USE of SEMANTIC TECHNOLOGIES AT Agence France-Presse (AFP)

Stéphane GUERILLOT

13 place de la Bourse
75002 PARIS, France
Stephane.guerillot@afp.com

Abstract. Adding value to content of various media types and selling and delivering customised content to various types of customers can be the definition of any news agencies' business today. But the insertion of taxonomies in the information workflow is a challenge that needs prototyping and experimental phases because of the impact on the production cycle itself.
With the increasing pressure of real-time information, this is a practical example of semantic web applications in an "adapt or die" context.

1. Introduction

A Long Tradition of Newsgathering

AFP is the world's oldest established news agency, founded in 1835 by Charles-Louis Havas, the father of global journalism.
Today, the agency continues to expand its operations worldwide, reaching millions of individuals via thousands of subscribers such as radios, televisions, newspapers, administrations… from its main headquarters in Paris and regional centers in Washington, Hong Kong, Nicosia and Montevideo. All share the same goal: to guarantee top quality international service tailored to the specific needs of clients in each region.

AFP provides a 24/7 worldwide multilingual news coverage and produces daily an average of 2 millions words, 1000 photos, 50 news graphics, as well as video reports and multimedia products for the Web. French, English, German, Spanish, Portuguese and Arabic are the languages used.

A news agency is to report about the news, the facts with speed and reliability. *Accuracy* is the guiding principle. It has to get to the heart of the issues and to provide coverage and analysis of world events. As well as general news from around the

world, AFP is offering economic and financial news, sports coverage, human interest stories, celebrity news, science, culture, new technology, lifestyle and offbeat items.

To do so, AFP is relying on 1.200 reporters, 200 photo-reporters and 2.000 stringers based in 165 countries.

2. The Challenge

With the Web development, news agencies are no longer THE player for news and information management. As a direct impact on our day to day production cycle, our customers as well as our own journalists are expecting more and more functionalities and help to work with the huge pressure of the information, its volume and its continuous flow.

It is always on and those "users" can easily be submerged by the amount of information to process and manage. As an example, our French clients are receiving more than 1200 news items a day and between 500 to 800 digital pictures. Our Image Forum online database is offering more than 7 million pictures plus an even wider selection of the daily production.

The characteristics of the search engines on the Internet can lead us to believe that information search and accurate match are now possible in an efficient way and that full text search tools are no longer the ultimate solution in our business environment.

Our Clients (encompasses the internal clients – our journalists – and our subscribers) have several access tools to the information including Web platforms and News Editorial systems and their requirement is going *beyond the language barrier* and the information *nature*.

They want to:

- Subscribe to selections by themes,
- Browse quickly and efficiently through a large corpus of multimedia documents, including archives.

Our main goal is to provide, within AFP, the solutions within its production cycle and help for the implementation of the changes required to offer those potentials to our Clients via our next generation of Multimedia Editorial System.

3. Search and Filters

In the news business, the search and implementation of real-time filters are based upon concepts and topics.

The selection is made on information of a known nature (Text, Picture…) and precise categories which can be taken from a list of Subject Matters (such as *alpine skiing* or *cinema*) or Named Entities of different classes (*City, Person's Name*). Sometimes it could also be related to a specific Genre of information such as *magazine* or *obituaries*.

So Nature, Subject and Genre are the main criteria for selection of a news item.

We should then offer:

- On our real-time and archive platforms the possibility to quickly find information regarding the Vatican, Benoit XVI, Ségolène Royal, Tom Cruise, the Oscar 2006 awards as easily as the information related to football or bio technologies;
- To easily browse between concepts as per example between Antoine Deneriaz, the Subject detail of alpine skiing and an Event such as Turino 2006;
- To allow our Clients to simply create either on our platforms or even in their own Information System, the search interface and alert filters based on the news items metadata as made available by AFP.

4. What is available today?

A news agency is a Factory for News and the Journalists are, as in many cases under pressure to release their production as quickly as possible and accurately. They are willing to enrich their reports if the tools made available for them are efficient, simple to use and do not interfere too much with the constant request made to them to beat the clock.

Our content specificities are essential when it relates to mark-up and enrichment.

1. AFP is working from production to distribution in 5 different languages. All the production, in every language, is made available to each and every desk. They constantly are exchanging information, asking for details, explanations or follow-up.
   It is then essential to provide them with search and selection tools operating through multiple languages at a same time.
2. Most of the news items are related to subject matters (such as *Politics*, *Art*, specific *Sports*) or persons (G.W. Bush…) or events (European summit) or organisations (Political party, Microsoft…) or "products" such as *Da Vinci Code*.
   The news items are often short and concise. The main issue is mentioned in the first two paragraphs which means the first 100 words. Then the rest of the news item is more about the context or background information.
3. AFP is producing between 5 and 10.000 documents per day. During the peak hours we could have one document validated every second.

Production:

Up to now, the “production” is organised by Nature. We have multiple production lines running in parallel. For example, Text and Photo are managed and processed with their own systems from the journalist (photographer), the keyboard and the camera, to the desk.

The main driving line is then concentrated on what is called the *slug* or *slug line*. It is limited to 24 or 64 characters (depending on the distribution data format) and contains a set of keywords including named entities and controlled vocabulary taken from standard lists. This set is adapted permanently to follow the news focus.

For the Text production line, the slug line can be specific to one desk and or a language and or the desk specialisation (such as sport, business…). Because of this required flexibility it is impossible to use them as a universal and reliable reference in the general news domain but only in the sport and the economy and finance production lines.

On the Economic & Finance desk, we have also included an automatic parser to search for company or organisations names appearing in the news item to generate the ISIN code and by that make the Client search or indexing more reliable and accurate.

<Slug> Japan-IT-camera-company-Canon <Slug>

<Title> Canon latest to pull out of film cameras </Title>

TOKYO, May 25, 2006 (AFP) - Canon Inc. said Thursday it would stop developing film cameras, joining a growing number of high-tech firms pulling out of the sector as digital cameras take over.

"The situation is very difficult for new (film-based) cameras," Canon president Tsuneji Uchida told Jiji Press in an interview.

The announcement by Japan's largest digital camera maker followed similar exits from film cameras by rivals Nikon and ailing Konica Minolta.

…

hih/sct/mtp

Our Text Editorial systems then provide with semi-automatic categorisation based on the words entered in the *slug line*.

Our Photo Editorial system is also providing for additional manual categorisation to bring consistency with the slug used in the Text services plus the inclusion of person’s names, locations, precise subject matters or details and keywords taken from our AFP taxonomy for Images.

In both cases, we are using the IPTC News Code taxonomy (hierarchical with 3 levels).

Distribution:

AFP news services are either distributed as complete services in a define language (managed by one editorial desk) or as a selection of news item sorted by one of our filtering tool.

This last tool is working from concepts (topics) that are defined with the words used in our unstructured text and stored in a reference database. They are then used against each news item validated in our production line and are nearly applicable across any defined language (but Arabic).

To show the limits of such a solution, the continuous evolution of our *slug* words can lead to complex situations: If "Katarina" is used in a slug line during the hurricane reports in 2005, how should it be considered in the future? Is it then becoming a generic term or a synonym?

When made available to our Clients, they want, from their interface, to apply automatic filtering for alerts and to ease their work. Imagine how you will deal with thousands of news items arriving in your mail box on a daily basis. With a standard full-text searching tool you might find too much hits and even worse miss some important ones.

When applied on Photo services and because of the caption writing style (often reduced to one paragraph and less than 150 words) this is critical.

If you add that the end user's interface is most of the time offering the multi-criteria or advanced search as an option and that the Boolean logic or operators are not well managed by the human being behind the screen, you would understand the value of enriched content applied to News.

5. Our project

The classification and enriching services enable the association of Metadata to the documents.

Some of the information is filled through assisted input (scroll, combo-box) by the user, other information can be assumed from the content.

These services depend exclusively on the document's content and are independent of the editorial platform; they should be accessible through the network and can be called by the users whether they are producers, editors or archivist.

To allow the users to work offline, the necessary tables/data to run these services locally are downloaded on the editor's workstation.

The classification and enriching services are the same for the production and editorial functions.

They must comply with the timely constraints that are part of the user's tasks/job.

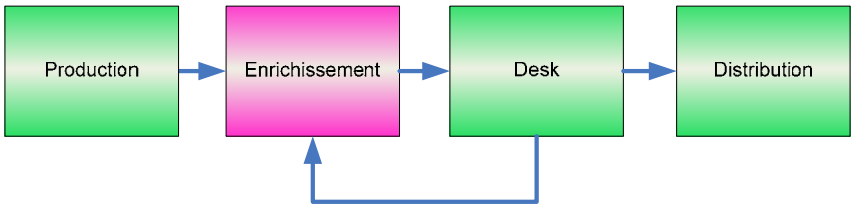
Automatic indexation

To answer the need of search be concepts in various technical environment, within AFP platforms as well as at our Client's, we have concluded that we had to enrich the information at the desk level.

It is this information which is later on in the distribution cycle made available to our Clients, on line, on the Web or within their professional applications.

Several projects were studied and one prototype is based on a purely automatic enrichment. Those enriched documents can then be sent to our selection engine with an acceptable percentage of errors.

If we later on wish to refine this process we shall have to include in the loop a control by the journalists and, ideally this would be done without delay during the validation process.



Learning process

The initial project is also based on the use of a selection of less than 100 elements from the 1300 terms already defined in the IPTC taxonomy – News Codes.

For each category a set of carefully selected news items in a semi-automatic mode. They will be used as references or learning corpus by the system to assign one or multiple categories by using similarities.

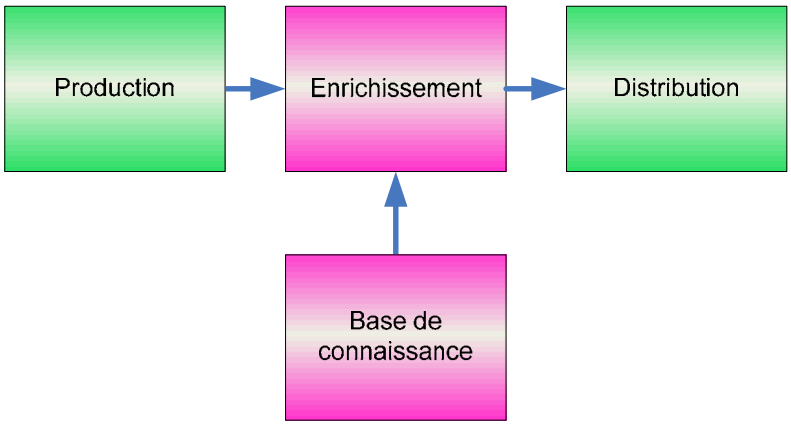
A post processing refinement using additional rules could be added for some categories when the first process is not conclusive.

The benefits of a Knowledge base

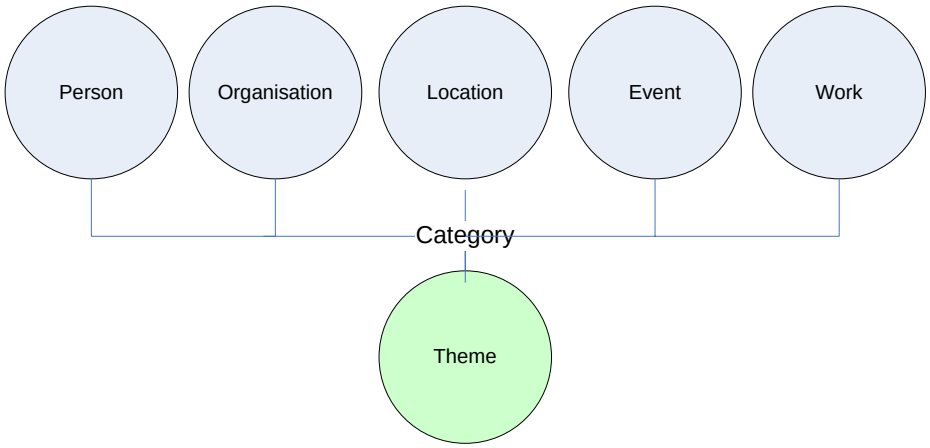
A knowledge base is included with a set of selected Named Entities. It is similar to a dictionary with locations, organisations, companies, persons, events… as needed when appearing within one of the editorial system at AFP.

Each entity can have properties such as alias or descriptions in different languages or specificities such as the year of birth for a person.

This knowledge is bringing the necessary consistency and reliability needed for all the content processed by our various systems.



Those named entities can also be connected through a semantic network with complex topics. This is essential if we want to connect multiple elements such as a company name to a location, a person, his role within a company…



The knowledge database is then a reference for the generation of interrelated content with internal and external links.

The system is also able to infer Named Entities not pertaining to the Knowledge database or relations between Named Entities from linguistics rules. Those potential new entries could then be proposed as new candidate to an administrator which is essential in a constantly moving environment as the one found in the News industry.

6. Conclusion

Adding value to content of various media types and selling and delivering customised content to various types of customers can be the definition of any news agencies' business today.

We know how to gather and produce the basic information but our main goal is now to increase the value of our products and services to our Clients - being "mono" or "multi" media - by a wide and efficient use of metadata within our news items and links (internal and external).

As we have at the same time to safeguard the specificities of our business and remain highly competitive with speed and accuracy in our reports, this project is in fact forcing a redesign of our Editorial systems in a broad sense.

We have anticipated the current trend and for the past 6 years have been heavily involved in the development of new standards at the IPTC level as well as in enhancing cooperation with manufacturers on pilot projects.

Two different – but also complementary strategies – have been tested to reach this target.

- Rely on the journalists' willingness to enrich the production
- Implement new tools to apply the content enrichment in the background with standard taxonomies and ontology.

At the moment most of the people involved in the projects have the feeling that the standardisation will bring a lot of benefits to both the internal and the external Clients and that the cost for it will be acceptable (especially the cost seen by the producer at his end of the information chain).

We still expect that those new developments will also help increase the service quality as perceived by the Subscribers as it should ease browsing through the massive amount of information brought to them and allow for an efficient access to the pertinent documents.

One of the key for the success of these new features is also the possibility to include links between news items of the same nature as well as of different nature leading to real Multimedia content management as early as possible in the production chain.

AFP is committed to support standards, particularly those from the IPTC and the W3C.

After the implementation of NewsML™ as our standard format for distribution of XML and multi-media content, our strong implication in the development of the NewsML 2G new format should be seen as a clear sign of the news agencies to jump on the semantic Web bandwagon.

Glossary

Desk

A desk is an editorial entity in charge of receiving the production (news items) from the field or the specialised production department, to sort them, to edit them and to validate them:

- The sorting action consists in removing the information that will not be used in an editorial product managed by the desk
- The editing action consists in checking that the copy conforms to the Agency's editorial rules and is adapted to the customer's requirements. The editor can enrich a news item by adding some background information. Corrections, translation and truncation are common actions carried on the desks. Any change that could possibly change the angle or meaning of the text is made in accordance with the producer.

Document

A document is an editorial object from one nature (text, photo, graphic, video and multimedia) which follows a manufacturing process carried out by journalists and/or photographers.
It has a particular Type (NewsItem IPTC).
It goes through various stages that correspond to its editorial lifecycle:
in production, produced, in edition, published, archived, deleted.

Event

An event is defined as « what happens and what can be potentially covered » (as news).

The event can either be planned or not. The unpredictable can take priority over the entirety of planned events.

Explicit or dynamic collection

A collection is either explicit or dynamic. An explicit collection is a set of structured elements; each of these elements can be a document or a collection.

A dynamic collection contains in addition and executable query on a corpus. Once the query has been executed, the collection that contains the result of the query becomes explicit.

Explicit Editorial Link

An explicit link is an editorial object linking two editorial objects together. It is created by actors in order to enrich the editorial products and to bring them added-value. It has a source, a destination, and a lifetime (optional).

Editorial Product

An editorial product is composed by a document stream or by collections prepared by the journalists and photographers and validated for their delivery to clients (stream service, Internet Journal…).

IPTC

The International Press Telecommunications Council (IPTC) was founded in 1965 to safeguard the telecommunication interests of the world's press. IPTC develops and maintains technical standards to improve the free exchange of news which are adopted by virtually every major news provider world wide. www.iptc.org

Metadata

Metadata is data associated to the document (but is not a part of its strict content) and enables to classify the document with various criteria.

News Codes

Is an IPTC standard to assign metadata values from predefined common sets (Subject Codes & Qualifiers, Genre, news status, news types …)

NewsML™

NewsML is a media independent IPTC standard for describing news in an electronic service environment. NewsML defines an XML based language for expressing the structure of news, associated metadata, and relationships between news, throughout their lifecycle.

The current NewsML version is v1.0, ratified in October 2000 by IPTC members.
Its new generation NewsML 2G to be released in 2007 will allow to package content across media and content types. www.newsml.org

Process

A process is the entire chain of tasks performed by a group of actors.

Production Services

The coverage of various news events is made by the journalists and photographers from production services and gives birth to the creation of news items. The same process applies through all media.

The collected information is sorted, re-read, corrected, adapted, sometimes rewritten in the production department prior to being sent to the desks.

The basic principle at production level is to create and complete the documents, as fast as possible, according to AFP's rules.

The guiding and assistance functions of the editorial system help the journalists and the photographers to respect these rules.

Acknowledgments.

This document was made available thanks to the contribution of Laurent Le Meur, head of the MULTIMEDIA LAB of AFP, chairman of the IPTC News Architecture Working Party.