

Europe Media Monitor (EMM)

<http://emm.newsbrief.eu>

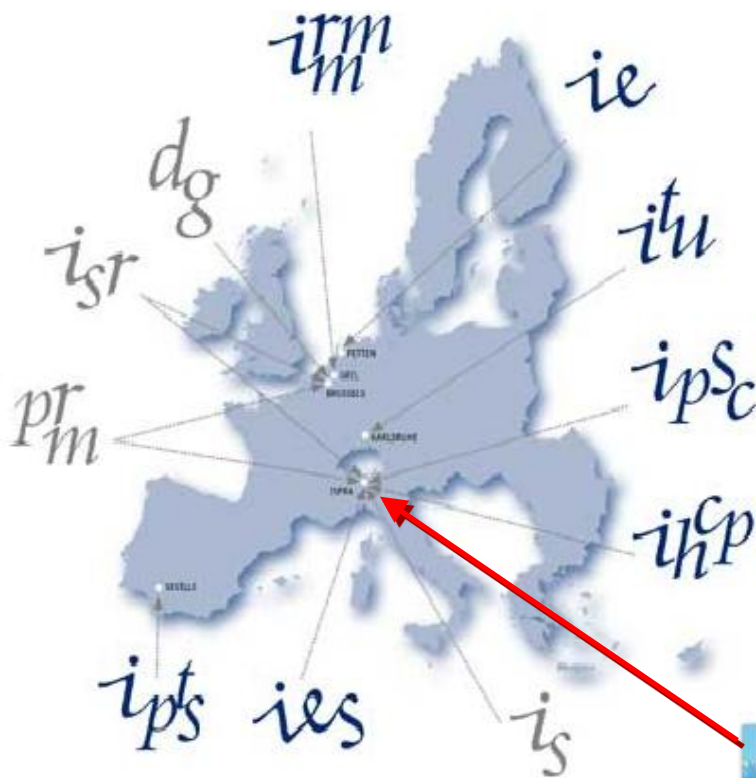


Erik van der Goot & Optima Team

European Commission – Joint Research Centre (JRC)
Institute for the Protection and Security of the Citizen (IPSC)
Global Security and Crisis Management (GLOBESEC)
Open Source Text Information, Mining and Analysis Action

- **The Joint Research Centre of the European Commission**
- **EMM: Where and why did it all start, rationale**
- **EMM: System overview, functionality**
- **EMM in depth, development and system architecture**

- **Joint Research Centre is a General Directorate of the European Commission**
With more than 2500 staff the JRC is one of the largest DGs in the Commission
- **Technical&Scientific support for law and policy making**
- **5 locations in Europe, originally linked to Nuclear Research (except for Sevilla)**
- **Largest site (2000+ staff) in Ispra, Italy (Lago Maggiore)**



BRUSSELS (BE)

The Directorate General (DG)
The Institutional and Scientific Relations Directorate (ISR)
The Programme and Resource Management Directorate (PRM)

GEEL (BE)

The Institute for Reference Materials and Measurements (IRMM)

KARLSRUHE (DE)

The Institute for Transuranium Elements (ITU)

ISPRA (IT) [Download the Ispra site Brochure \(English - Italian\)](#)

The Institute for the Protection and Security of the Citizen (IPSC)
The Institute for Environment and Sustainability (IES)
The Institute for Health and Consumer Protection (IHCP)
The Ispra site Directorate (IS)

PETTEN (NL)

The Institute for Energy (IE)

SEVILLE (E)

The Institute for Prospective Technological Studies (IPTS)



- **EU Commission Media Monitoring (until 2001/2002)**
 - Traditional cut and paste for printed press only
 - Monitoring of incoming news wires (e.g. Reuters, AFP)
 - Simple keyword based filtering of wires
 - Manual selection of printed press items
 - Human classification of items

- **Potential problems**
 - Not 'real-time' for mainstream media: printed press typically once a day
 - Limited coverage: not all media is printed
 - Inaccurate and incomplete classification: subjective and limited number of categories
 - Labour intensive and expensive: limited number of articles per reviewer per day, requires topical knowledge and requires language knowledge

- **Challenges (as seen in 2002)**
 - Enlargement (+10 countries, 15→25): more media, more languages
 - More use of electronic publishing (media)
 - Electronic distribution of media monitoring results (web+mobile)
 - Automatic alerting functions
- **New approach: EMM - a one stop shop for Media Monitoring**
 - Facilitate (not replace) human Media Monitoring activities
 - Provide monitoring of on-line sources and other digital channels (e.g. news wires)
 - Improve coverage, number of languages, analysis.
 - Apply automatic categorization and further analysis to all sources
 - Provide new services like automatic e-mail, sms, mobile editions etc.
 - **Provide editorial system to manage the information and produce newsletters etc.**

Important: EMM is *not* Yet Another Internet Search Engine

Wide coverage

Many sources

- Local, Regional, National and International coverage

Many languages

- Multilinguality & cross-lingual information access

Fast coverage

High frequency monitoring of sites, some sites every 5 minutes

Detect new articles

Use RSS where possible (in 2001??)

Use HTML news pages

Analyse full content

Extract 'meaningful' text from article HTML





Italian - German



English - French



Spanish - Portuguese



Some EMM statistics

Gathers approximately **100,000 new news articles per day**

In **>42 languages**

From ~ 2600 news portals world-wide (roughly 5500 pages/RSS feeds), plus 20 news wires and some specialist sites

Classifies all news according to **hundreds of subjects and countries.**

(~1000 category definitions, ~30.000 key word patterns and pattern combinations)

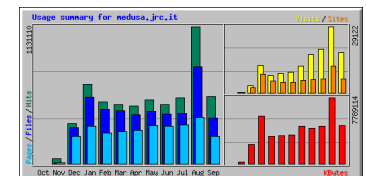
24 / 7: Runs 24 hours per day, 7 days a week.

Started in 2002

Up to 1.5 Million hits per day for public websites (EMM NewsBrief, MediSys, NewsExplorer)

But more importantly, 25.000 visitors/day on the public sites

Developed, built, maintained and run by the JRC.



- **EMM**
 - EU Commission Media Monitoring
- **MediSys, monitoring of health related open source information**
 - Collaboration with ECDC, EFSA, DG SANCO, GPHIN, Chinese Health Authorities
<http://medisys.newsbrief.eu>
- **HEWS II, Humanitarian Early Warning**
 - Collaboration with WFP, UNHCR
- **Africa News Monitor (EMM for Africa)**
 - Collaboration with European Parliament and UNDESA (APKN)
- **And an increasing number of customer installations in EU Agencies, various EU and member state institutions, and beyond**

MediSys
 Home | Diseases | Bioterrorism | Other | Search | Advanced Search

Most Active Topics

- Biodefense**
- In combination with: Iran, Islamic Republic Of;**
- Toxic**
- In combination with: Iraq;**
- Coma**
- In combination with: Italy;**

Today's Hot Topics

Last 24h Alert Statistics for all alerts

Country/Region	Alert Count
Iran-BioDefence	9
Iran: Toxic	9
Italy: Coma	6
Alghamstan:Counter...	11
Germany:Acid/...	6
Haiti:BioDefence	5
Uganda:ALDS-HIV	7
Greece:Shin/flu	6
Italy: Shin/flu	12
Italy: Shin/flu	16

Recent disease incidents provided by the University of Helsinki

Disease	Time	Location	Cases
Swine Flu	2009.08.30	France	(6) 209-438 cas
Influenza	To date	USA/Georgia	four deaths
Swine Flu	2009.08.30	USA	(6) 7 nouveaux cas
Pneumonia	Sunday	Israel	The woman
Swine Flu	last week	UK	5,000 new patients

Orla Influenza fa paura, c'è un giovane in coma
 Colpito un ventiquattrenne di Parma. I primi sintomi dopo il ritorno dalle vacanze sulla riviera romagnola. Ora è ricoverato in terapia intensiva a Monza. I medici: «Il ragazzo respira soltanto grazie alle macchine»...

NewsDesk Service (a.k.a. RNS) Editorial Interface

1. Section 1

Barroso will Gusebauer als Vice [1] de de
 EU-Kommissarpräsident Barroso wunscht sich den irischen EU-Kandidat als neuen Stellvertreter. Die Regierung will davon noch nichts wissen.

EU-Kommission will sich keinen Maulkorb verpassen lassen [1] de de
 Irlands Ministerpräsident Bertinotti hatte gefordert, dass sich nicht mehr jeder EU-Kommissar sondern nur Kommissionspräsident Barroso oder sein Sprecher Lattebergem öffentlich äußern dürfe. Dies wurde jetzt aus Brüssel klar abgelehnt. (ap) Die Forderung des irischen Ministerpräsidenten.

EU Presidency and Barroso must react to Berlusconi immediately, says Schulz [1] de de
 Martin Schulz, Leader of the S & O Group in the European Parliament, said: "The Italian Prime Minister, Silvio Berlusconi, today attacked the European institutions. He showed intolerance towards the EU Commission's immigration policy. We threatened to hold up the workings of the European Council..."

Barroso's reputation needs to be restored [1] de de
 A delegation from employer' organization BEC today met in Brussels with senior European Union officials to outline the view of Irish employers on the serious economic challenges facing both Ireland and the EU.

Participation of José Manuel Barroso at the BEC meeting [1] de de
 Participation of José Manuel Barroso, President of the EC, at the Irish Business and Employers Confederation (BEC) meeting.

Bec calls for Yes vote on Lisbon [1] de de
 Business group Bec has called for a Yes vote in the Lisbon Treaty referendum and said the restoration of Ireland's international reputation is vital to reduce the indirect cost of government borrowing' and the strain this is putting on the public finances.

Barroso Wiederwahl rückt näher [1] de de
 Nielsen und Machi Der Postenbesitzer um Top-Positionen in der EU geht in die nächste Runde. Kommissionspräsident Barroso will nächste Woche in den nächsten Wahlkreis "gehen", wie es bei den europäischen Sozialdem. Am 18. September wird er im EU-Parlament wahrscheinlich wiedergewählt.

"Kürbis" Kommissar von Reinhard Gitter: "Europas Elfen sind made gemorend" [1] de de
 In den Europäischen Union herrscht Ernährungsnot. Das ist fast = Wein (OT5) = in Brüssel herrscht wieder reges Pfl. Treiben. Kommissionspräsident Barroso will wieder gewählt werden, während sein Arbeitsprogramm vorliegt, auch Kommissare sind politischem Ausgang.

Arto Linn: Kallioa kaatumineen webi Saksienaste koe presidentiksi [1] de de
 Rahvala ja Riigikogu liige ja eurovalimis kandidaat Arto Linn loob Saksiaa lõunaeurovõitluse võidul viie lesta nappimise peenestööks, mis omakorda süüdistab heid Edgar Saksienaste lõunaeurovõitluse kandidatuur.

Streit um Flüchtlingspolitik: Berlusconi kritisiert EU-Kommission [1] de de
 Der italienische Ministerpräsident Silvio Berlusconi hat die Entscheidung der EU-Kommission und ihre Generäle kritisiert. Am Freitag der Gendarmen hat Beginn des Zweiten Weltkriegs darfs Berlusconi mit einer Blockade der Europäischen Union. Der sozialistische Fraktionsvorsitzende Martin Schulz kündigte umgehend ein Misstrauensvotum an.

Berlusconi affaccia i commissari Ue: «Facciamo, però solo il presidente» [1] de de
 Nel mirino l'opinione di Bruxelles, c'è il blocco il Consiglio europeo. La ragione, siamo davvero sorpresi.

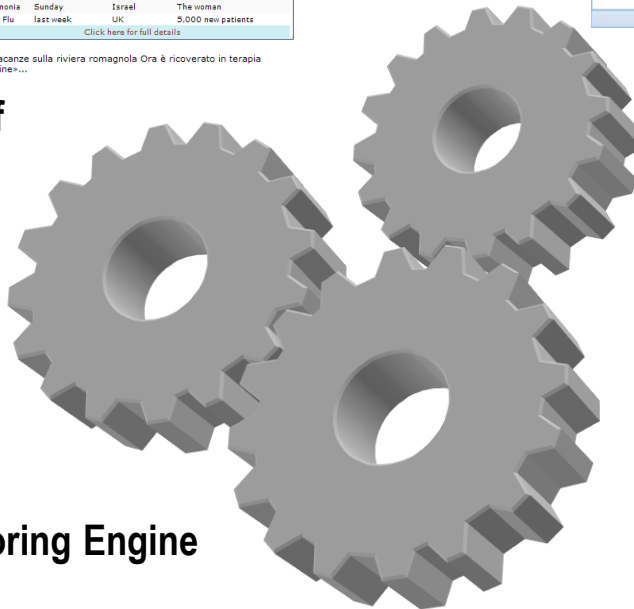
Berlusconi deign Et te bakkeren [1] de de
 De balansa posier Berlusconi heeft gisteren jedgeit de besluitvorming in de Europese Unie te bakkeren als hij veer streekte bijg van de Europese Commissie.

Putin deduce el aniversario de la Guerra Mundial expandiendo a la URSS de sus crímenes [1] de de
 Los actos que se celebraron este martes en Gdansk (Polonia) para conmemorar el inicio de la Segunda Guerra Mundial, han trascendido mercados por los discursos desatados entre Rusia y Polonia, después de que Putin se negase a condenar la actuación de la URSS en la contienda.

Intronati, parli il portavoce di Barroso: "I casi è chiuso, non siamo infelicitati" [1] de de

MediSys Newsbrief

NewsDesk Service (a.k.a. RNS) Editorial Interface



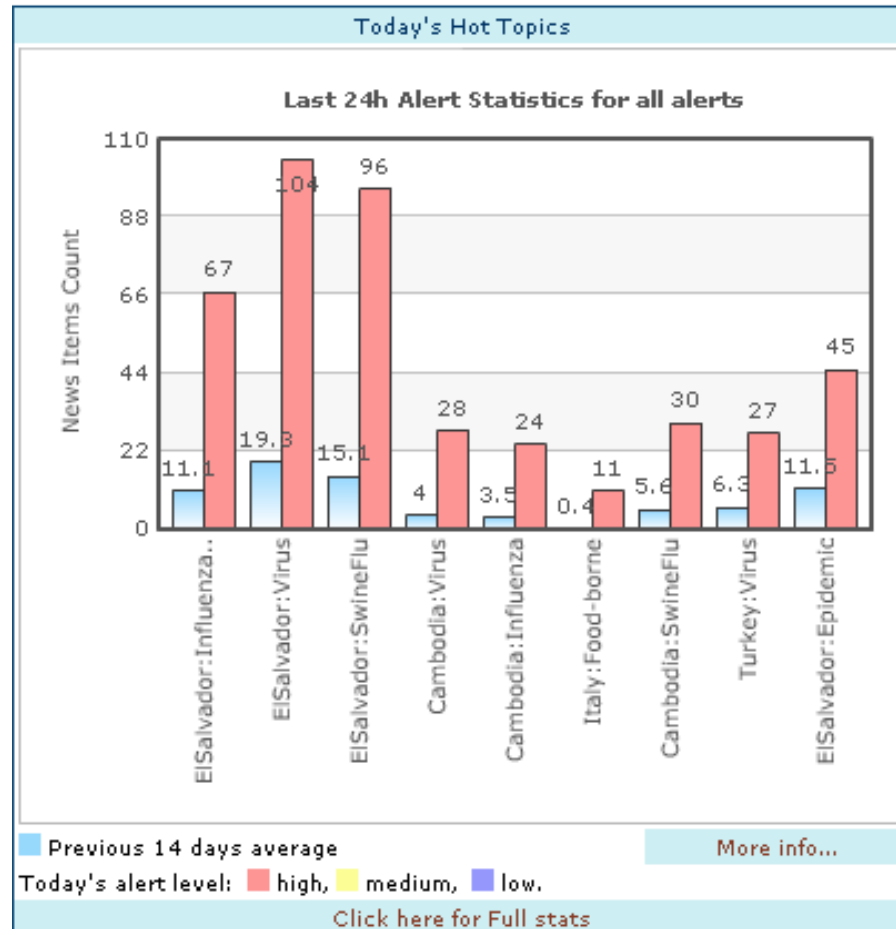
- *Automatic language recognition*
Based on continuously updated language specific frequency tables (Analyser)
Normally part of channel metadata but some channels have multiple languages
- *Automated information/entity extraction*
500.000 persons and organizations based on continuously updated list of entities, many language specific synonyms.
Based on pattern recognition. i.e. no **new** entities detected here.
Allows user-defined entities.
Current (JRC only) entity recognizer produces new set of entities each day.
- *Quote Extraction*
Automatically find and extract quotes from all incoming articles in 12 languages
(currently ar,bg,da,de,en,es,et,fr,it,nl,no,pl,pt,ro,ru,sl,sv,sw,tr)
- *Geotagging*
Based on homegrown harmonised multilingual geo-data set, about 600.000 place name variants in most languages covered by EMM, mostly national capitals, regional capitals and provincial capitals.
Work ongoing to further integrate available resources and increase number of recognised names to 1.500.000

- *Tonality/Sentiment detection*
Simple bag of words approach, range from very negative to very positive, corrected for long term source bias, interesting for following reporting trends per category.
Currently in 5 languages
Ongoing research/development for entity related sentiment detection.
- *Duplicate detection*
Detect duplicates using character trigrams from title and description.
Detect duplicates in clusters using word count vectors based on full text.
- *Powerful Categorization Engine (a.k.a. Alerts)*
Based on user defined keywords/patterns
Boolean combinations, proximity, wildcards
Support for Arabic and similar (automatic noun-prefix processing) Support for Chinese and similar (no whitespace)
Categories can be overlapping, no ontology
- *Metadata categorization (a.k.a. Filters)*
Allows selection of articles based on any previously assigned meta-data.

- *Automated information linking (clustering)*
Incremental topic based clustering and storytracking, geolocation.
Sliding 10 minute interval incremental clustering on last 4 hours of news. (Top Stories on front page)
Bottom up hierarchical average linking clustering
Use simple word/document vectors based on maximum 200 words per item
Eliminate stopwords using frequency tables and document set entropy analysis
Variable distance cut-off criterion based on matrix density
- *Indexing/Search*
Index full text and most metadata. Provide powerful search on all indexed items/metadata
Uses Lucene.
- *Statistics/Trend analysis*
Quantitative analysis of reporting.
Maintain simple count statistics.
Maintain co-occurrence statistics for certain category classes
Detect increase in reporting for categories and category combinations
Build/update language frequency tables
Maintain statistics on category coverage per source

- *Automatic detection of breaking news*
Cluster growth rate
Flux of articles per category
- *Event extraction*
Language independent event grammars used to parse clusters using language dependent resources to fill the grammar slots.
Currently for 5 languages (en, fr, it, pt, ru), violent events, humanitarian events, natural disasters
- *Blog Monitoring*
Monitor selected blogs for posts and comments. Track changes and additions. Perform analysis on hyperlinks found across blogs.
Currently studying the use of twitter feeds

Detect abnormal flux of reporting for a particular country/category combination



EMM Team: 4 full time staff

2002

2004

2006

2009

2010

ODIN

linkr

AMM

ANM

EMM/RNS



Domain specific application

MediSys

First version 2005

News Explorer

EMM System redesign



Redesign based on EMM



RNS redesign
NewsDesk

Continuous development of system and new features

EMM Team: 23 full time staff

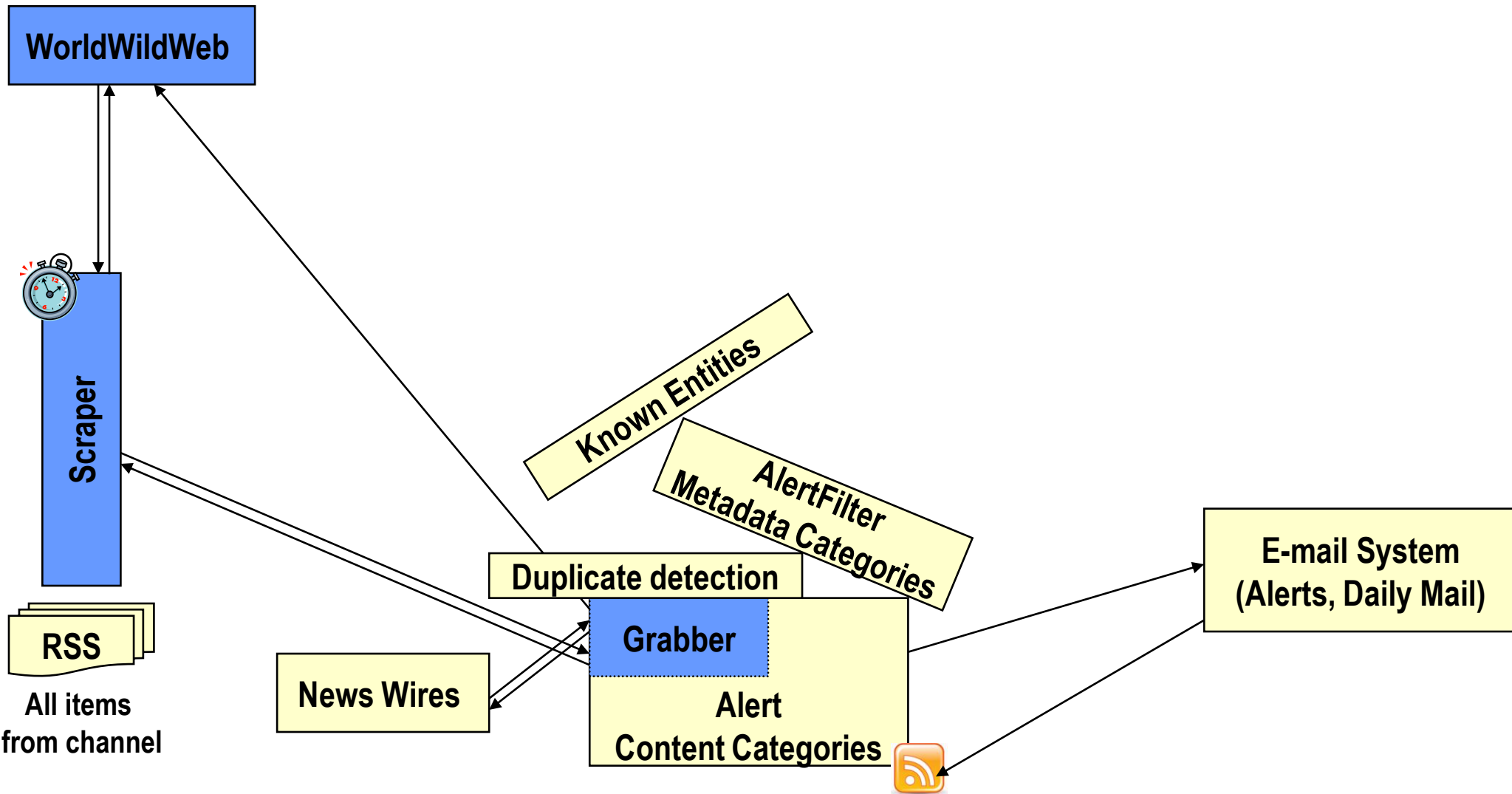
- Monitor main online media
- Monitor news wires from wire providers
- Provide automatic categorization in predefined categories (initially similar to existing newswire categorization)
- Replace and extend the traditional cut and paste activities
- Provide automatic notification (e-mail, sms) of important articles
- Support processing of all relevant languages
- Operate 'near real-time'
- Avoid copyright issues

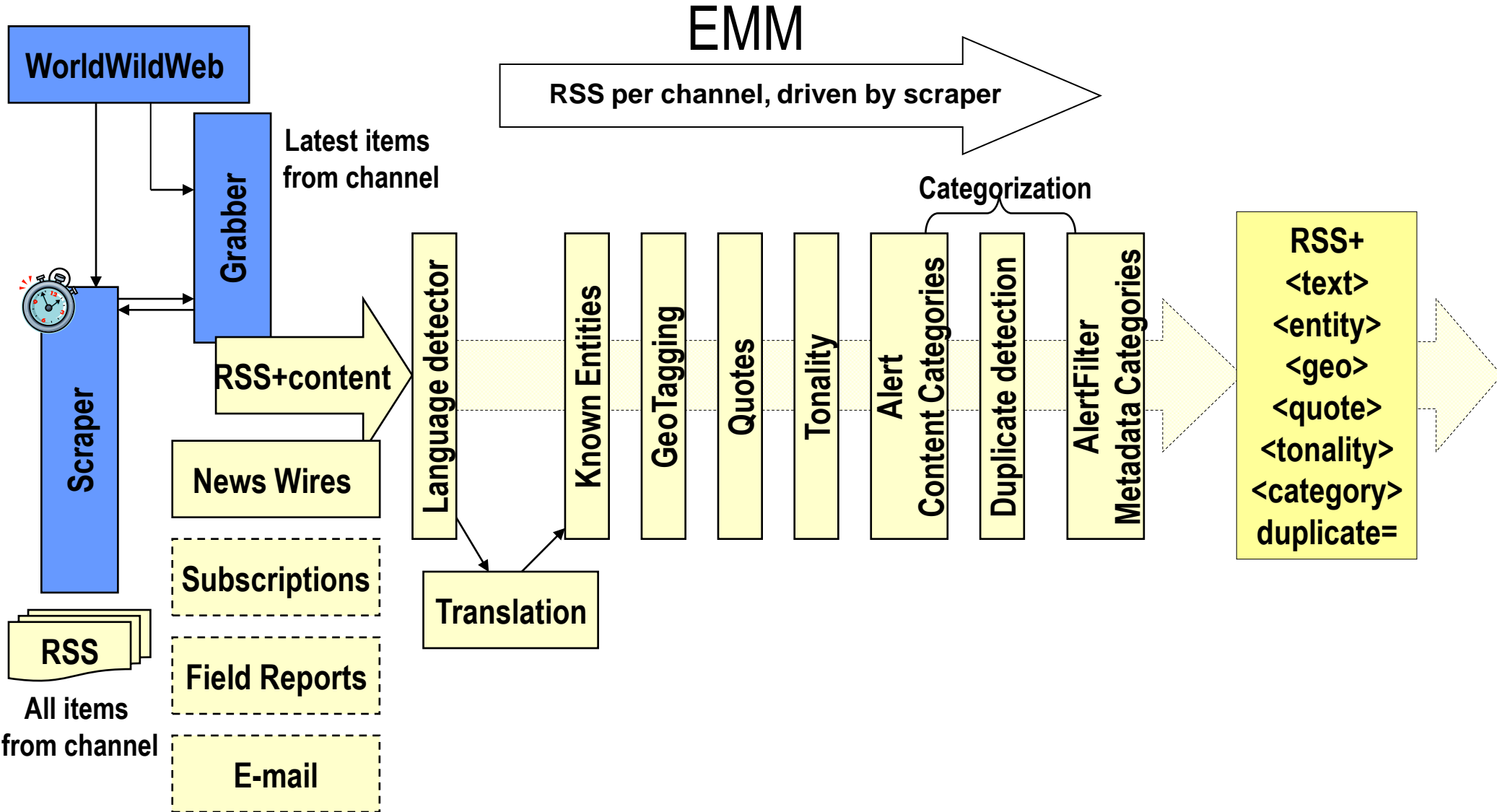
Restrictions: no budget, no people, no hardware

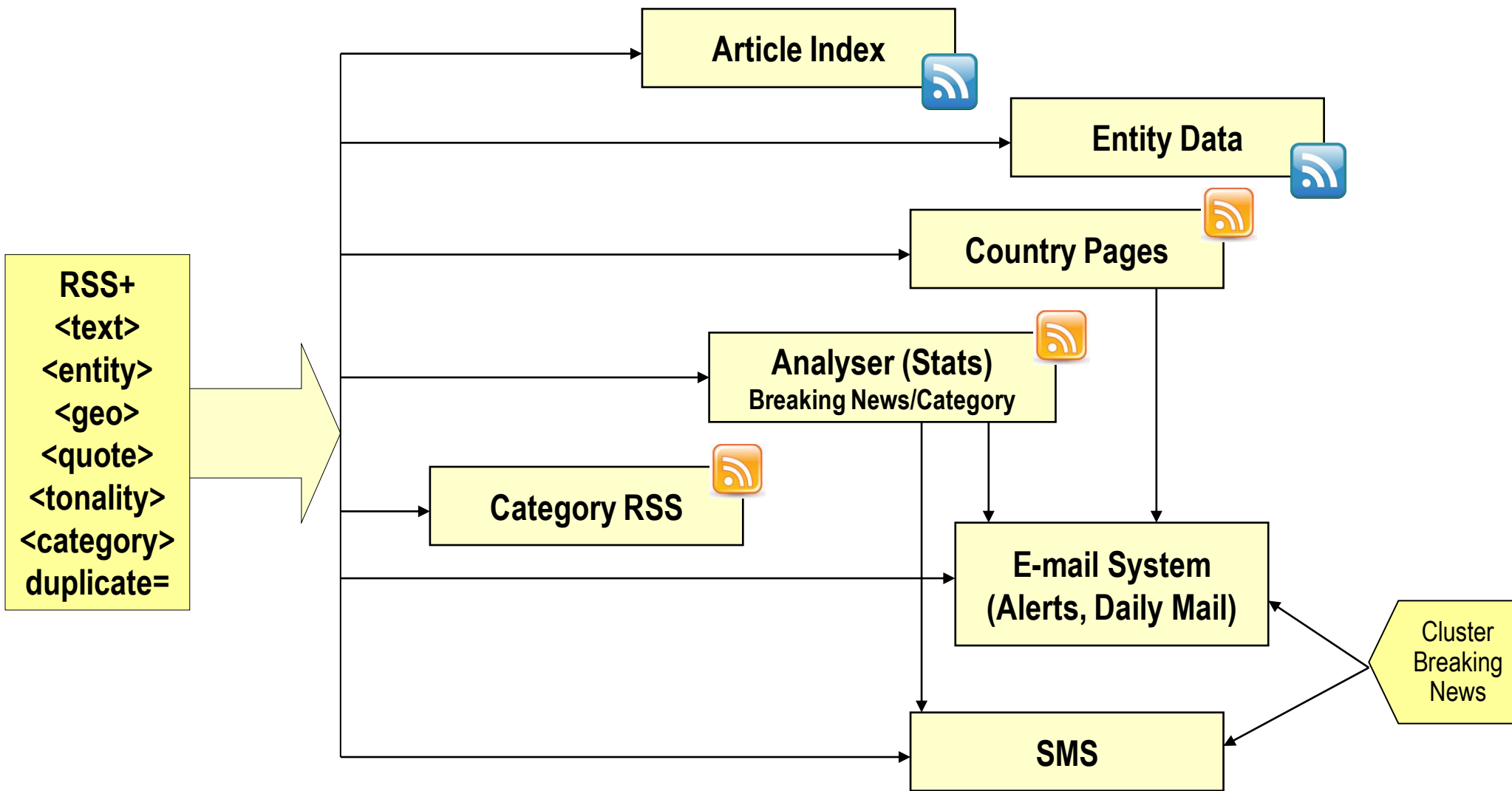
- Good support for languages, web stuff, easy programming + fast development, multithreaded
→ Java
- Modular, communicating processes, scalable, multi-processor/multihost → web services, Tomcat
- Simple data representation, simple communication between processes, flexible → XML, RSS.
Normalize all data on input to RSS
- OS: Microsoft Windows (2000) Server

Develop from scratch, keep software footprint small and performance high.

- *Inter process communication*
Simple HTTP. Sufficient, no overheads (no SOAP abuse)
- *Custom HTML scanner/parser.*
Not much available in 2002 dealing with real word HTML. Different purpose from usual HTML processors, no rendering, 'just' datamodel. Map all incoming to Unicode (UTF-16) Map HTML-entities to Unicode.
- *Main article text extraction.*
Proprietary algorithm (patent pending) based on custom scanner/parser. Universal, no training data needed
- *Custom categorisation system.*
Inspired by parallel finite state technology. Transitions embedded in datamodel. 'Infinitely' scalable. Processing time linear with size of text.
- *Dedicated home-grown mailer/subscription manager*



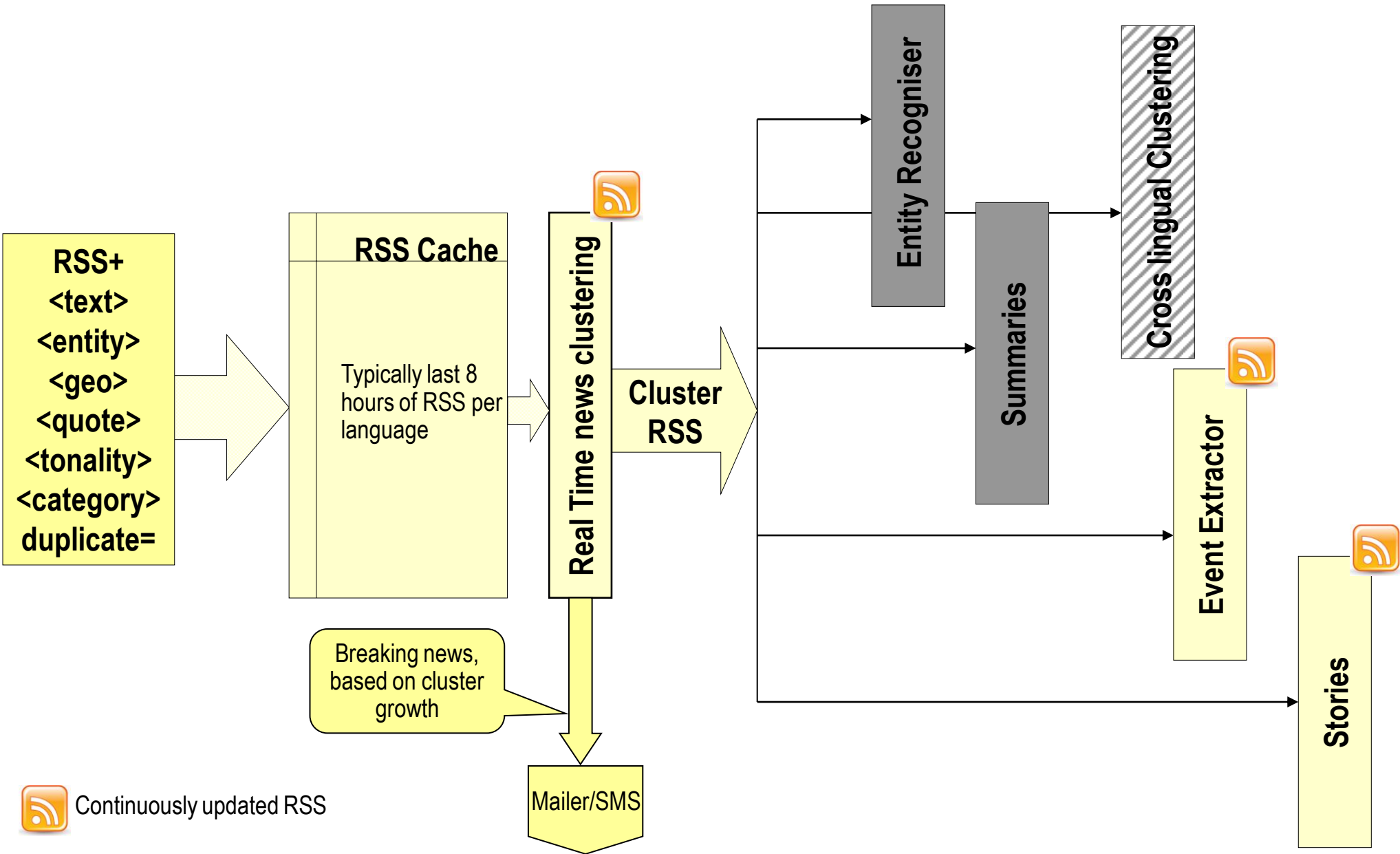


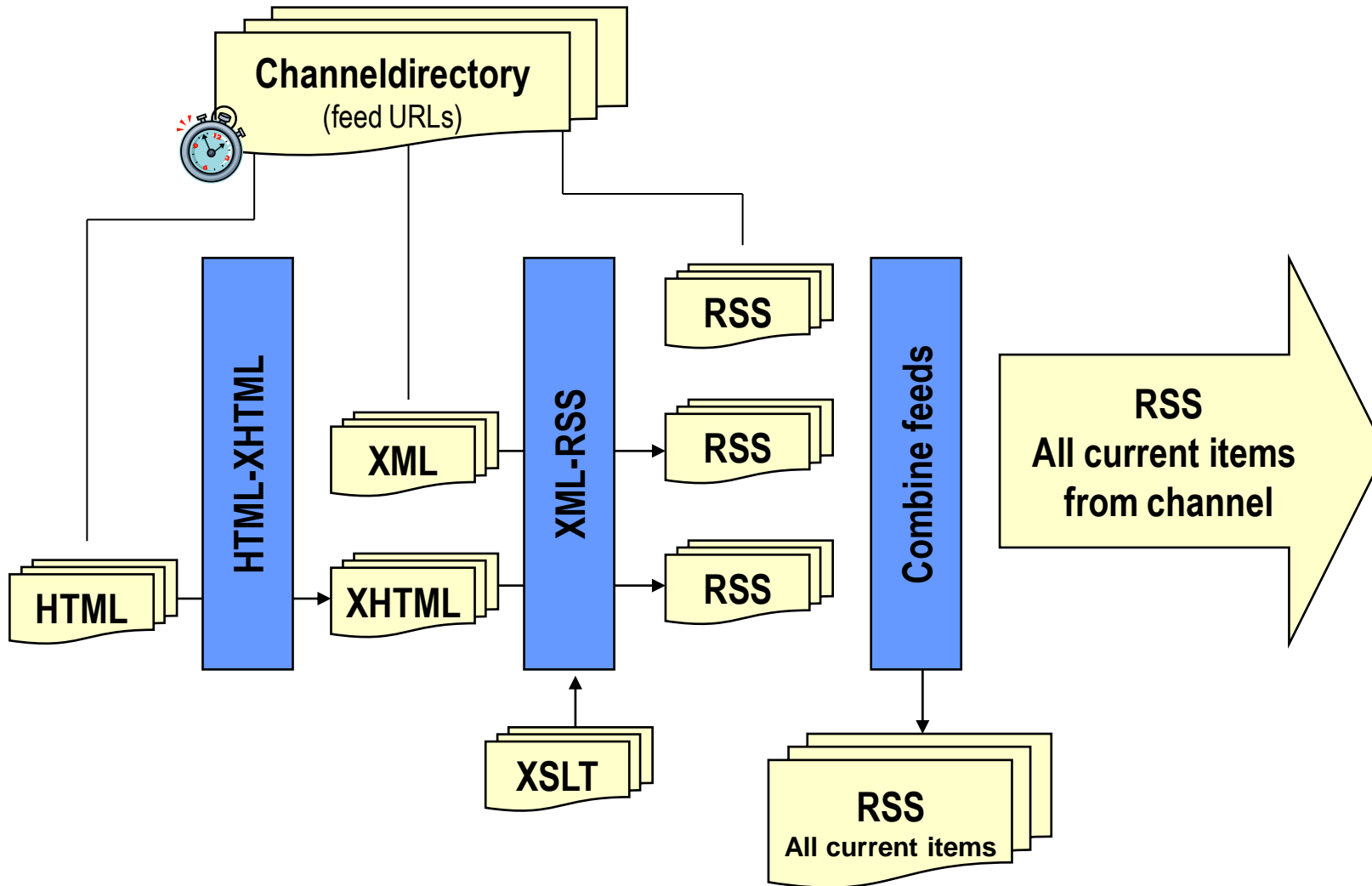


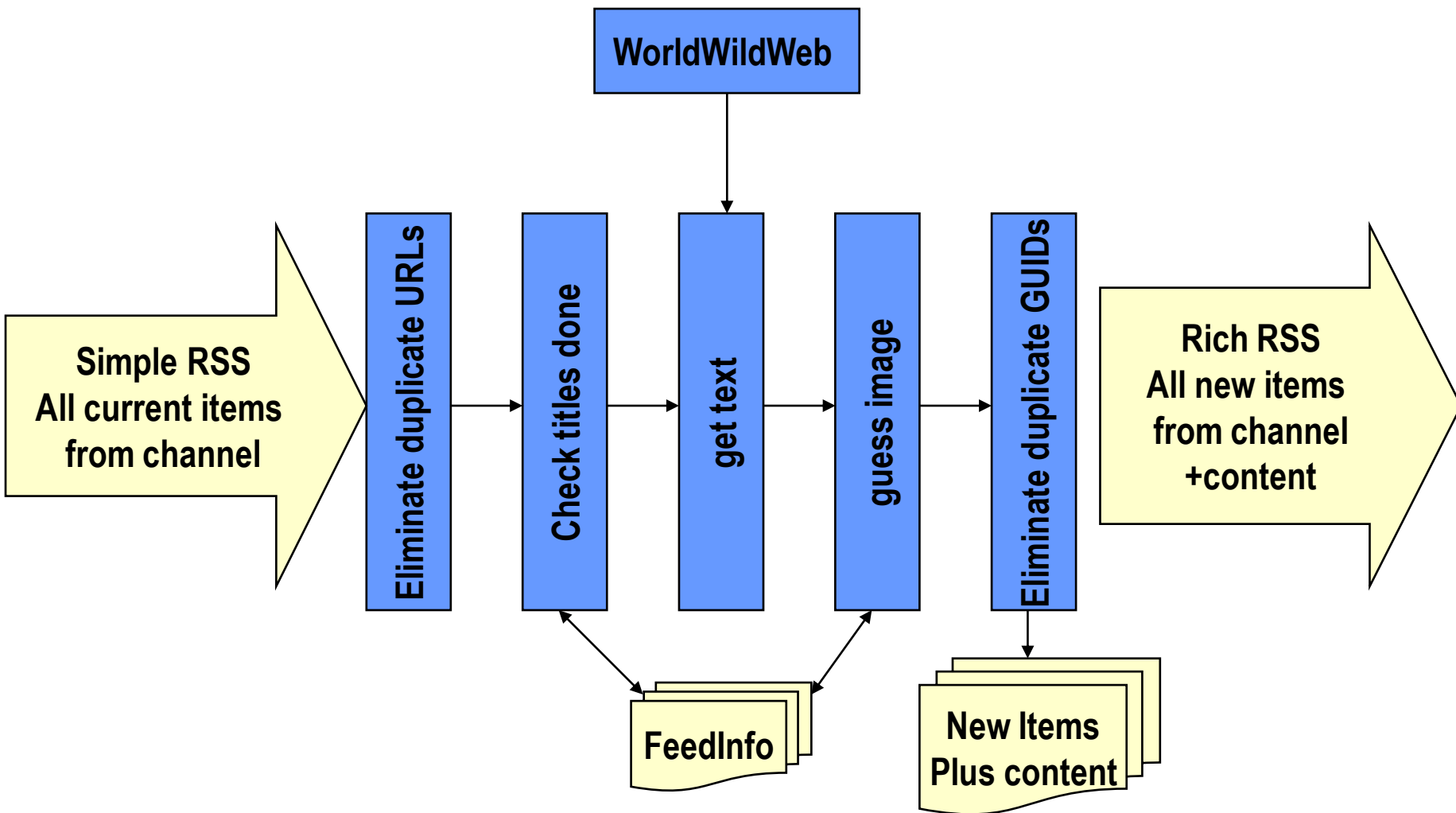
Continuously updated RSS

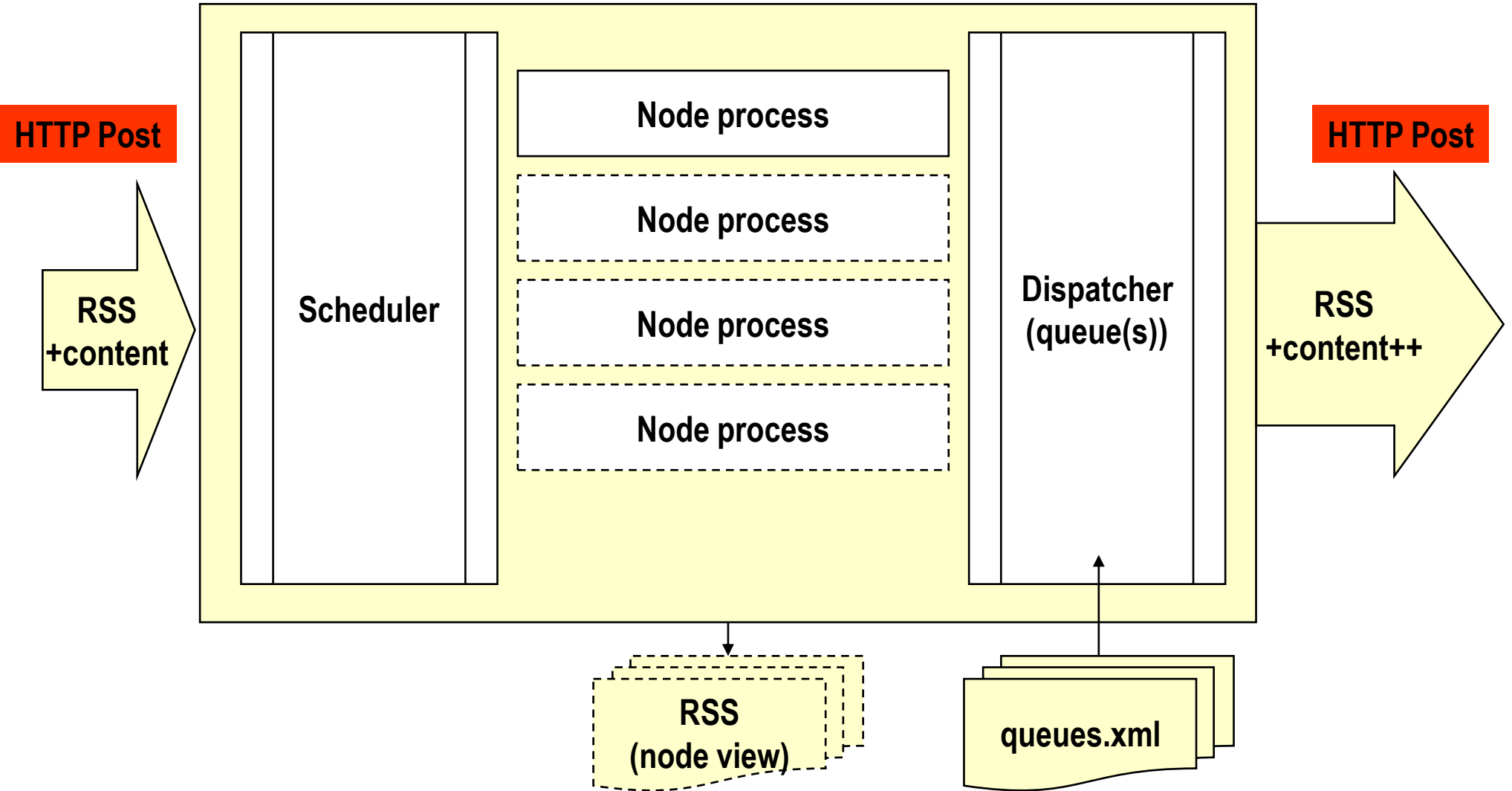


RSS on demand









```
public void doPost(HttpServletRequest request, HttpServletResponse response)
    throws ServletException, IOException
{
    String sId = null;
    String sRSS = null;
    sId = request.getParameter("id");
    sRSS = request.getParameter("xml");

    if (sId != null && !sId.equals(""))
    {
        logger.info("received job for "+sId);
        CachedJob job = CachedJob.makeJob(jobQ, sId, sRSS);
        int nJobsInQ = scheduler.schedule(job);
        response.setStatus(response.SC_OK);
    }
    else
    {
        logger.error("could not process request for id "+sId);
        response.setStatus(response.SC_BAD_REQUEST);
    }
}
```

```
public void run()
{
    while (bRunning)
    {
        CachedJob job = (CachedJob)scheduler.next();
        bRunning = job != null;
        if (bRunning)
        {
            try
            {
                logger.info("processing job "+job.getId());
                RSS rss = rssParser.parse(job.getReader());
                Vector<RSSItem> items = rss.getItems();

                // do something with our items
                // for (RSSItem item:items){}

                dispatcher.send(rss.getGuid(), rss.toString());
                job.delete();
            }
            catch(Exception e)
            {
                logger.error("error processing RSS for "+job.getId(), e);
                job.fail();
            }
        }
    }
}
```

Feel free to browse and use our public sites:

EMM NewsBrief/NewsExplorer

<http://emm.newsbrief.eu>

MediSys

<http://medisys.newsbrief.eu>

If you use any of our feeds on your website, please acknowledge EMM