# Explorer des actualités multimédia dans le web de données

Raphaël Troncy

# Explorer des actualités multimédia dans le web de données

Raphaël Troncy[1]

CWI Amsterdam, Science Park 123, 1098 XG Amsterdam, The Netherlands
raphael.troncy@cwi.nl

**Résumé** : Pour faciliter l'échange des actualités, l'IPTC (*International Press Telecommunication Council*) a développé l'Architecture NewsML (NAR) composée d'un modèle XML pour représenter les métadonnées et de vocabulaires contrôlés (IPTC News Codes) pour catégoriser les dépêches de presse. D'autres formats de métadonnées spécifiques au multimédia peuvent être utilisés conjointement mais cela pose des problèmes d'interopérabilité puisque les modèles XML fermés sous-jacents empêchent en particulier de lier ces métadonnées à d'autres connaissances disponibles sur le web. Dans cet article, nous proposons un environnement unique pour chercher et naviguer dans des contenus multimédia d'actualités contextualisés. Nous présentons une ontologie OWL pour l'Architecture NewsML, liée à d'autres ontologies multimédia. Nous montrons comment les métadonnées fournies par les journalistes peuvent être automatiquement enrichies par des méthodes de traitement automatique de la langue et du signal multimédia, pour ensuite être liées à des connaissances formalisées dans le web de données. Nous fournissons des recommandations quant à développer une ontologie à partir d'un schéma et d'en formaliser les connaissances implicites.

**Mots-clés** : Actualités multimédia, Web de données, Interface d'exploration.

## 1 Introduction

Dans le cycle de vie d'une actualité, l'information est généralement : *i)* produite par une agence de presse, un journaliste indépendant ou citoyen, *ii)* consommée et enrichie par un quotidien, un magazine ou un diffuseur radio-télévisé et finalement *iii)* livrée à des utilisateurs finaux. Les dépêches de presse sont typiquement accompagnées de métadonnées et d'une description brève du contenu pour faciliter leur indexation et leur recherche dans des archives. Cependant, la plupart de ces métadonnées se perd à cause de problèmes d'interopérabilité entre les différents acteurs des processus métier de production des actualités. De plus, les interfaces utilisateur de consultation utilisent rarement ces métadonnées. Par conséquent, les utilisateurs sont souvent obligés d'utiliser des environnements qui, pour une recherche donnée, contiennent une grande quantité d'information non pertinente, souvent redondante ou peu fiable, avec un accès insuffisant à de la connaissance de contexte pour comprendre ces actualités.

Notre objectif à long terme est de créer un environnement qui permettrait aux utilisateurs de voir les relations causales, logiques et sémantiques entre des actualités mul-

timédia prises individuellement, en utilisant leurs descriptions formalisées et la connaissance disponible sur le web. Notre approche consiste à créer des modèles de connaissance pour améliorer les problèmes d'interopérabilité dans toute la chaîne de production des actualités. Le problème que l'on cherche à résoudre dans cet article couvre les deux extrêmes de cette chaîne de production : comment représenter le sens des métadonnées tout au long du flux d'information et quelle conséquence cette modélisation a sur l'interface utilisateur finale.

Notre contribution est double. Nous présentons tout d'abord la modélisation d'ontologies OWL pour les langages de descriptions standardisés par l'IPTC, nous convertissons les vocabulaires contrôlés dans un thésaurus SKOS et nous montrons comment les métadonnées peuvent être automatiquement enrichies et intégrées à de la connaissance formalisée disponible sur le web. Nous généralisons cette approche et nous fournissons quelques recommandations quant à modéliser une ontologie formelle à partir d'un schéma. Nous discutons ensuite les décisions de modélisation en essayant d'évaluer leurs conséquences sur les interfaces utilisateur. Nous présentons finalement un prototype qui permet de chercher et de naviguer dans des actualités à partir de leurs descriptions formelles (Troncy, 2008).

L'article est structuré de la manière suivante. Nous introduisons brièvement dans la section suivante les standards principaux utilisés par l'industrie des médias. Nous discutons dans la section 3 des méthodes existantes pour construire des ontologies à partir de schémas et nous présentons les différentes tentatives pour intégrer ontologie pour représenter les actualités et ontologie multimédia. Nous détaillons dans la section 4 les étapes que nous préconisons pour construire une infrastructure sémantique pour les actualités. Pour démontrer l'adéquation de cette infrastructure, nous présentons un système de recherche sémantique pour des actualités multimédia (section 5) avant de conclure et d'ouvrir quelques perspectives à ces travaux (section 6).

# 2 Les standards pour l'actualité et le multimédia

## 2.1 Les standards pour l'actualité

Historiquement, l'IPTC a développé les formats NITF[1] et NewsML pour décrire la transmission, la structure et le contenu des informations d'actualités. Ces langages XML ont, cependant, démontré leurs limites pour décrire des actualités de plus en plus multimédia, et ont souvent été jugés trop verbeux. L'IPTC a donc récemment produit l'architecture NAR[2] qui fournit un cadre général pour une seconde génération de spécifications (G2). NAR est un modèle générique qui définit quatre objets principaux (`newsItem`, `packageItem`, `conceptItem` and `knowledgeItem`) ainsi que les opérations et traitements associés à leurs structures. Des langages spécifiques pour décrire des actualités (NewsML G2) ou des événements (EventsML G2) étendent ensuite cette architecture. Ainsi, l'élément générique `newsItem` est spécialisé pour prendre en considération les différents médias (dépêche textuelle, image, clip vidéo).

[1] News Industry Text Format : `http://www.nitf.org/`

[2] `http://www.iptc.org/NAR/`

IPTC maintient finalement un ensemble de vocabulaires contrôlés appelés *IPTC NewsCodes* qui sont utilisés pour catégoriser les dépêches d'actualités. Le thésaurus *Subject Code* contient par exemple 1300 termes organisés sur trois niveaux hiérarchiques pour décrire le sujet principal de chaque dépêche.

## 2.2 Les standards pour le multimédia

Bien que NAR défini des concepts pour représenter différents médias (textuel, photo, audio, vidéo, graphique, animation), une multitude d'autres standards sont utilisés par l'industrie des médias (Hausenblas *et al.*, 2007). Ainsi, les photos prises par les journalistes contiennent des métadonnées EXIF fournies par l'appareil spécifiant les caractéristiques de la photo (e.g. taille, orientation) ou des informations liées à sa prise (e.g. focale, temps d'exposition, flash). Kanzaki[3] et Norm Walsh[4] ont tous deux proposés une ontologie RDFS de EXIF et fournissent un service pour extraire et convertir ces métadonnées contenues dans l'en-tête des images.

Ces métadonnées techniques sont généralement complétées avec d'autres standards dont le but est de décrire le contenu. DIG35 est par exemple une spécification de l'I3A (*International Imaging Association*) qui définit un schéma XML pour représenter les paramètres de l'image, les informations de création, ce que l'image représente (qui, quoi, quand et où), ou encore les droits associés à l'image. En collaboration avec l'université de Ghent, nous avons récemment proposé une ontologie pour ce format[5] dont la modélisation suit les mêmes principes que nous exposerons dans la section 4. XMP fournit un modèle RDF natif pour décrire la gestion, les droits et le contenu d'images en ré-utilisant le format du Dublin Core. IPTC a lui-même intégré XMP dans son propre format de métadonnées pour les images.

Une vidéo peut être décomposée et décrite en utilisant le standard MPEG-7 (MPEG-7, 2001). Ce langage fournit un ensemble important de descripteurs pour décomposer un média, gérer les métadonnées de catalogage, représenter les caractéristiques de bas niveau du signal ou encore définir des concepts plus abstraits. L'ambiguïté et le manque de sémantique formelle de MPEG-7 ont déjà été mises en avant, et plusieurs ontologies OWL ont été proposées et récemment comparées (Troncy *et al.*, 2007). L'ontologie COMM (*Core Ontology for Multimedia Annotation*) propose par exemple une nouvelle conceptualisation du standard en utilisant DOLCE comme ontologie de haut niveau et en créant de nouveaux patrons de conception (*design patterns*) pour le multimédia (Arndt *et al.*, 2007). Chez les diffuseurs, l'EBU[6] a récemment adopté l'architecture NAR pour décrire les vidéos en fournissant quelques extensions pour décrire plus finement le contenu et gérer les droits associés.

En conclusion, la chaîne de production d'actualités utilise de nombreux standards et formats, souvent basés sur XML, mais intrinsèquement fermés, ce qui conduit à des problèmes d'interopérabilité. De plus, ces formats empêchent d'utiliser de nouveaux vocabulaires contrôlés qui n'auraient pas été prévus au préalable, ou plus généralement

[3] `http://www.kanzaki.com/ns/exif`

[4] `http://sourceforge.net/projects/jpegrdf`

[5] `http://multimedialab.elis.ugent.be/users/chpoppe/Ontologies/`

[6] European Broadcaster Union : `http://www.ebu.ch`

de la connaissance disponible sur le web. Nous proposons d'utiliser les langages du web sémantique pour faciliter l'intégration de ces standards. En se basant sur l'état de l'art de la construction d'ontologies décrit dans la prochaine section, nous présentons notre infrastructure sémantique et nous formulons des recommandations pour modéliser une ontologie formelle à partir de schémas existants.

# 3 Etat de l'art

## 3.1 Convertir des schémas et des thésauri pour le web sémantique

Les productions en matière de construction d'ontologies sont nombreuses. Uschold et Grüninger s'intéressent par exemple à l'ensemble du processus de conception de l'ontologie et de son cycle de vie, tandis que METHONTOLOGY propose de modéliser l'ontologie au niveau des connaissances en utilisant des représentations intermédiaires. D'autres travaux insistent sur la conceptualisation des taxonomies de concepts et de relations (Troncy & Isaac, 2002), tandis que la méthodologie ARCHONTE se veut constructiviste et préconise un retour au corpus textuel pour sélectionner des concepts et des relations dont leur sens est normalisé avant d'être formalisé (Bachimont, 2004). Toutes ces méthodologies, cependant, ne considèrent pas le cas supposé plus simple où un schéma semi-formel (diagramme UML, schéma XML, thésaurus) modélisant les connaissances du domaine, existe mais doit être formalisé pour être utilisable sur le web sémantique.

Une méthode générale pour convertir des thésauri en SKOS a également été proposée (Assem *et al.*, 2004). Cette méthode préconise quatre étapes : préparation, conversation syntaxique, conversion sémantique et standardisation. Notre méthode s'inspire de ces recommandations mais en ajoute de nouvelles pour conceptualiser l'ontologie à partir de schémas semi-formels. L'alignement entre des thésauri convertis en SKOS et des ressources du web sémantique a été discuté (Tordai *et al.*, 2007). Nous utilisons l'outil AnnoCultor[7] issu de ces travaux pour convertir les vocabulaires contrôlés de l'IPTC en thésauri SKOS.

## 3.2 NewsML et les ontologies multimédia

Plusieurs travaux ont tenté de construire une ontologie des actualités. Le projet européen NEWS[8] propose une ontologie RDFS multilingue à partir des formats NITF et NewsML et du thésaurus *SubjectCodes* (Fernández *et al.*, 2006). Le projet Neptuno[9] suit la même approche et propose une ontologie RDFS construite également à partir de NewsML et du thésaurus *SubjectCodes* aligné avec d'autres vocabulaires contrôlés internes à l'agence de presse espagnole (Castells *et al.*, 2004) dans le but de décrire son archive. MESH[10] est finalement un projet européen en cours qui met l'accent sur l'analyse multimédia pour enrichir automatiquement les métadonnées et construire des

[7] `http://sourceforge.net/projects/annocultor`

[8] `http://www.news-project.com/`

[9] `http://seweb.ii.uam.es/neptuno/`

[10] `http://www.mesh-ip.eu/`

résumés personnalisés. Une ontologie des actualités semble avoir été développée mais celle-ci n'est pas encore disponible.

A la différence de ces projets, notre approche consiste à dissocier les thésauri utilisés pour valuer les propriétés des métadonnées, de l'ontologie qui décrit la gestion et le contenu des dépêches selon le point de vue du journaliste. Cette séparation fournit une architecture plus flexible où, par exemple, le thésaurus *SubjectCodes* peut être indépendamment aligné avec d'autres vocabulaires contrôlés. Nous publions également ces thésauri sur le web sémantique, en fournissant pour chaque terme un URI déréférençable[11]. Finalement, notre ontologie est conforme aux standards les plus récents de l'IPTC (NAR) et nous la lions à d'autres ontologies suivant les bons principes du web (sémantique).

La méthodologie *XML Semantics Reuse* consiste à transformer automatiquement un schéma XML en une ontologie OWL[12]. Elle a été utilisée dans le domaine du journalisme pour convertir les formats NITF, NewsML et MPEG-7 ainsi que le thésaurus *SubjectCodes* de l'IPTC en OWL/RDF (Garcia *et al.*, 2008). L'ontologie résultante, cependant, ne formalise pas la sémantique informelle de ces standards puisque celle-ci n'est pas présente dans les schémas XML originaux (Troncy *et al.*, 2007). De plus, la transformation automatique re-créée des structures complexes imbriquées (e.g. des éléments intermédiaires correspondants à des types ou des ensembles d'éléments) qui n'ont pas lieu d'être dans une ontologie. Nous préconisons au contraire de re-modéliser l'ontologie en suivant les bonnes pratiques détaillées ci-dessous.

# 4 Une infrastructure sémantique pour les actualités

NAR est un modèle générique pour décrire le contenu des actualités, et la manière de les gérer et de les échanger. Ce modèle partage assez naturellement les mêmes principes que le web sémantique :

– les actualités sont des ressources distribuées qui ont besoin d'être identifiées de manière unique et pérenne ;
– les actualités sont décrites avec des vocabulaires contrôlés et partagés.

NAR est cependant défini à l'aide d'un schéma XML et par conséquent, sa sémantique implicite n'est pas formellement représentée (par exemple, un `NewsItem` peut être un `TextNewsItem`, un `PhotoNewsItem` ou un `VideoNewsItem`). L'extension du modèle à d'autres standards est également compliquée puisque il est difficile d'établir l'équivalence de deux éléments XML. Nous décrivons dans la suite les étapes nécessaires pour modéliser notre infrastructure ontologique[13].

## 4.1 Étape 1 : modéliser l'ontologie NAR

La première étape consiste à formaliser la sémantique implicite des standards de l'IPTC. Bien que ces modèles existent sous forme de diagrammes UML, leur "ontologi-

[11] Le déréférencement consiste à accéder à la représentation d'une ressource identifiée par une URI. L'expression *deferencing an URI* a fait l'objet d'une longue discussion au sein du W3C pour finalement aboutir à la résolution communément appelée httpRange-14.

[12] Voir le projet ReDeFer : `http://rhizomik.net/redefer`

[13] Les différentes ontologies sont disponibles à `http://newsml.cwi.nl/ontology/`

sation" n'est pas triviale. Nous discutons ci-dessous quelques décisions de modélisation.

**Aplatir la structure XML.** XML Schema permet de définir des structures relativement riches, mais il est limité quant à définir leur sémantique puisque le langage fournit principalement un système de typage pour des données structurées. Ainsi, le modèle NAR contient un certain de nombre de structures intermédiaires, sorte de container dont le but est juste de regrouper des éléments sans avoir un sens particulier. Ces structures ne doivent pas être représentées dans l'ontologie puisqu'elles ne seront pas instanciées et correspondront à des noeuds anonymes dans le graphe RDF. Pour conceptualiser l'ontologie, nous préconisons donc d'aplatir le schéma XML en ne gardant que les propriétés qui seront instanciées.

**Indiquer la provenance.** Les assertions contenues dans les dépêches ont souvent besoin d'être réifiées. Par exemple, un éditeur enregistré comme `team:md` peut indiquer qu'une dépêche a été catégorisée `diplomatie` le `2005-11-11T08:00:00Z`, ce qui se traduit en RDF par :

```
{<> nar:subject cat:11002000} dc:creator   team:md ;
                  dc:modified  ''2005-11-11T08:00:00Z'' .
```

Le mécanisme de réification de RDF n'a cependant pas de sémantique formelle dans la théorie des modèles. Pour représenter la provenance des informations, nous préconisons donc d'utiliser la technique des graphes liés et nommés où les relations entre graphes sont décrites à l'aide de requêtes SPARQL et de vues (Schenk & Staab, 2008).

## 4.2 Étape 2 : la lier avec d'autres ontologies

Comme nous l'avons vu dans la section 2.2, d'autres standards tels que EXIF, Dublin Core, XMP, DIG35 ou MPEG-7 sont utilisés par l'industrie des médias. Ces standards ont généralement été traduits, ou existent nativement en OWL/RDF et peuvent donc s'intégrer naturellement à notre architecture. Par conséquent, nous préconisons d'ajouter des axiomes dans l'ontologie pour expliciter les relations entre concepts provenant d'ontologies différentes mais qui se recouvrent partiellement. Ainsi, l'ontologie NAR contient les axiomes suivants :

```
nar:subject  owl:equivalentProperty  dc:subject
nar:Person   owl:equivalentClass     foaf:Person
```

Les moteurs de recherche du web sémantique tels que Sindice[14], Watson[15] ou Falcon[16] peuvent être utilisés pour découvrir de nouveaux concepts ou propriétés partageant le même sens que ceux définis dans notre ontologie et auxquels ils pourraient être liés.

## 4.3 Étape 3 : convertir les IPTC News Codes en thésauri SKOS

Les IPTC *NewsCodes* sont définis dans 36 thésauri, de tailles variables, dont les termes sont utilisés dans les métadonnées décrivant les actualités. Bien que ces termes

[14] `http://sindice.com/`

[15] `http://watson.kmi.open.ac.uk/WatsonWUI/`

[16] `http://www.falcons.com.cn/`

soient parfois organisés en taxonomie, la relation de subsomption n'est jamais explicite mais encodée dans les noms de termes. Ainsi, "cancer" (`cat:07001004`) est plus spécifique que "maladie" (`cat:07001000`) qui est lui même plus spécifique que "santé" (`cat:07000000`) simplement parce qu'ils partagent les mêmes quatre premiers chiffres. Nous avons converti ces thésauri en SKOS en explicitant cette relation de subsomption à l'aide des constructeurs `skos:narrower` et `skos:broader`.

Cette compatibilité RDF nous permet de définir plus en avant certains concepts de l'ontologie NAR en terme de `owl:Restriction` : la valeur d'une propriété peut être un `skos:Concept` ou doit provenir d'un `skos:ConceptScheme` particulier. Par exemple, la propriété `nar:subject` ne peut avoir comme valeur qu'un terme provenant du `skos:ConceptScheme` *SubjectCodes*.

Finalement, nous avons exposé[17] tous ces thésauri sur le web sémantique en suivant le guide des bonnes pratiques promues par le W3C[18]. Chaque terme est donc identifié par un URI déréférençable. Ainsi, une requête http dont le type est `Accept:text/html` retournera la description XML originale de l'IPTC tandis que le type `Accept:application/rdf+xml` retournera la version SKOS/RDF utilisable par les machines du thésaurus.

## 4.4 Étape 4 : enrichir automatiquement les métadonnées

Une fois l'ontologie NAR modélisée et liée à d'autres ontologies populaires sur le web sémantique, la conversion des métadonnées de chaque dépêche en RDF selon notre architecture est triviale. Cependant, nous préconisons une ultime étape dont le but est d'enrichir automatiquement les métadonnées en suivant les principes des données liées[19]. Ainsi, nous utilisons des techniques de traitement automatique de la langue et du signal multimédia pour extraire d'avantages de métadonnées (Figure 1).

Le traitement automatique de la langue consiste à extraire les entités nommées telles que les personnes, les organisations, les lieux, les marques, etc. à partir de la légende d'une photo ou d'une dépêche textuelle. Nous utilisons désormais le service OpenCalais[20] après avoir expérimenté avec les plate-formes GATE[21] et SPROUT[22]. Une fois les entités nommées extraites, nous les alignons avec des ressources disponibles dans le web des données, à savoir, avec Geonames pour les lieux, ou avec DBPedia pour les noms de personnes et d'organisations. Le traitement du signal fournit également un autre type de métadonnées utilisé ultérieurement pour classifier les résultats d'une requête. Ainsi, il est possible d'effectuer une classification non supervisée de toutes les images montrant le footballeur Zinedine Zidane en utilisant les descripteurs de texture et d'histogramme de couleur, et ainsi de différencier les photos où il apparaît en costume pour recevoir un prix et où il est sur le terrain.

[17] `http://newsml.cwi.nl/NewsCodes/`

[18] `http://www.w3.org/TR/swbp-vocab-pub/`

[19] Linked Data : `http://linkeddata.org/`

[20] `http://www.opencalais.com/`

[21] `http://gate.ac.uk/`

[22] `http://sprout.dfki.de/`

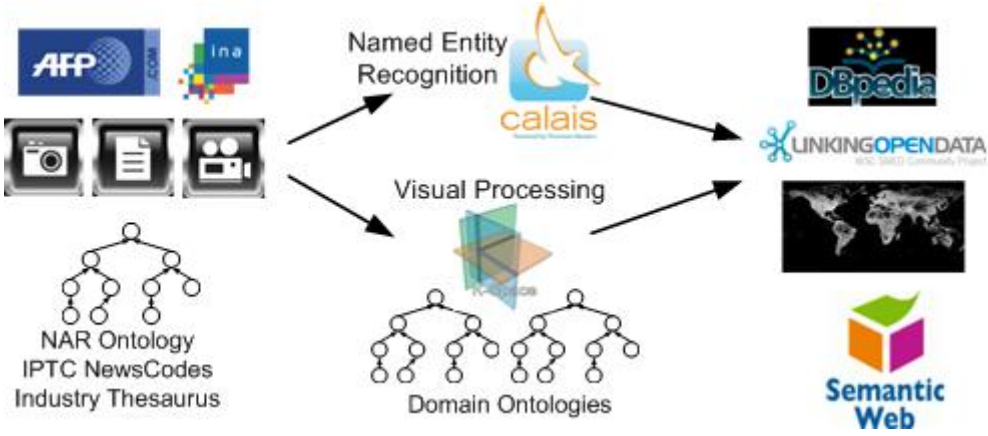


FIG. 1 – Enrichissement automatique des métadonnées des actualités

# 5 Explorer des actualités multimédia

Dans le but de démontrer l'utilité de notre architecture sémantique, nous présentons dans cette section un prototype pour chercher et explorer des actualités multimédia. Nous utilisons Cliopatria[23], une plate-forme basée sur SWI-Prolog comprenant un entrepôt pour des données RDF, l'implémentation des langages de requêtes SeRQL/SPARQL, les bibliothèques d'interface utilisateur de Yahoo ! (YUI) et des routines pour la recherche sémantique (Hildebrand *et al.*, 2006; Wielemaker *et al.*, 2008). A la différence d'*Exhibit* de Simile, ClioPatria offre une architecture client-serveur qui retourne des objets auxquels il est possible d'appliquer des styles de présentation et des stratégies d'interaction personnalisés. Nous présentons dans la suite les données utilisées pour notre expérimentation et nous montrons comment la formalisation des métadonnées permet d'obtenir des dimensions pour présenter les résultats d'une recherche ou pour guider l'utilisateur dans une navigation par facettes.

## 5.1 Données utilisées

Le jeu de données utilisé dans notre expérimentation comprend : 100000 dépêches de presse en anglais et en français, 2557 photos et 8 heures de journaux télévisés couvrant la période juin et juillet 2006 (Tableau 1). Suivant les quatre étapes détaillées dans la section 4, nous avons analysé ces dépêches pour enrichir automatiquement les métadonnées. Ainsi, l'analyse de la légende des 2557 photos fournit 217 personnes connues dans DBPedia et 426 lieux connus dans Geonames. L'évaluation manuelle des résultats montre que le service Geonames a tendance à toujours trouver d'abord une ville américaine pour chaque requête. Les dépêches contiennent toutefois la plupart du temps un nom de ville et de pays, ce qui permet d'avoir une reconnaissance précise du lieu mentionné dans la dépêche. Les quelques erreurs que nous avons observées pro-

[23] http://e-culture.multimedian.nl/software/ClioPatria.shtml

viennent d'un mauvais typage de l'entité nommée (e.g. `Australia` a parfois été typé comme une `Person`). Nous sommes actuellement en train d'évaluer des algorithmes plus sophistiqués tels que IdentityRank (Fernández *et al.*, 2007) pour minimiser ces problèmes d'homonymie.

| Description | Nb de triplets RDF |
|---|---|
| Ontologies générales : NAR, NewsML-G2, DC, VRA, FOAF | 7,336 |
| Ontologies de domaine et BC (football) | 104,358 |
| Thésauri : IPTC NewsCodes, Thesaurus INA | 34,903 |
| Ressources externes : Geonames, DBPedia | 53,468 |
| Fil des dépêches de l'AFP en anglais (Juin et Juillet 2006) | 804,446 |
| Photos de l'AFP de la coupe du monde 2006 | 61,311 |
| Journaux télévisés archivés à l'INA (Juin et Juillet 2006) | 1,932 |
| **Total** | 1,067.754 |

TAB. 1 – Nombre de triplets RDF chargés dans ClioPatria

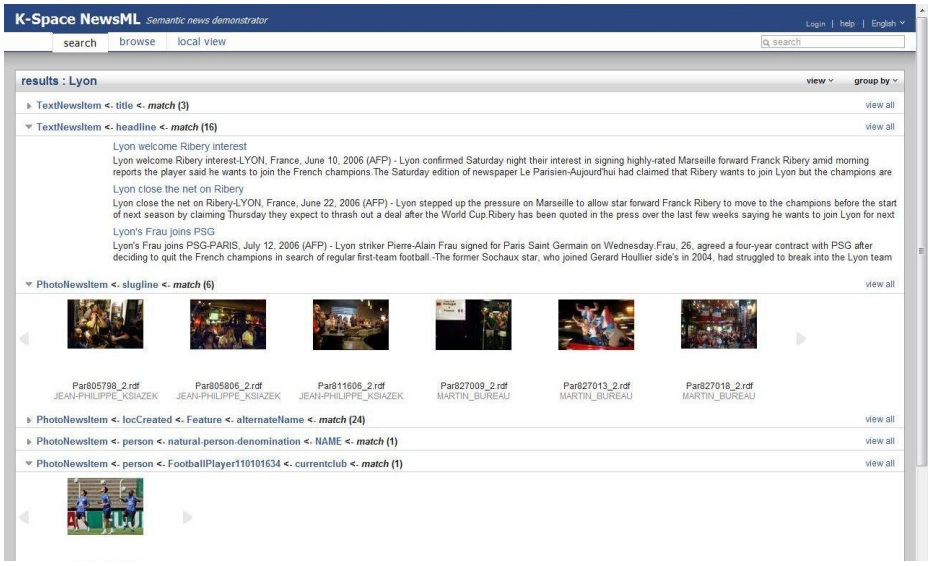


FIG. 2 – Recherche avec le terme "Lyon" dans notre moteur de recherche sémantique

## 5.2 Recherche sémantique d'actualités

La figure 2 montre le résultat de notre système pour la requête "Lyon". Les actualités sont groupées selon le chemin dans le graphe RDF qui mène à la propriété pour laquelle la valeur est pertinente pour la requête. Dans notre cas, le système retourne les dépêches où "Lyon" apparaît dans les propriétés `title`, `headline`, `slugline`,

etc. De plus, le type des dépêches permet de personnaliser leur présentation. Ainsi, les trois premières lignes d'une dépêche textuelle seront affichées alors que les images sont présentées sous la forme d'un carrousel.

Le dernier groupe affiché dans la figure 2 contient une photo unique avec trois footballeurs. De manière surprenante, les métadonnées de cette photo ne contiennent pas le terme "Lyon". En revanche, la légende mentionne le joueur Juninho Pernambucano, personne connue dans DBPedia qui fournit des informations supplémentaires sur cette personne telles que sa date de naissance ou les clubs dans lesquels il a joué, et en particulier son club actuel : "Lyon". Cet exemple illustre la puissance (et les limites) de notre système : il est désormais possible de trouver des documents ayant une relation, même lointaine, avec la requête et d'expliciter cette relation (le graphe RDF). Ce résultat figurera toutefois en queue d'un classement de pertinence dû à la longueur du graphe.

De manière similaire, la requête "Saksamaa" retourne un groupe de 679 photos pour lesquelles ce terme n'apparaît jamais dans les métadonnées. L'explication provient du fait que toutes ces photos ont été prises pendant la coupe du monde en Allemagne, une entité nommée reconnue comme un lieu et donc lié à la base Geonames, qui fournit également le libellé du pays dans toutes les langues, *Saksamaa* signifiant Allemagne en éthiopien.

### 5.3 Exploration sémantique des actualités

En plus d'une interface de recherche, notre prototype contient un navigateur à facettes pour mieux explorer le contenu d'une archive. Les facettes correspondent à certaines propriétés jugées d'intérêt particulier dans les métadonnées et sont configurables par l'utilisateur. Ainsi, nous avons défini une vue dédiée au football qui contient les propriétés `subject`, `slugline`, `locCreated`, `location` et `person`. Le lien avec la base de connaissance Geonames permet de proposer des vues plus riches pour présenter les actualités. La figure 3 montre par exemple toutes les photos du joueur Zinedine Zidane sur une carte à l'emplacement où les photos ont été prises (drapeau bleu) ou à l'emplacement mentionné dans les dépêches (drapeau rouge).

Cette vue présente parfois un réel intérêt. Deux séries de photos dont les mots clés sont `FBL-WC2006-MATCH64ITA-FRA` et `FBL-WC2006-MATCH64-FRA-ITA` semblent en effet concerner le même événement. Leur affichage sur une carte permet immédiatement de comprendre ce qui les différencie : les dépêches ont été produites soit en Italie, soit en France adoptant un point de vue inévitablement biaisé de la finale.

## 6 Conclusion et perspectives

Nous avons décrit dans cet article une méthode composée de quatre étapes pour construire une infrastructure basée sur les ontologies pour les actualités. Ces recommandations sont complémentaires des patrons de conception ontologique et des guides de bonnes pratiques pour publier des ontologies dans le web de données. Nous avons discuté nos décisions de modélisation : au niveau ontologique, nous préconisons d'aplatir les structures XML quand l'ontologie est construite à partir d'un schéma et de ré-utiliser le plus possible les vocabulaires existants ; au niveau des instances, nous recommandons

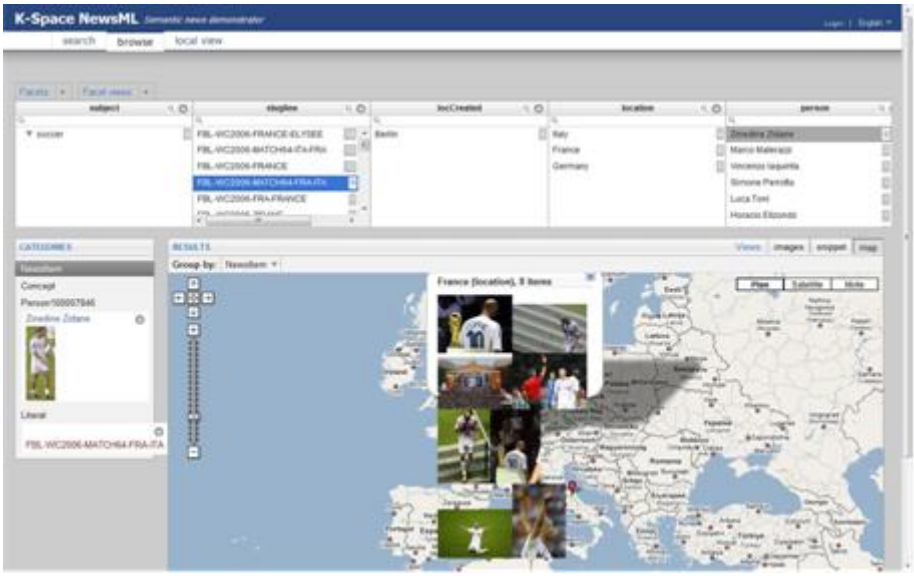


FIG. 3 – Explorer les photos prises pendant la finale de la coupe du monde

d'enrichir et de lier les métadonnées avec des thésauri SKOS ou de la connaissance formalisée disponible dans le web de données telle que les bases Geonames ou DBPedia. L'ontologie NAR est actuellement en cours de revue par l'IPTC et pourrait être approuvée prochainement. Nous avons présenté un prototype pour chercher et explorer des actualités. Les interfaces utilisent la richesse des métadonnées sémantiques pour grouper, classer et présenter le résultat des requêtes. Cet environnement est publiquement disponible à `http://newsml.cwi.nl/explore/search`.

Le temps est une dimension fondamentale dans le domaine des actualités et notre système possède également des vues temporelles. Raisonner sur des données temporelles est néanmoins un problème complexe. Nous planifions d'inclure prochainement une ontologie du temps[24] et les relations temporelles de l'ontologie DOLCE dans le but de proposer une vue des dépêches agrégées par sujet, par jour, mois ou année. Notre système fonctionne actuellement avec des données statiques qui ont été préalablement analysées. Une évolution naturelle consiste à proposer un environnement dynamique alimenté par des fils de dépêches en continu. Finalement, une évaluation de notre prototype par des journalistes de l'AFP est planifiée.

## Références

ARNDT R., TRONCY R., STAAB S., HARDMAN L. & VACURA M. (2007). COMM : Designing a Well-Founded Multimedia Ontology for the Web. In *$6^{th}$ International Semantic Web Conference (ISWC'07)*, p. 30–43, Busan, Korea.

[24] `http://www.w3.org/TR/owl-time/`

ASSEM M. V., MENKEN M. R., SCHREIBER G., WIELEMAKER J. & WIELINGA B. (2004). A Method for Converting Thesauri to RDF/OWL. In *3rd International Semantic Web Conference (ISWC'04)*, p. 17–31, Hiroshima, Japan.

BACHIMONT B. (2004). *Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle*. Habilitation à diriger des recherches, Université de Compiègne.

CASTELLS P., PERDRIX F., PULIDO E., RICO M., BENJAMINS R., CONTRERAS J. & LORÉS J. (2004). Neptuno : Semantic Web Technologies for a Digital Newspaper Archive. In *1st European Semantic Web Symposium (ESWS'04)*, p. 445–458, Heraklion, Crete.

FERNÁNDEZ N., BLÁZQUEZ J. M., ARIAS J., SÁNCHEZ L., SINTEK M., BERNARDI A., FUENTES M., MARRARA A. & BEN-ASHER Z. (2006). NEWS : Bringing Semantic Web Technologies into News Agencies. In *5th International Semantic Web Conference (ISWC'06)*, p. 778–791, Athens, Georgia, USA.

FERNÁNDEZ N., BLÁZQUEZ J. M., SÁNCHEZ L. & BERNARDI A. (2007). IdentityRank : Named Entity Disambiguation in the Context of the NEWS Project. In *4th European Semantic Web Conference (ESWC'07)*, p. 640–657, Innsbruck, Austria.

GARCIA R., PERDRIX F., GIL R. & OLIVA M. (2008). The semantic web as a newspaper media convergence facilitator. *Journal of Web Semantics*, **6**(2), 151–161.

HAUSENBLAS M., BOLL S., BÜRGER T., CELMA O., HALASCHEK-WIENER C., MANNENS E. & TRONCY R. (2007). Multimedia Vocabularies on the Semantic Web. W3C Multimedia Semantics Incubator Group Report. `http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies/`.

HILDEBRAND M., OSSENBRUGGEN J. V. & HARDMAN L. (2006). /facet : A Browser for Heterogeneous Semantic Web Repositories. In *5th International Semantic Web Conference (ISWC'06)*, p. 272–285, Athens, Georgia, USA.

MPEG-7 (2001). Multimedia Content Description Interface. ISO/IEC 15938.

SCHENK S. & STAAB S. (2008). Networked Graphs : A Declarative Mechanism for SPARQL Rules, SPARQL Views and RDF Data Integration on the Web. In *17th International World Wide Web Conference (WWW'08)*, Beijing, China.

TORDAI A., OMELAYENKO B. & SCHREIBER G. (2007). Semantic Excavation of the City of Books. In *Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM'07)*, p. 39–46.

TRONCY R. (2008). Bringing the IPTC News Architecture into the Semantic Web. In *7th International Semantic Web Conference (ISWC'08)*, p. 483–498, Karlsruhe, Germany.

TRONCY R., CELMA O., LITTLE S., GARCÍA R. & TSINARAKI C. (2007). MPEG-7 based Multimedia Ontologies : Interoperability Support or Interoperability Issue ? In *1st International Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies (MAReSO)*, Genova, Italy.

TRONCY R. & ISAAC A. (2002). DOE : une mise en oeuvre d'une méthode de structuration différentielle pour les ontologies. In *13th Journées d'Ingénierie des Connaissances (IC'02)*, p. 63–74, Rouen, France.

WIELEMAKER J., HILDEBRAND M., OSSENBRUGGEN J. V. & SCHREIBER G. (2008). Thesaurus-based search in large heterogeneous collections. In *7th International Semantic Web Conference (ISWC'08)*, p. 695–708, Karlsruhe, Germany.