscribers>
                <subscriber id="EMM">
                        <subscriberRef id="European Parliament"/>
                        <subscriberRef id="HotNews"/>
                        <subscriberRef id="Court of Justice"/>
                        <subscriberRef id="Court of Auditors"/>
                        <subscriberRef id="Committee of Regions"/>
                        <subscriberRef id="Commission"/>
                        <subscriberRef id="Agencies and Offices"/>
                </subscriber>
                <subscriber id="Committee of Regions">
                        <alertRef id="CommitteeOfRegions"/>
                </subscriber>
                <subscriber id="Court of Justice">
                        <alertRef id="CourDeJustice"/>
                </subscriber>
                <subscriber id="Court of Auditors">
                        <alertRef id="ECA"/>
                </subscriber>
```

```
<subscriber id="United Nations">
        <alertRef id="KofiAnnan"/>
        <alertRef id="PeaceKeeping"/>
        <alertRef id="SecurityCouncil"/>
        <alertRef id="UNbodies"/>
        <alertRef id="BorderDisputes"/>
</subscriber>
<subscriber id="Commission">
        <subscriberRef id="Commissioners"/>
        <subscriberRef id="DGs and Services"/>
        <alertRef id="ECnews"/>
</subscriber>
<subscriber id="Commissioners">
        <alertRef id="JoseBarroso"/>
        <alertRef id="MargotWallstrom"/>
        <alertRef id="GunterVerheugen"/>
        <alertRef id="JacquesBarrot"/>
        <alertRef id="SiimKallas"/>
        <alertRef id="FrancoFrattini"/>
        <alertRef id="VivianeReding"/>
        <alertRef id="StavrosDimas"/>
        <alertRef id="JoaquinAlmunia"/>
        <alertRef id="DanutaHubner"/>
        <alertRef id="JoeBorg"/>
        <alertRef id="DaliaGrybauskaite"/>
        <alertRef id="JanezPotocnik"/>
        <alertRef id="JanFigel"/>
        <alertRef id="MarkosKyprianou"/>
        <alertRef id="OlliRehn"/>
        <alertRef id="LouisMichel"/>
        <alertRef id="LaszloKovacs"/>
        <alertRef id="NeelieKroes"/>
        <alertRef id="MariannFischerBoel"/>
        <alertRef id="BenitaFerreroWaldner"/>
        <alertRef id="CharlieMcCreevy"/>
        <alertRef id="VladimirSpindla"/>
        <alertRef id="PeterMandelson"/>
        <alertRef id="AndrisPiebalgs"/>
</subscriber>
<subscriber id="DGs and Services">
        <subscriberRef id="ADMIN"/>
        <subscriberRef id="AGRI"/>
        <subscriberRef id="BUDG"/>
        <subscriberRef id="COMP"/>
        <subscriberRef id="DEV"/>
        <subscriberRef id="ECFIN"/>
        <subscriberRef id="EAC"/>
        <subscriberRef id="EMPL"/>
        <subscriberRef id="TRANSPORT"/>
        <subscriberRef id="ENERGY"/>
        <subscriberRef id="ELARG"/>
        <subscriberRef id="ENTR"/>
        <subscriberRef id="ENV"/>
        <subscriberRef id="RELEX"/>
        <subscriberRef id="FISH"/>
        <subscriberRef id="SANCO"/>
        <subscriberRef id="INFSO"/>
        <subscriberRef id="DGIT"/>
        <subscriberRef id="MARKT"/>
        <subscriberRef id="INTERP"/>
        <subscriberRef id="JLS"/>
```

```
        <subscriberRef id="JRC"/>
        <subscriberRef id="PRESS"/>
        <subscriberRef id="REGIO"/>
        <subscriberRef id="RTD"/>
        <subscriberRef id="TAXUD"/>
        <subscriberRef id="TRADE"/>
        <subscriberRef id="AIDCO"/>
        <subscriberRef id="EUROSTAT"/>
        <subscriberRef id="ECHO"/>
        <subscriberRef id="OLAF"/>
</subscriber>
<subscriber id="Agencies and Offices">
        <subscriberRef id="MaritimeSafety"/>
        <subscriberRef id="EEA"/>
        <alertRef id="EuropeanOmbudsman"/>
        <alertRef id="eurofound"/>
        <subscriberRef id="Trade Marks and Designs"/>
        <subscriberRef id="United Nations"/>
        <subscriberRef id="OSHA"/>
</subscriber>
<subscriber id="HotNews">
        <alertRef id="UKPresidencyEU"/>
        <filterRef id="Iran-NuclearReprocessing"/>
        <alertRef id="Referendum"/>
        <alertRef id="Eurobarometre"/>
        <alertRef id="EuropeanConstitution"/>
        <alertRef id="Iraq-EU"/>
        <alertRef id="Romania-EU"/>
        <alertRef id="Croatia-EU"/>
        <alertRef id="Turkey-EU"/>
        <alertRef id="Bulgaria-EU"/>
</subscriber>
```

alertRef refers directly to an alert definition
suscriberRef refers to a grouping of alerts into a category.

The same subscriber file is also used by the WAP site to define the hierarchy of links available to WAP browsers.

The NewsBrief contains a free text search interface to allow users to search the index of recent articles. The index normally goes back about one month of articles. The search results are returned in RSS 2.0 format and formatted into paged HTML pages for the user.

The NewsBrief website receives heavy accesses and a cached versioning system has been adopted to cope with the traffic. The main pages for each language are generated automatically each 10 minutes. This avoids each and every user causing an XSLT transformation, thereby easing the load on the machine.

The Commission version of the NewsBrief includes presentations of the RNS generated EMM Panorama and the 30 or so Press Reviews generated each day from the capitols. These are published as XML files and transformed by XSLTs into the HTML version visible on the NewsBrief.

The NewsBrief is currently available in 25 languages – all languages of the EU. Each language is supported by an independent breaking news system and a filter on the

Alerts for that given language. The Alert system has been modified to ensure that at least 2-3 articles are available in the less monitored languages. This is to avoid that English dominates the alert content to the exclusion say of Maltese.

A schematic overview of the NewsBrief application is given in the figure

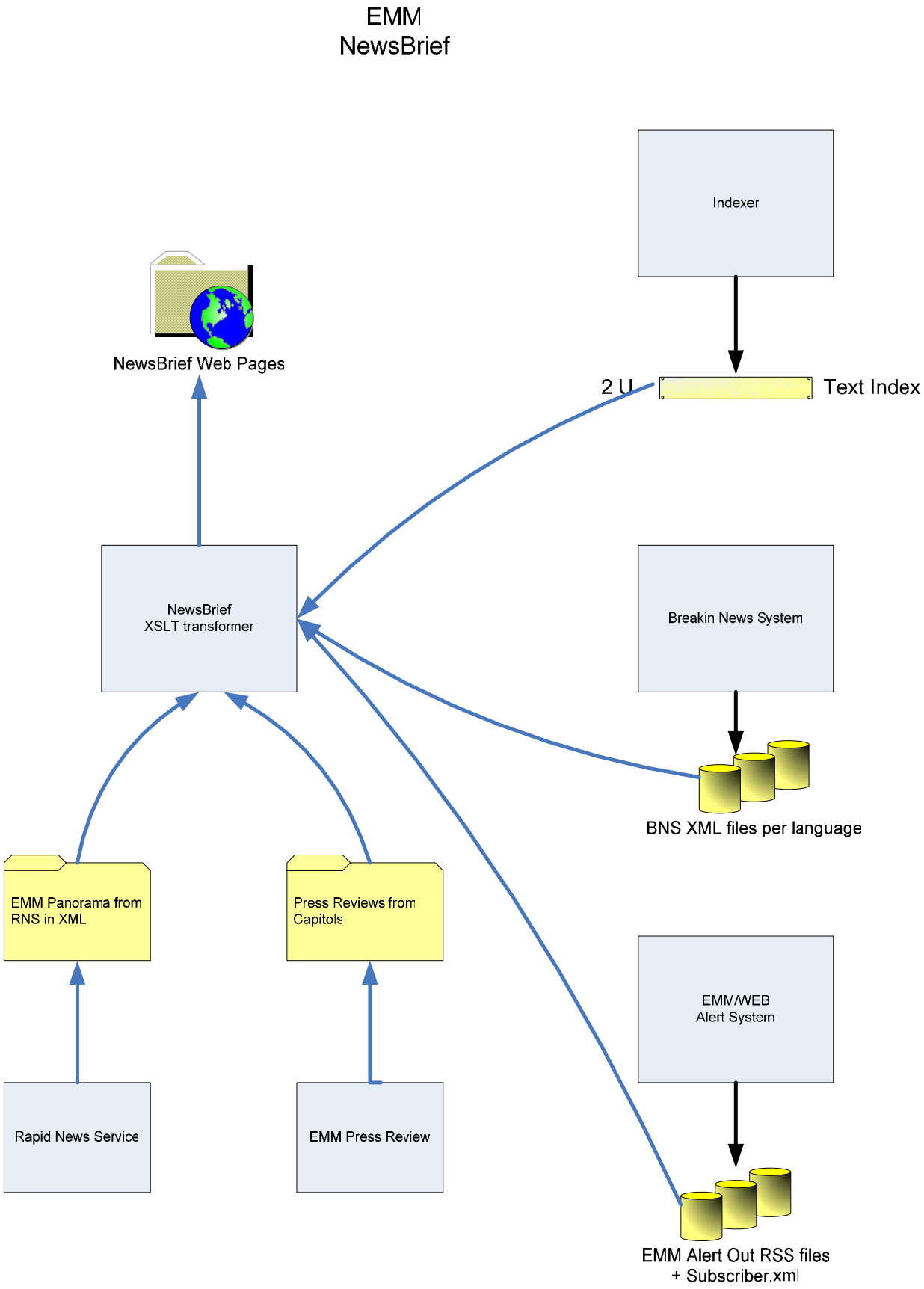

Figure 24: Schematic Overview of EMM NewsBrief application. The public version does not include EMM panorama or Press Review.

The Newsbrief allows users to subscribe to email alerts. By navigating to a given alert file the user clicks on the “suibscribe to this alert” link. This interface maintains the alert subscription details found in an XML file within the webapp AlertMail held on the Newsbrief service. For security reasons the software needs to ensure that the user’s email address is valid. Therefore it sends an email to the email address given by the user instructing them to click on a URL within the email. This URL is generated dynamically by the software and is based on any existing information for that user.

The generated web page offers the user two choices for the selected alert. They can either subscribe to an immediate alert whenever an article is found in that category or they can subscribe to a daily update. In practice many emails per day can result from the first choice, so it is recommended to start with a daily summary. The interface also allows the user to unsubscribe or change subscriptions from any existing alerts.

This is illustrated in the figure below

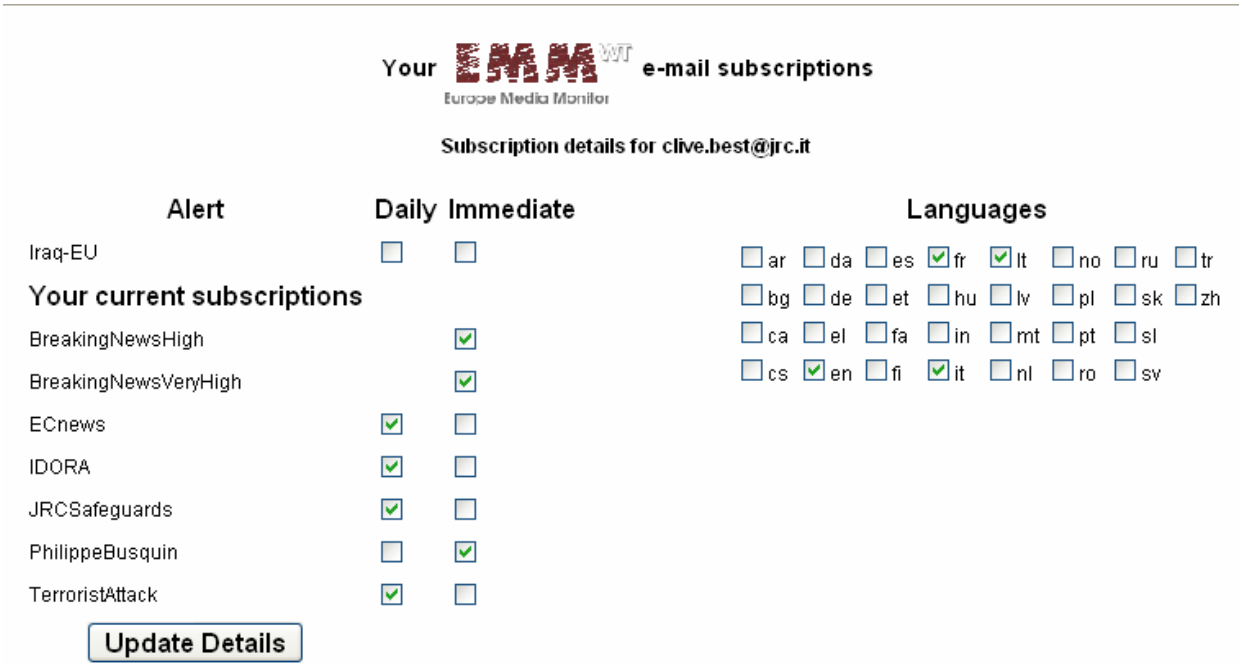

Figure 25: The email subscription interface

5.12 Rapid News Service

The Rapid News Service is a system specifically designed for use by DG PRESS news team. It aims to provide them with a single interface to live incoming news reports from the News Agencies and selected web sites to enable them to perform the following tasks :

- Alert key personnel by email to a major news report
- Alert key personnel by SMS message of a major event
- Edit and publish the twice daily EMM Panorama using News Agency wires.
- Maintain mailing lists of key personnel
- Maintain an automatic Roster for out of hours SMS alerts from a special alert SMSAuto.
- Allow users access to edit specific alert definitions (including SMSauto)
- Eventually to allow the editing and production of the Revue de Press using press cuttings from the EMM Review system in the same way as for EMM Panorama.

RNS accesses a selection of News Agency Wires and selected TV/Radio sites through an AlertFilter. This filter removes articles which trigger the Sport alert to reduce the large number of Sport results. RNS also has access to the Web alerts and can search the index.

Members of the News Team monitor the filtered articles and can manually send emails and SMS messages to lists of key persons. They can further filter by language and share the work between 2 or more members. One of the main tasks each day is the production of EMM Panorama. RNS has a drag and drop interface which makes it easy to select the content for EMM Panorama. Each issue is organised according to topic headings. These can be modified through the user interface. Much effort has gone into the user interface to make this as easy as possible.

When the user “publishes” EMM Panorama it is written out as an XML file for use by a number of systems. EMM Newsbrief reformats the latest edition for display. DG Press News Portal reformats the table of contents using a Javascript technique and the file is posted also to DG PRESS Avantgo synchronisation server for access over handheld Qtek devices.

EMM Panorama is converted to both WORD and PDF format using XSL/FO for printing and distribution inside the Commission.

The management interface is accessed through a special login. This interface . maintains the contact lists (XML file), defines user logins and which alerts they can view/edit, edits the on-call roster and edits the SMSAuto alert.

The Alert editor is a web based interface which permits to change alert definitions over through the RNS interface. It is based on a dictionary of terms. Terms can be added to or removed from this dictionary. The alert is defined as a set of patterns and weights or as a set of combinations of terms. Once an Alert has been updated and saved. It will be activated within one hour.

A user guide for RNS is available in a separate document.

RNS Overview

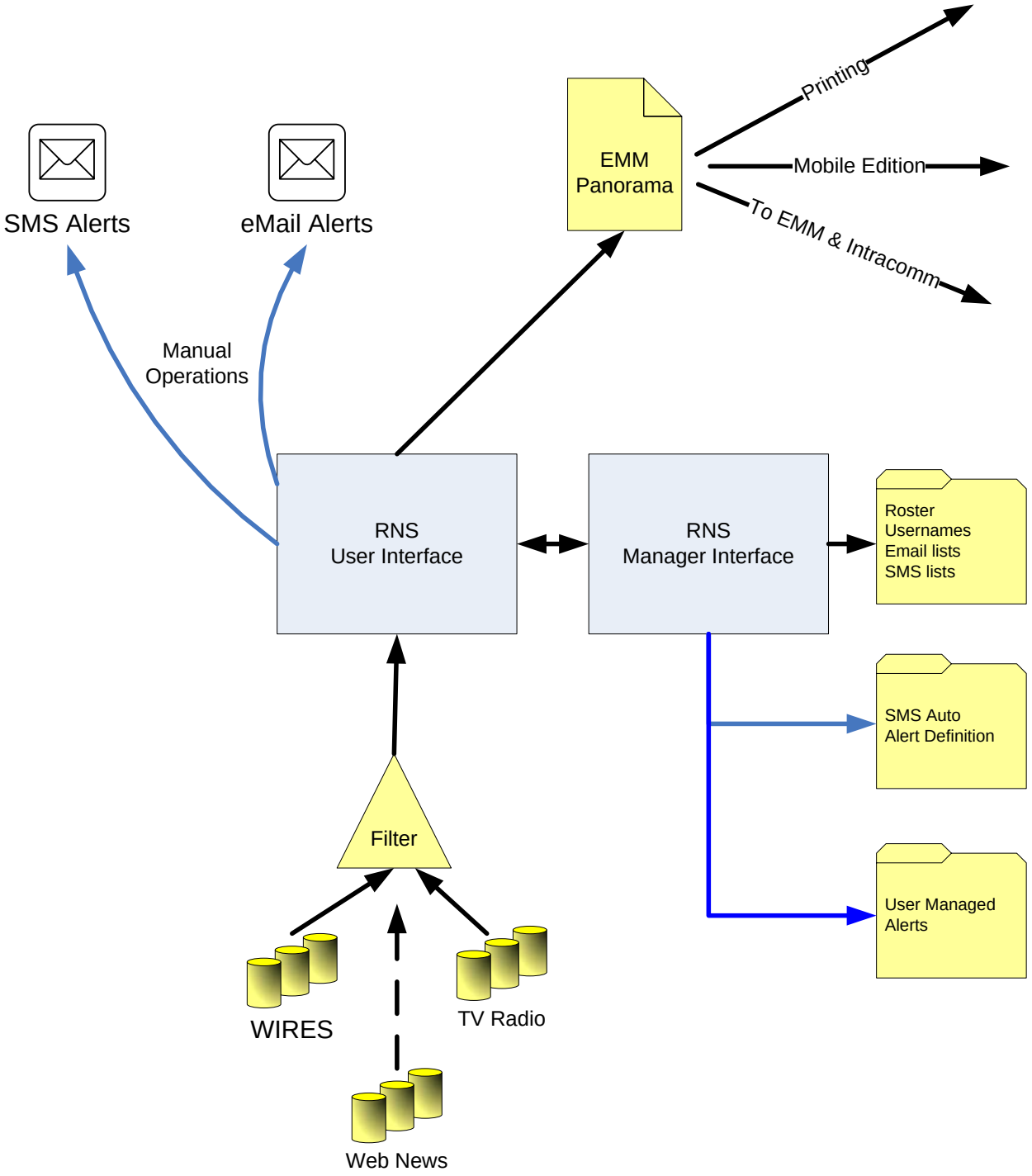

Figure 26. A Schematic Overview of the Rapid News Service (RNS)

5.13 EMM Review (Press Review)

The EMM review is a separate system from the rest of EMM to allow for the electronic editing and publishing of the daily reviews from the capitols of the national press and TV, radio. It is a web based authoring and editing system which is used by around 50 people around the world each day. Access is through controlled logins assigned by DG PRESS. About 30 delegations are usin g the system daily. Normally two reviewers are working from each capitol – audio visual and printed press reviews.

There is an administrator user who can create users and manage the topic classifications for the reviews. Each reviewer keeps a controlled list of sources eg. Newspapers and TV stations. The review is produced by entering each item one at a time, classifying it and labelling it with the source origin. The reviewer may also attach a press cutting as an image or as a PDF file. This is a scanned image of the original article. Once the reviewer is happy with the content s(he) publishes the content through the interface. This causes an XML version of the review to be written to the server and an email to be sent to Intracomm for automatic publishing to the Intracomm web site. The XML versions are then syndicated to EMM NewsBrief and to the PRESS news Portal, where they are displayed in HTML after an XSL transform.

All the reviews are held in a MySQL database. This allows DG PRESS staff to perform searches across all the reviews and to find reviews from any date in the past. The results of searches can then themselves be printed as reviews. Often spokespersons are interested in all references to a certain Commissioner or a certain subject and this allows a fast convenient summary to be generated.

The software is based on JSP pages running on Tomcat using a JDBC connection to a local MySQL database. It is not designed for heavy public usage, but rather for interactive updates by the reviewers. The reviews themselves are published on EMM and Intracomm for broad access purposes.

A full user guide for Press Review is available in a separate document.

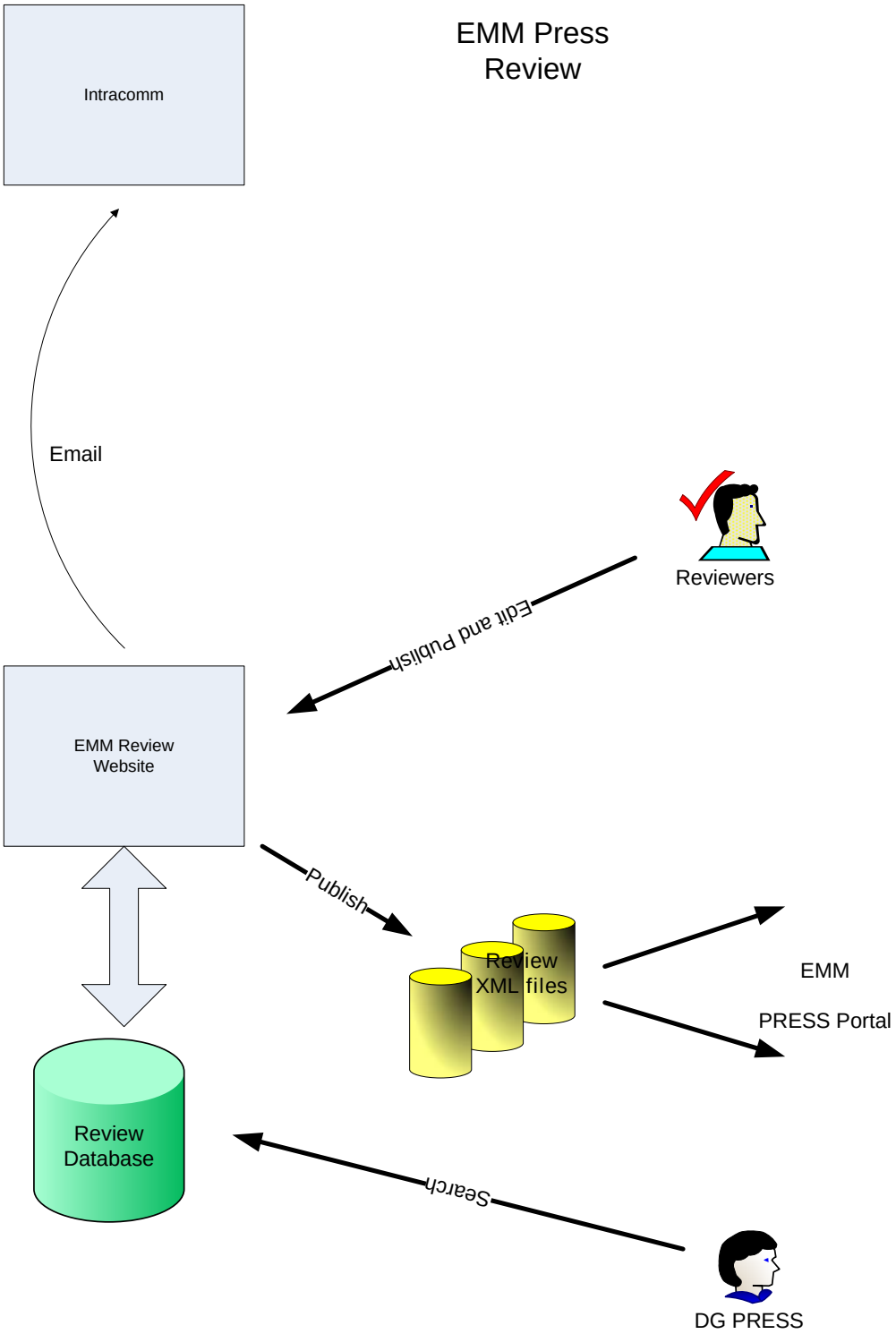

Figure 27. A schematic overview of the EMM Press Review System

5.14 EMM Wap Services

The software used in EMM Wap system uses the classic Model-View-Controller (M.V.C.) design pattern. This separates the application into three modules :-

Model : The data and the encapsulating processes which need to be represented and controlled in several ways. For the EMM system the data is stored in XML files.

View : The presentation layer. This is responsible for displaying the data to the user, whether it be on a PC, WAP phone or PDA. By using XSL Transforms (XSLTs)

Controller : The Programming Logic representing the business processes for controlling and processing the data.

Requests from the users are sent using HTTP to the 'Controller' servlet. The controller then interprets this request and calls the XSLT processor (implemented using a servlet) using the requested 'Model' (XML file) and 'View'(XSLT). The XSLT processor then generates the WML to be sent to the client.

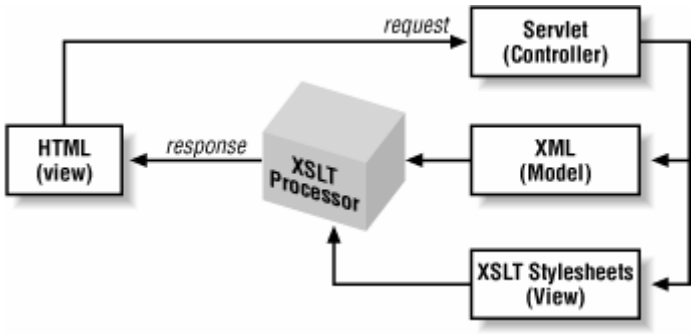

Figure 28. MVC architecture overview

Because the MVC design pattern was used the implementation of the EMM WAP site was made considerably easier.

5.14.1 Access EMM alerts and read the article text

As mentioned in the previous section a servlet is used to field HTTP requests from the internet. When a request is sent to the EMM site a check is made on the browser viewing the site. If the browser is a 'WAP' browser then the request if redirected to the EMM WAP Web Application. This then uses an XSLT processor to generate the top level WAP page (after displaying a splash screen) from an XML file. From the top level it is possible to navigate through the alert categories until an individual alert is reached:-

Figure 29. Navigation of the Alert Hierarchy

These pages are generated by recursively calling an XSLT passing the XML file of interest as a parameter. Once an alert had been selected a list of articles associated with that alerts is displayed:-

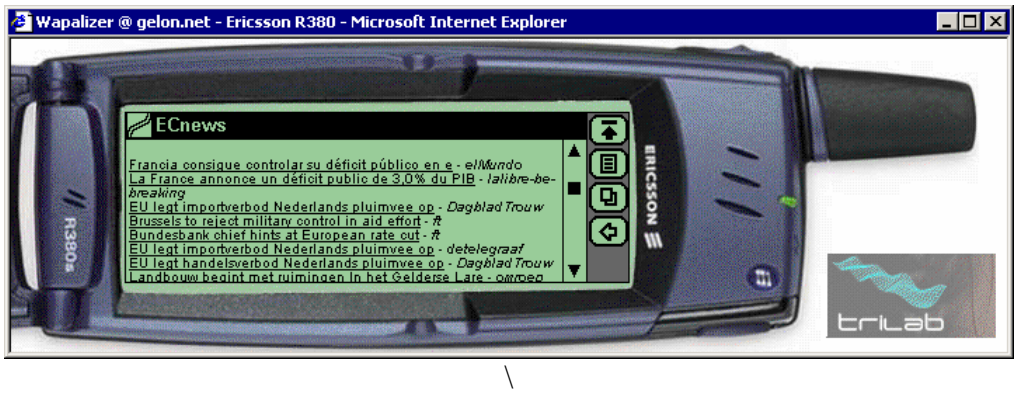

Figure 30. List of articles for an Alert

The user can then go on to read the title and description of an article. At this point the information displayed is still generated from XML files derived from the ‘EMM Alert System’. From this point the user can go on to read the full ext of the article. If the user selects ‘Read Article’ the text of the article is dynamically retrieved from the web site and translated into a WAP page. The process extracts the HTML from the web site, locates the main text of the article then converts it into ‘WML’ format.

Figure 31. Dynamically converting an article from a News Site to WAP format on a real phone

6 Accessing EMM content on third party web sites

The live content of EMM is held in RSS 2.0 files and can therefore be syndicated to other web sites. Today about 10 sites pick up and receive the content of EMM alerts in real time for redisplay on their site. There is an HTTP API which provides acces to the alert files and allows the user to filter the results according to the extra tags in the RSS file. Effectively this means that extra criteria can be applied such as Language, source etc. The interface is :

http://emm.jrc.it/Alert/filterAlertXML.jsp?ID="AlertName"

Extra filter parameters can be appended as HTTPET parameters as follows

&language=en - select only English articles (en,fr,de,es …..)
&feed=bbc,skynews - select articles ONLY from BBC OR skynews
&title=Iraq - select articles only where Iraq appears in the article
&similar="AnotherAlert" - selects articles which also triggered AnotherAlert

Parameters separated by commas mean that one value OR the other must appear.

Some example will help to understand the usage

http://emm.jrc.it/Alert/filterAlertXML.jsp?ID=CommunicableDiseases&language=en,fr&trigger=cholera

will return an RSS file of Communicable disease articles in Englisg and French where the keyword Cholera was found.

http://emm.jrc.it/Alert/filterAlertXML.jsp?ID=Conflict&similar=Iraq&language=en

filters English articles concerning Conflict AND Iraq.

The third party web site can format these RSS files into the HTNL content of their page. Ideally they cache the results at regular intervals , run the conversion and then include the content on the server. An alternative method which is easy to implement uses the users web client to render the results using Javascript. In this case EMM provides another interface – jsfilterAlert.

http://emm.jrc.it/Alert/jsfilterAlert.jsp?ID=ECnews for headline news on EC.

&mode=full will also return the descriptions.

The javascript code to render that into HTML or into scrolling headlines can be coppied from some example sites :
http://www.accent-network.org
http://sotw.jrc.it/regional/SouthAsia.html

The result for the second example is illustrated in figure

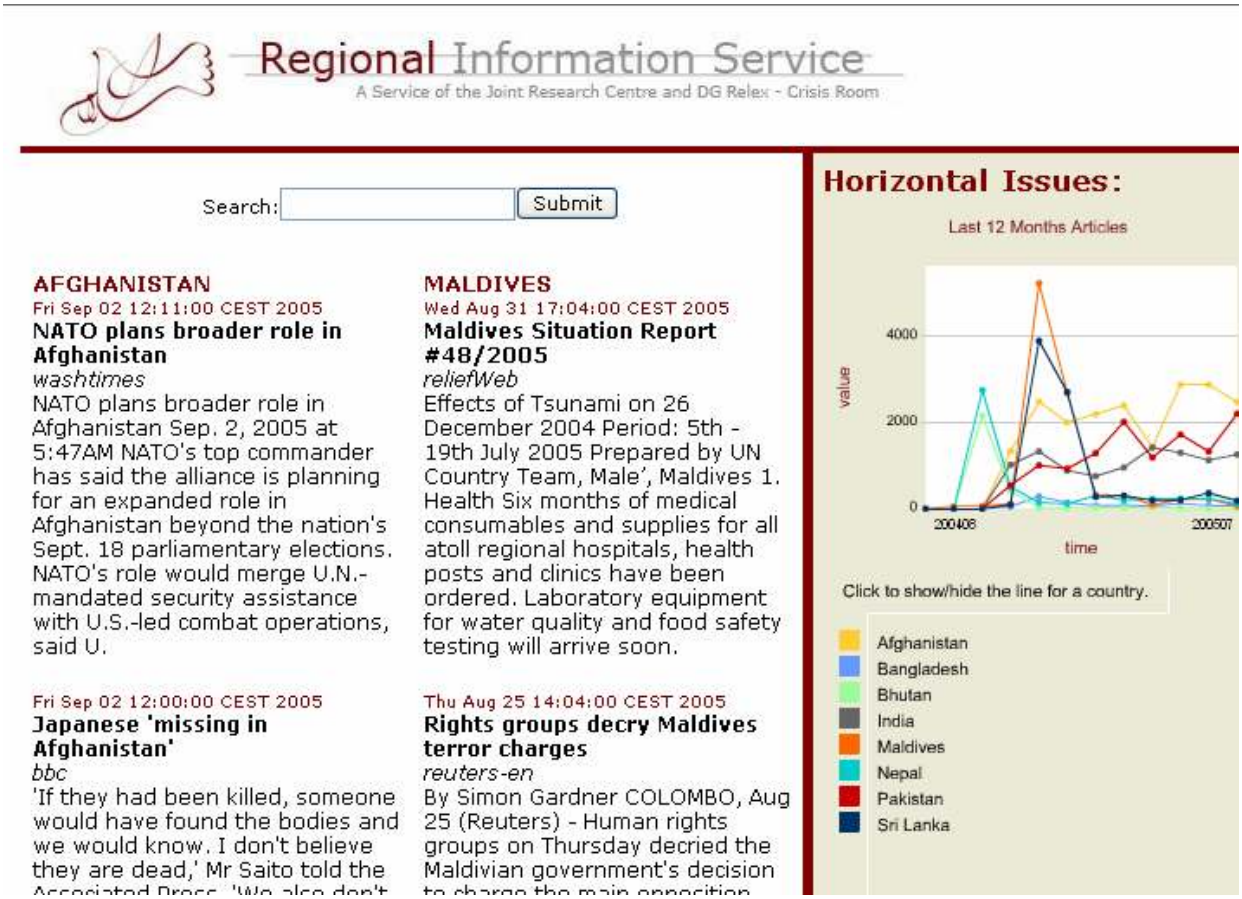

Figure 32. Illustration of the use of EMM news syndication to a third party web site

7 Conclusions

EMM is a mature service which has been in operation since 2002. It is being continuously improved and now forms a suite of services which support the Media Monitoring of the European Commission. Wider access to EMM is increasing in popularity through the public website at http://press.jrc.it Many EU agencies and external organizations depend on it for automatic news monitoring. It operates in over 30 languages and monitors news sites published around the world around the clock. The new News Explorer service can track major events across time and across languages. In particular it can monitor persons and organisations in the news and which other entities and topics they are most associated with.

8 Acknowledgement

The authors would like to acknowledge the support and encouragement of the ex Director of Communications at the European Commission – Niels Thogersen. Without his enthusiasm and encouragement EMM would never have reached the level of usage it currently enjoys.

European Commission

EUR **22173 EN** – DG Joint Research Centre, Institute **for the Protection and Security of the Citizen**
Europe Media Monitor - System Description
Best, Clive - Van Der Goot, Erik - Blackler, Kenneth - Horby, David - Garcia Domingo, Teofilo

Luxembourg: Office for Official Publications of the European Communities
2005 – 97 pp. – 21 x 29.7 cm
EUR - Scientific and Technical Research series; ISSN 1018-5593

Abstract

The Europe Media Monitor (EMM), has been developed by JRC on behalf of DG PRESS.DG PRESS drive its development by specifying their requirements at regular monthly meetings. Since January 2005 this service has been formalised through an Administrative Arrangement between DG PRESS and the JRC. EMM provides the software services which process incoming News Reports from News Agencies, Press Reviews from Capitols and the major web based news services in Europe. EMM began operations in May 2002 and has expanded its services over the intervening years to become Commission's primary source of live news related services. The front end web interface is through an automatically generated NewsBrief, which is updated every 10 minutes with the latest top stories. The internal Commission version of the NewsBrief is accessed by up to 20,000 users per day. The external public version of EMM which contains only open sources has found a growing number of customers in Europe and elsewhere. However, this main service is only one of a suite of other services that EMM provides. This document aims to describe these services and document the software architecture of EMM.

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

Publications Office
Publications.europa.eu