Europe Media Monitor (EMM)

http://emm.newsbrief.eu

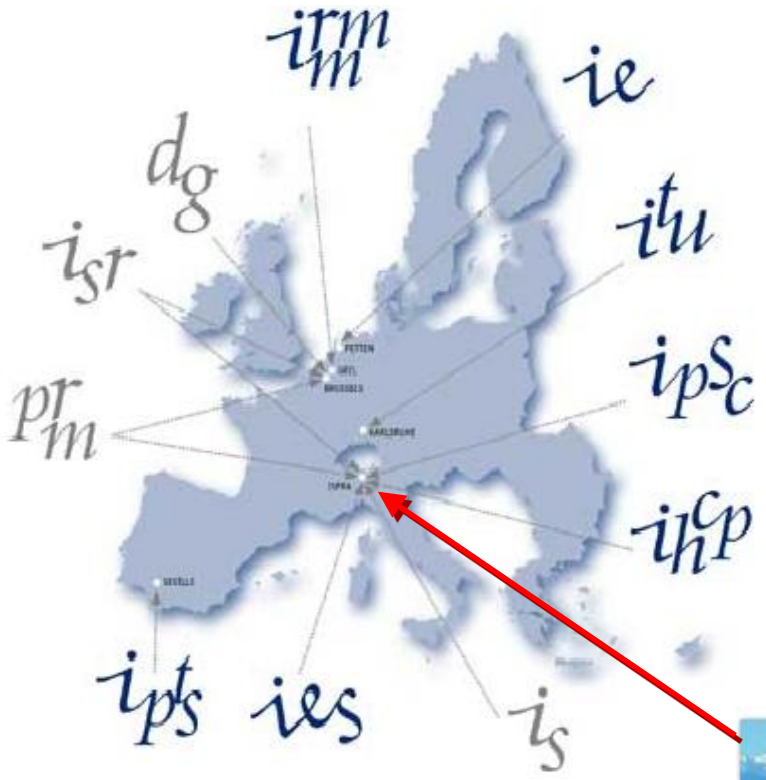
Erik van der Goot & Optima Team

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

Content

- 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

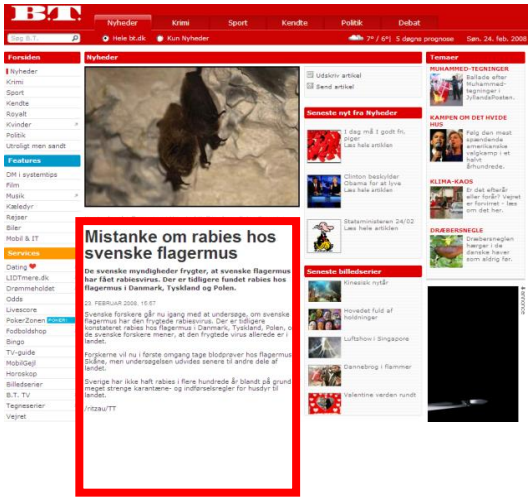
Fast coverage

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

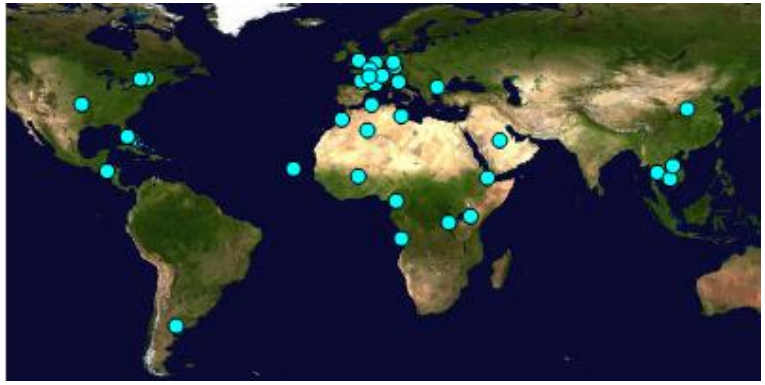
Extract ‘meaningful’ text from article HTML

Importance of multilingual information gathering

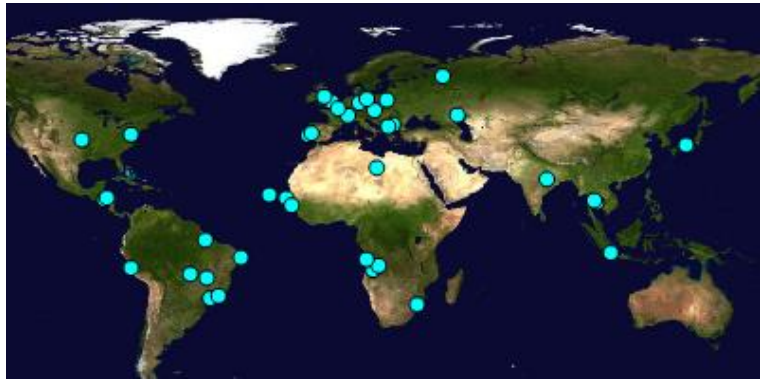
Locations mentioned in MedISys medical articles across languages

Italian - German

English - French



Spanish - Portuguese



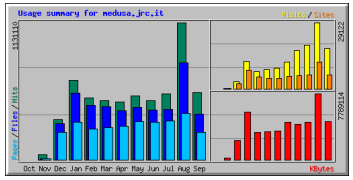
Some EMM statistics

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.

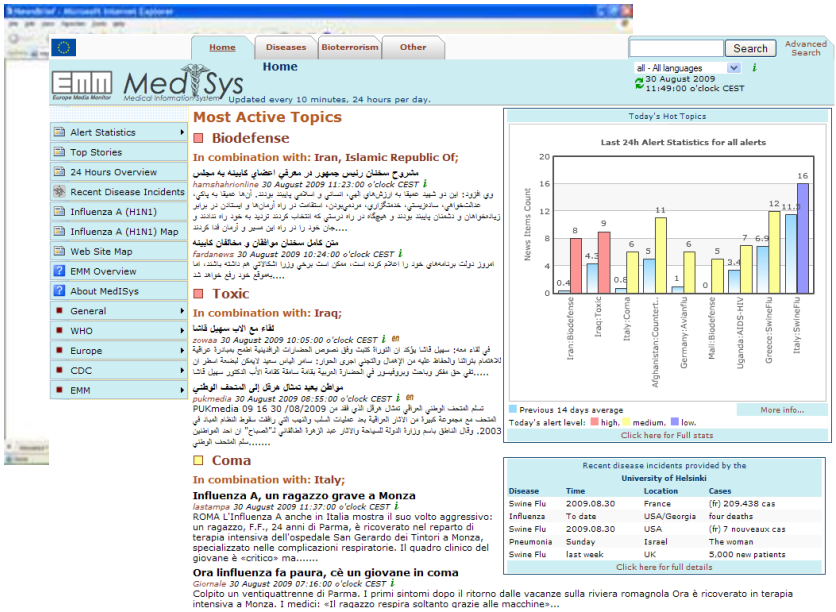
Started in 2002
Up to 1.5 Million hits per day for public websites (EMM NewsBrief, MediSys, NewsExplorer)
But more importantly, 25.000 visitors/day on the public sites

Developed, built, maintained and run by the JRC.

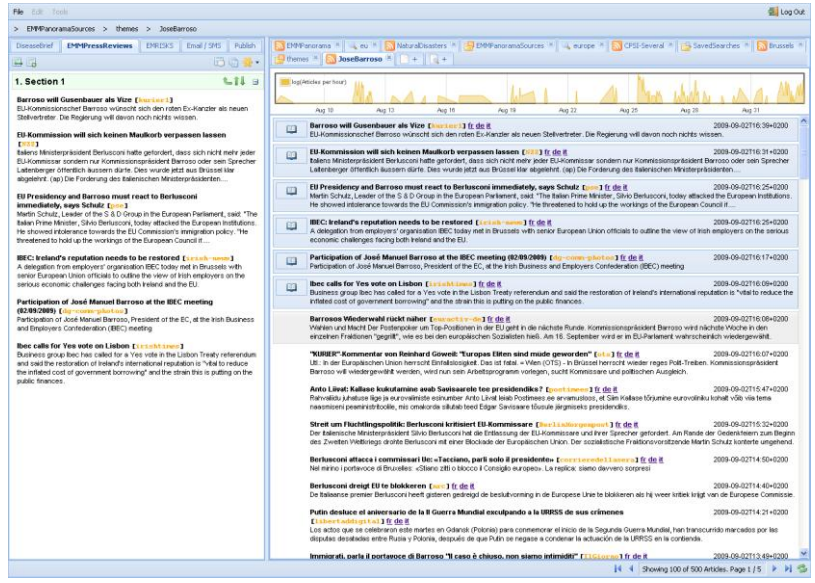


- **EMM**
  - EU Commission Media Monitoring
- **MediSys, monitoring of health related open source information**
  - Collaboration with ECDC,EFSA,DG SANCO, GPHIN, Chinese Health Authorities

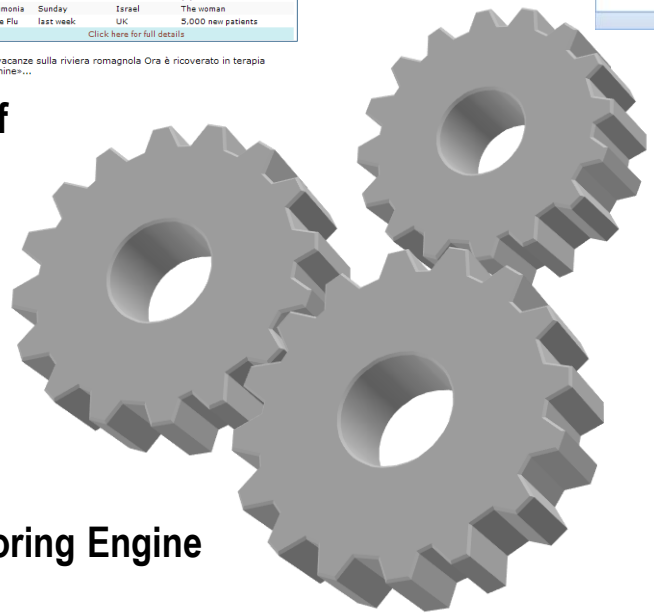
  http://medisys.newsbrief.eu
- **HEWS II, Humanitarian Early Warning**
  - Collaboration with WFP, UNHCR
- **Africa News Monitor (EMM for Africa)**
  - Collaboration with European Parliament and UNDESA (APKN)
- **And an increasing number of customer installations in EU Agencies, various EU and member state institutions, and beyond**



MediSys Newsbrief



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



EMM Open Source Monitoring Engine

- *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

… more features

- *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.

…and more features

- *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

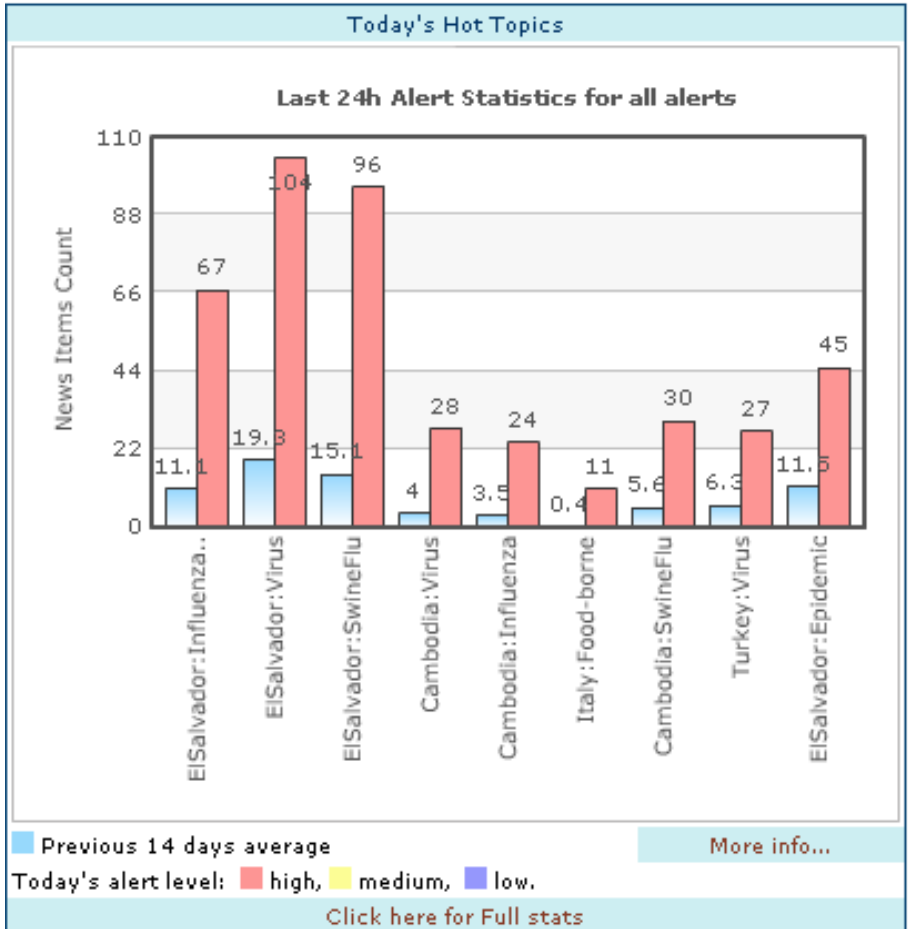
- *Event extraction*

  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



Development time line

EMM Team: 4 full time staff

EMM/RNS

2002

Domain specific application

MediSys

2004

ODIN

News Explorer

First version 2005

Continuous development of system and new features

2006

linkr

EMM System redesign

AMM

Redesign based on EMM

2009

ANM

RNS redesign
NewsDesk

2010

EMM Team: 23 full time staff

Initial Development Requirements

- 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.

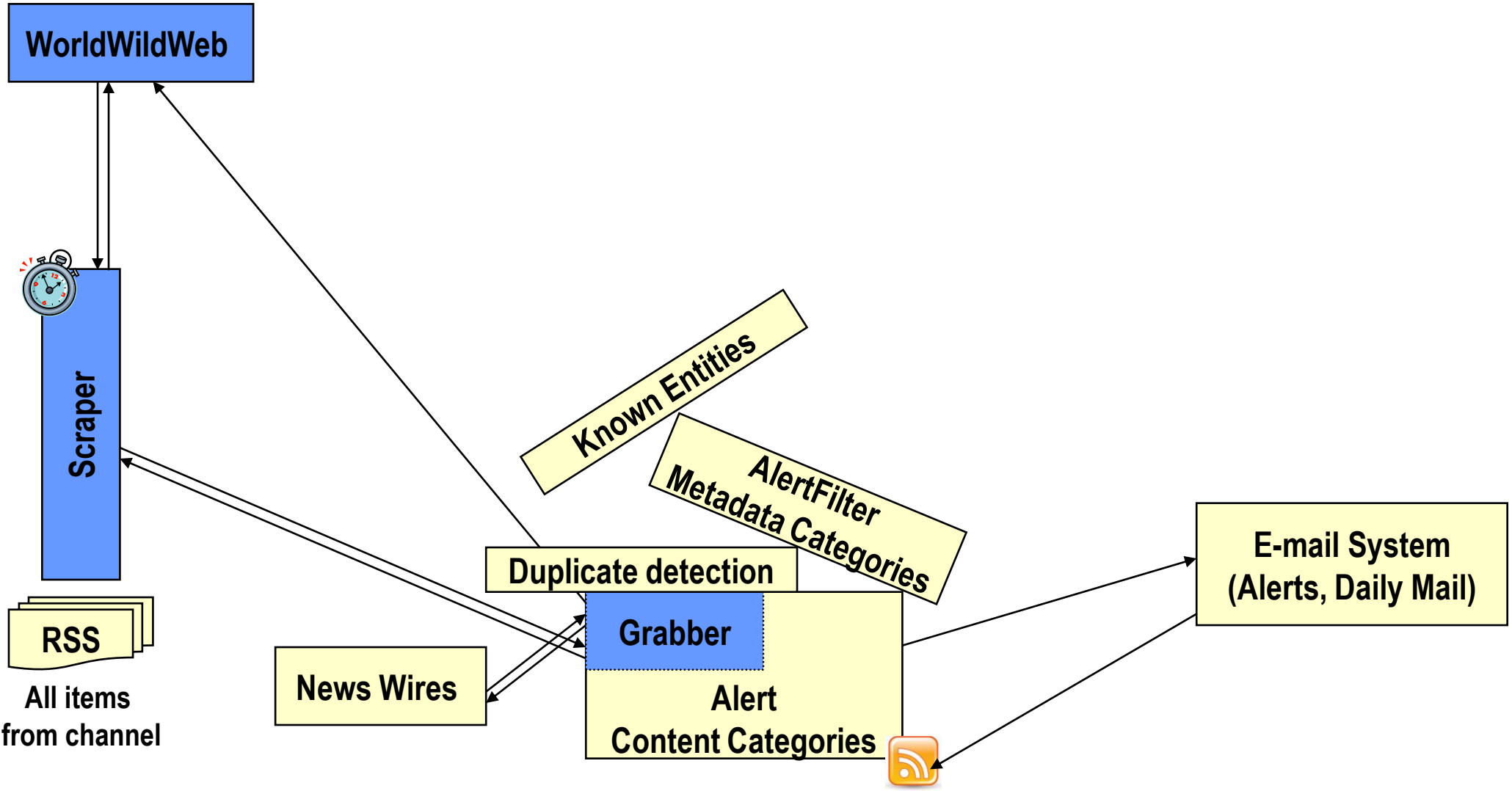
- *Inter process communication*
  Simple HTTP. Sufficient, no overheads (no SOAP abuse)

- *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.

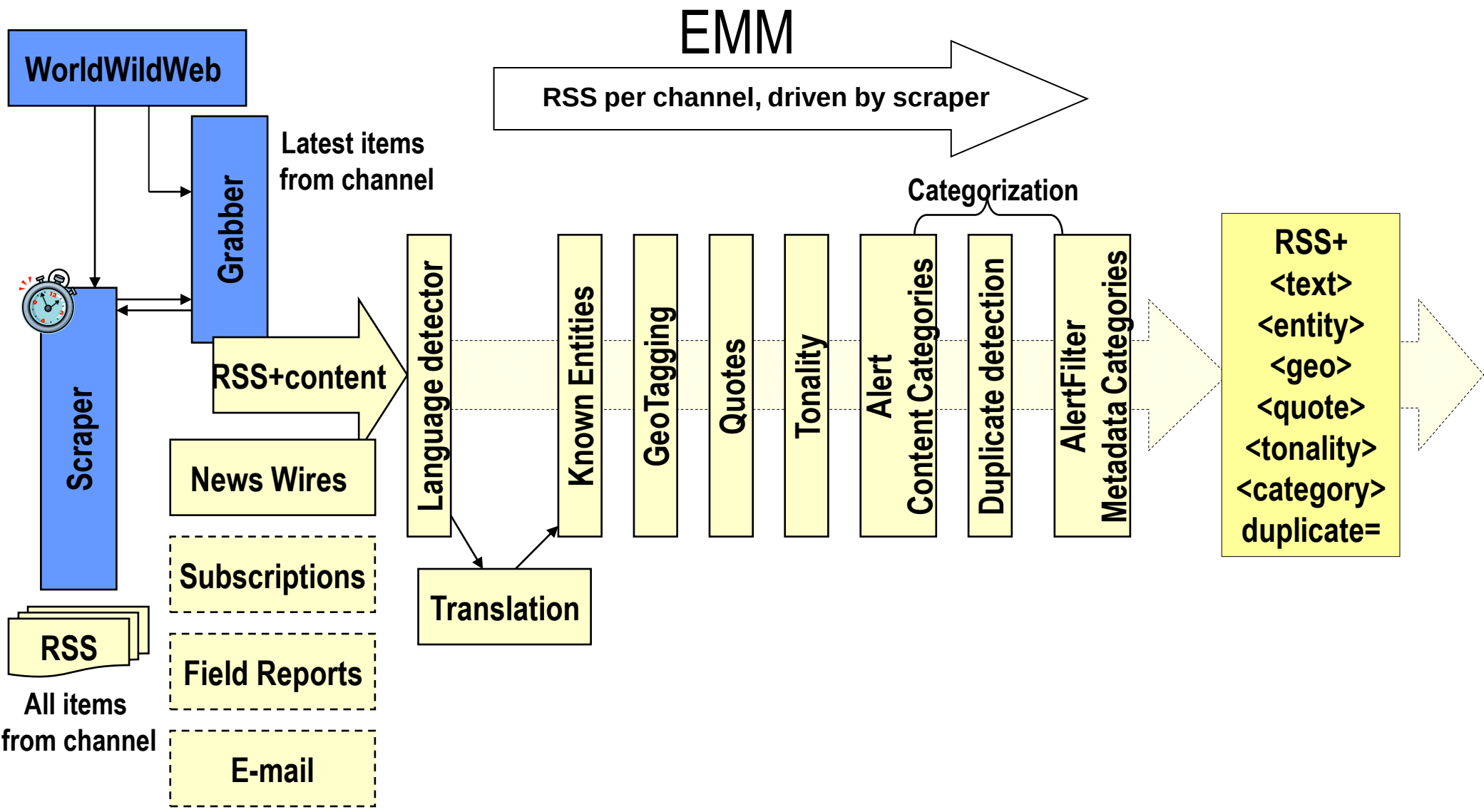
- *Main article text extraction.*
  Proprietary algorithm (patent pending) based on custom scanner/parser. Universal, no training data needed

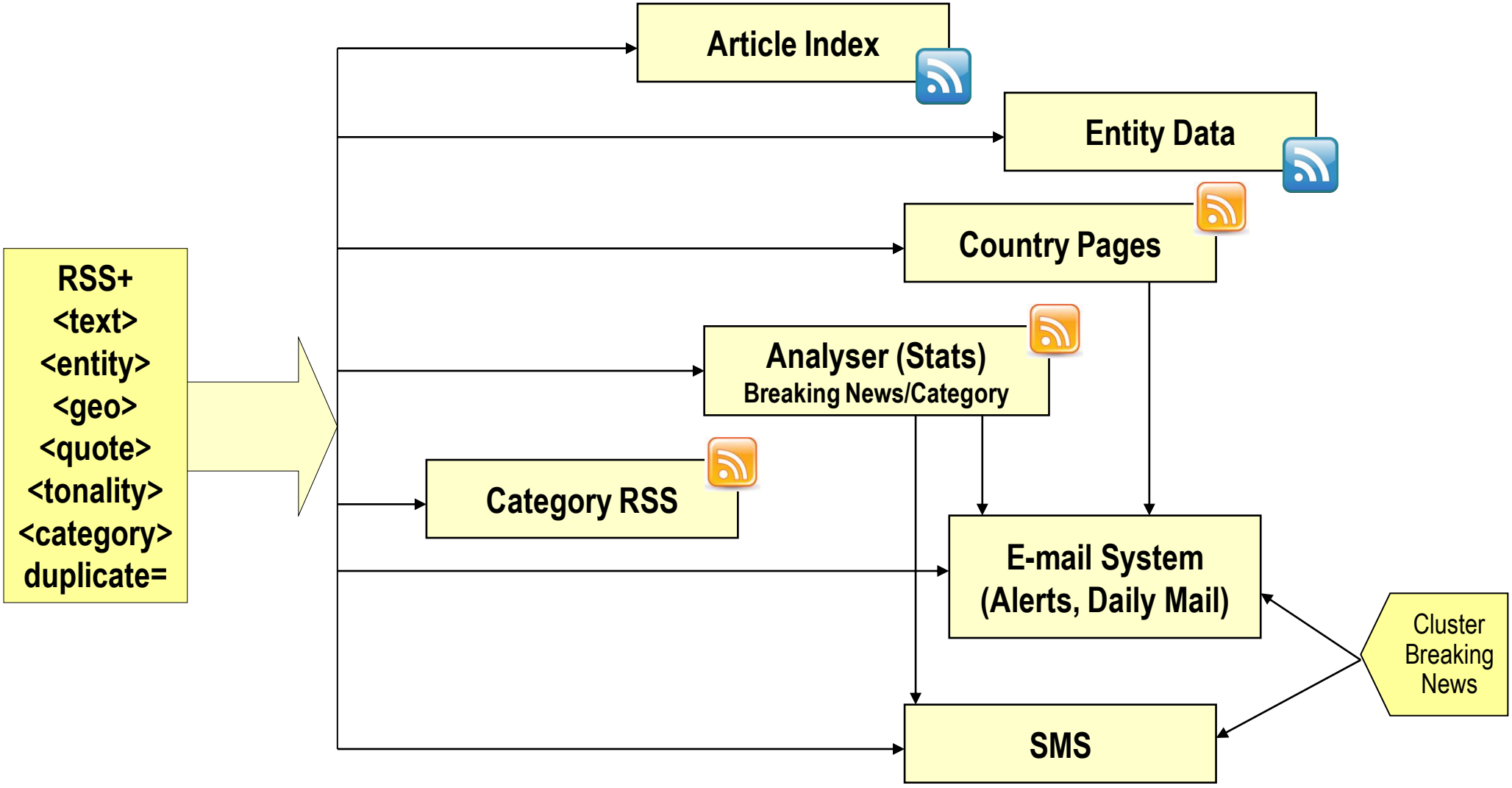
- *Custom categorisation system.*
  Inspired by parallel finite state technology. Transitions embedded in datamodel. 'Infinitely' scalable. Processing time linear with size of text.

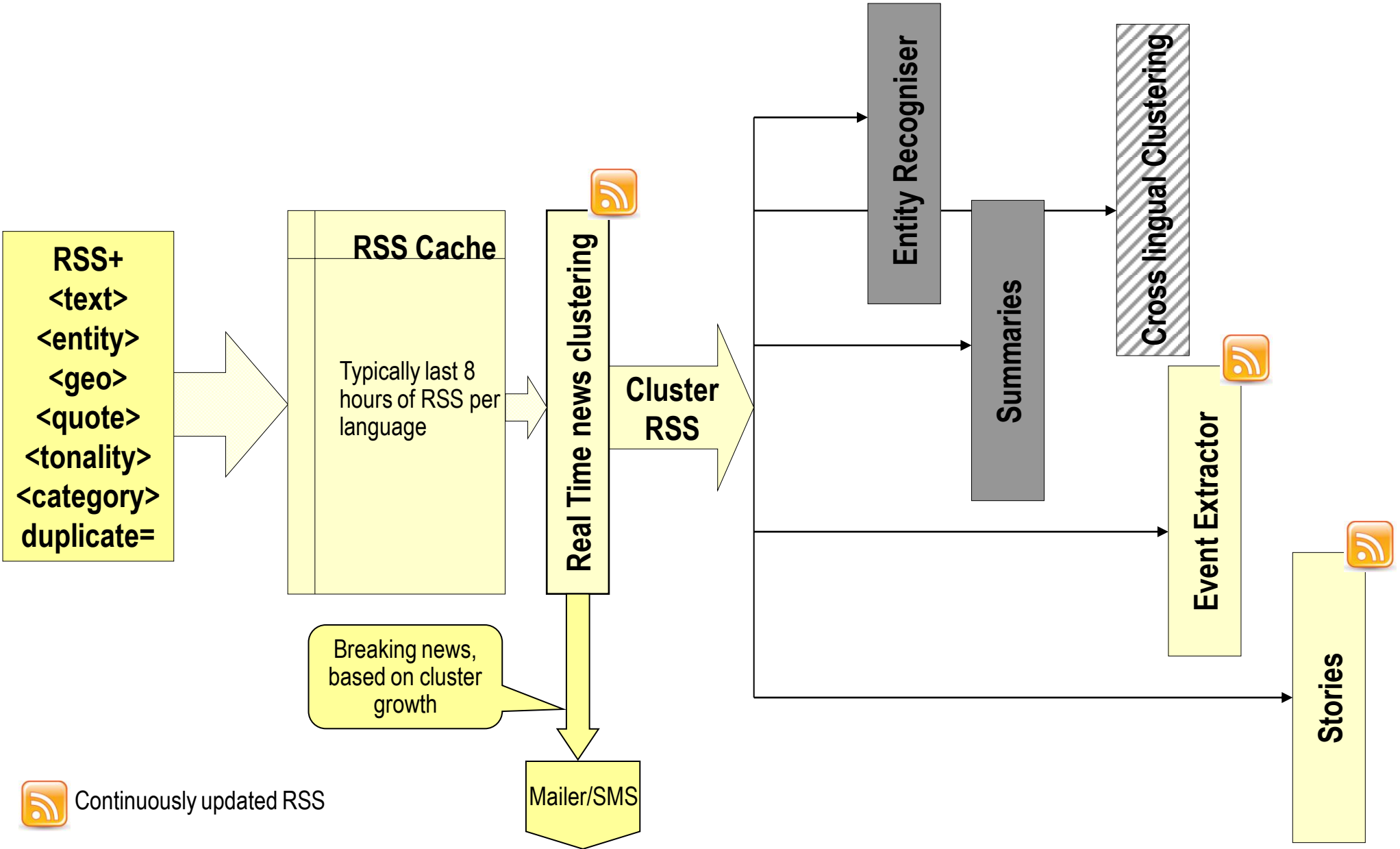
- *Dedicated home-grown mailer/subscription manager*


WorldWildWeb
Scraper
Known Entities
AlertFilter
Metadata Categories
Duplicate detection
Grabber
Alert
Content Categories
E-mail System
(Alerts, Daily Mail)
RSS
All items
from channel
News Wires

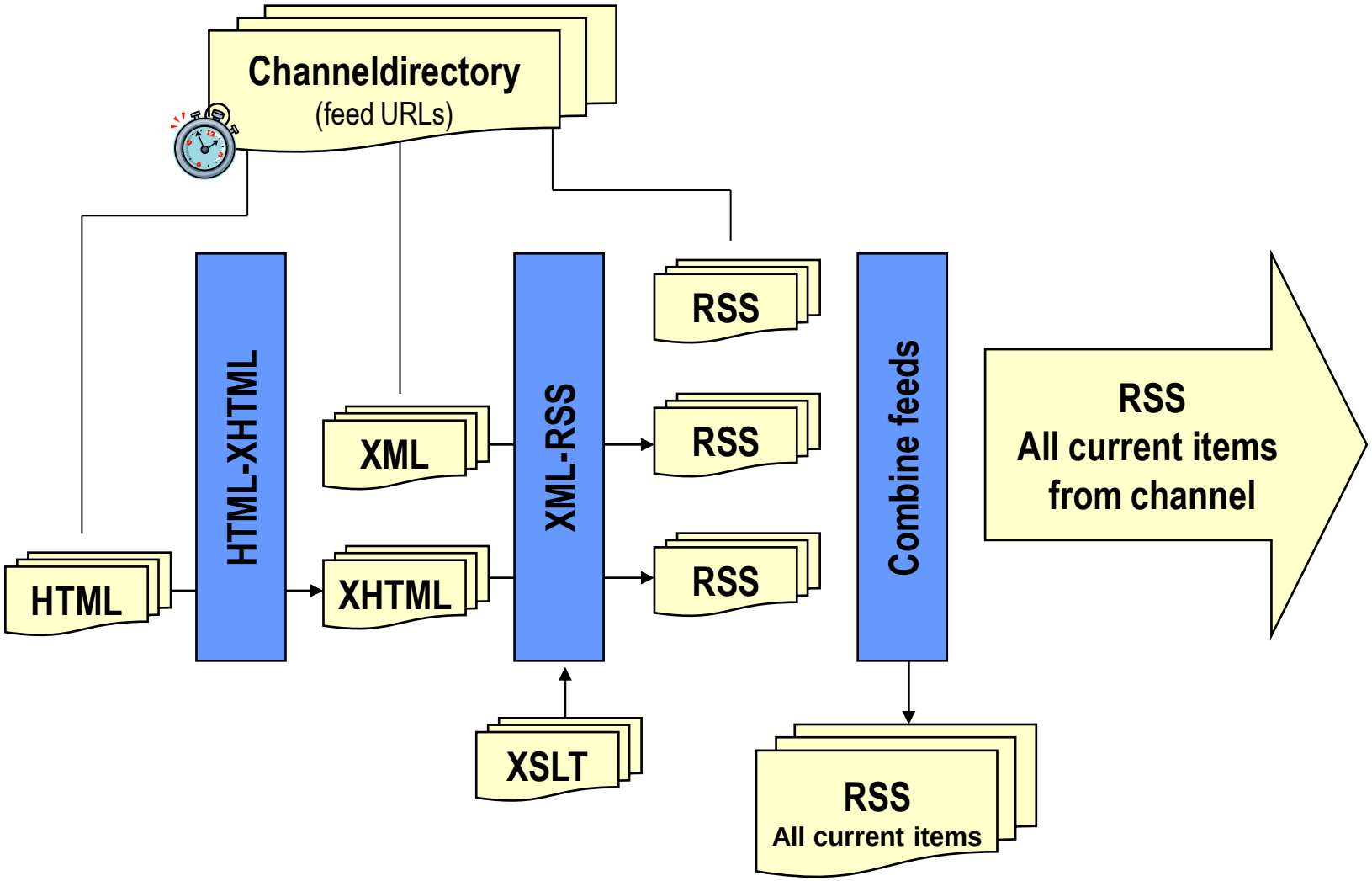
System Architecture: Processing Chain



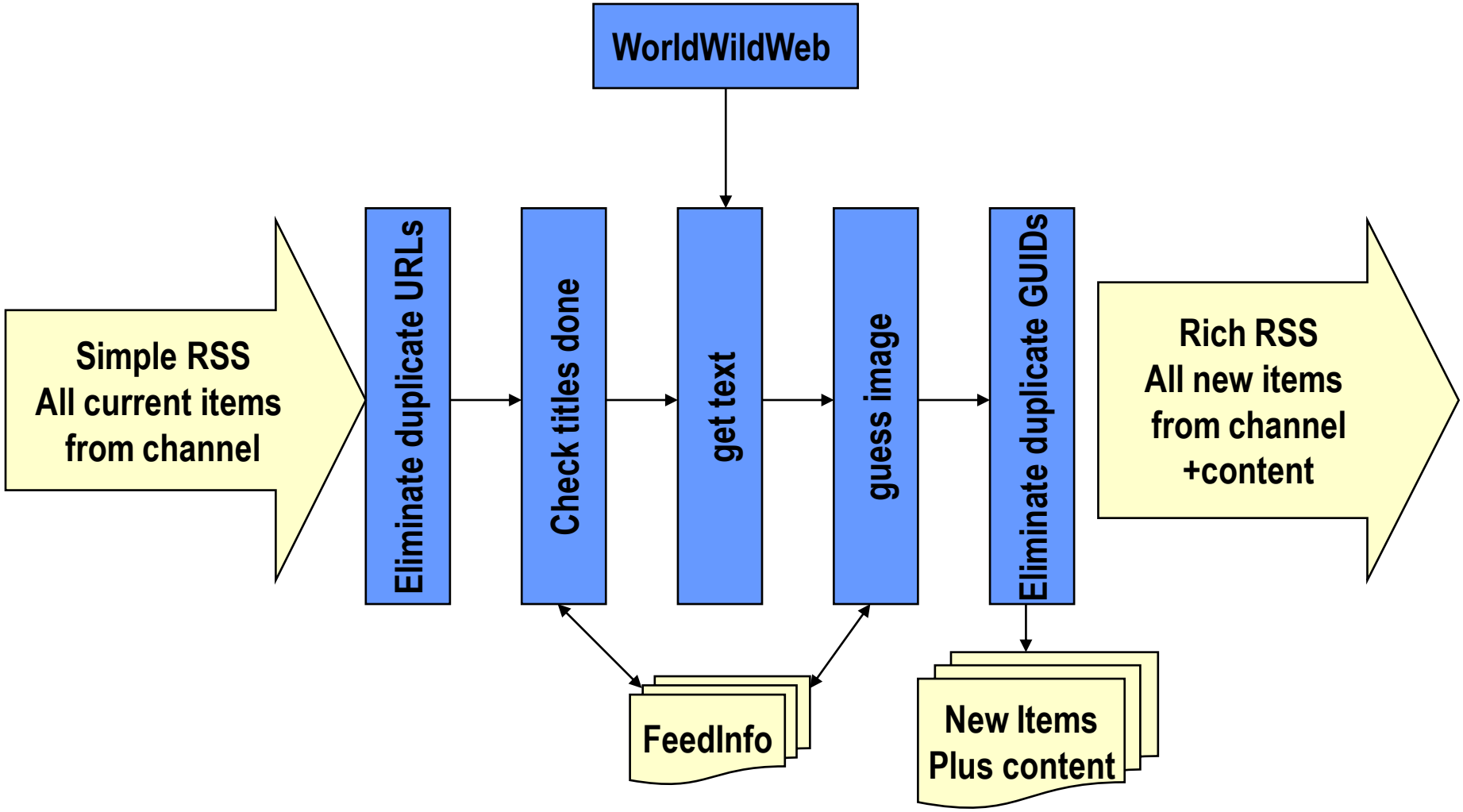

Article Index
Entity Data
Country Pages
RSS+
<text>
<entity>
<geo>
<quote>
<tonality>
<category>
duplicate=
Analyser (Stats)
Breaking News/Category
Category RSS
E-mail System
(Alerts, Daily Mail)
Cluster
Breaking
News
SMS
Continuously updated RSS
RSS on demand

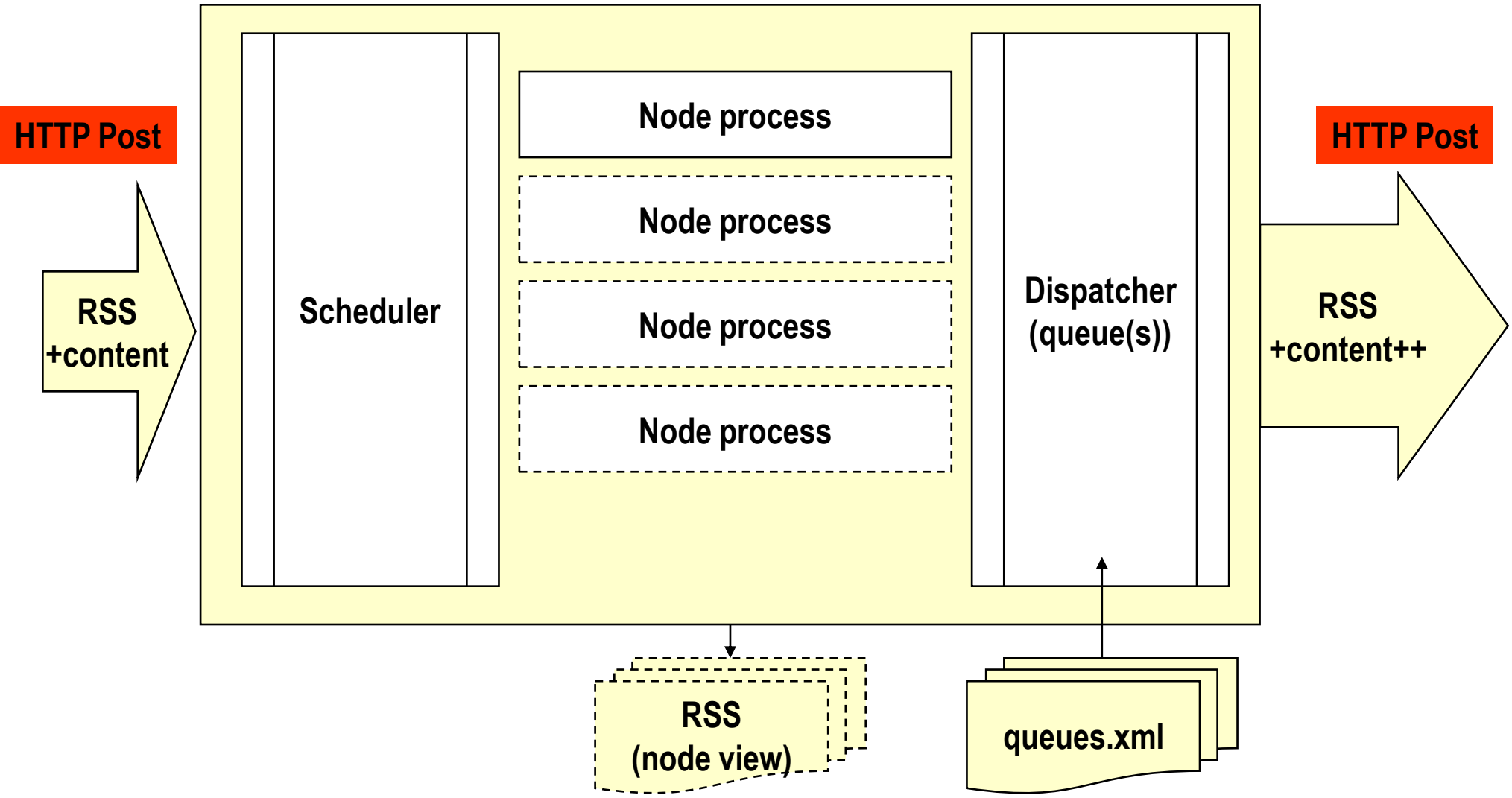

RSS+
<text>
<entity>
<geo>
<quote>
<tonality>
<category>
duplicate=
RSS Cache
Typically last 8 hours of RSS per language
Real Time news clustering
Cluster RSS
Entity Recogniser
Summaries
Cross lingual Clustering
Event Extractor
Stories
Breaking news, based on cluster growth
Mailer/SMS
Continuously updated RSS

Data gathering: Scraper



Data gathering: Grabber




HTTP Post
RSS
+content
Scheduler
Node process
Node process
Node process
Node process
Dispatcher
(queue(s))
HTTP Post
RSS
+content++
RSS
(node view)
queues.xml

Processing the POST


```
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();
       }
     }
    }
}
```

Thank you for your attention

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