Fouille au Corps des Media Français: Un exemple concret de Fouille Multimodale Transmedia

Marie-Luce Viaud, Agnes Saulnier, Denis Teyssou, Nicolas Hervé, Benjamin Renoust, Jérôme Thievre

Fouille au Corps des Media Français
Un exemple concret de Fouille Multimodale Transmedia

M.L. Viaud[1], A. Saulnier[1]
D. Teyssou[2], N. Hervé[1], B. Renoust[1], J. Thièvre[1]

1 - INA 4 Avenue de l'Europe, Bry/Marne
mlviaud@ina.fr; asaulnier@ina.fr; nherve@ina.fr;
renoust@gmail.com; jthievre@ina.fr
http://www.ina.fr
2 - AFP, 2 Place de la Bourse, Paris
denis.teyssou@afp.com
http://www.afp.com

Résumé: Numérique, réseaux et mobiles ont bousculé irréversiblement la production, la diffusion et la consommation des media d'actualité. L'Observatoire Transmedia est une plateforme de description, d'unification et d'analyse des actualités diffusées par la radio, la télévision, le web et l'AFP, sur la période mi 2011/fin 2013. La collaboration des chercheurs en informatique, SHS et professionnels de l'information a permis la conception, la validation et l'amélioration de la plateforme. La réalisation d'études -analyse de propagation, chronologies d'évènements, ou encore analyse des taux de reprise des dépêches AFP- nous a permis de répondre partiellement aux questions initiales du projet: Comment l'actualité aborde t'elle un sujet ? Qui produit l'information ? La multiplication des supports garantit-elle la diversité?

1 Introduction

"Hollande est élu président". Il est 18h45 le 6 mai 2012, l'AFP, défiant la loi, annonce l'issue de la présidentielle française à l'instant même où la Suisse et la Nouvelle Zélande publient les premiers résultats de sondages sur Internet. De 2011 à 2013, une partie de la PQR est rachetée et fusionnée. Quelle influence sur son contenu ? Printemps 2013, France Culture inaugure les journées de partenariat avec une presse écrite qui, malgré l'aide toujours plus conséquente de l'Etat, peine à trouver une respiration économique face à la multiplication des supports et des usages. Les chaînes d'information en continu changent la donne lors de la campagne présidentielle de 2012, en diffusant en direct meetings et interviews des candidats. 155 passagers d'un A320 atterrissent dans le fleuve Hudson, le photographe amateur remercie le ciel d'avoir été là au bon moment en voyant son image diffusée sur tous les media !

Si le numérique est sans conteste une révolution sociétale en marche, l'industrie de l'information en subit l'impact de plein fouet. Ses pratiques de production, de diffusion, de consommation et ses modèles économiques sont bouleversés, amenant de nouvelles possibi-

lités, de nouvelles contraintes, mais aussi brouillant les rôles et les frontières des acteurs éponymes.

Cet article est une présentation générale du projet OTMedia. Il s'adresse à un public large, aussi les modules scientifiques sont-ils décrits de manière à faire comprendre les enjeux, les problématiques et les difficultés rencontrées, sans rentrer dans les détails scientifiques très pointus[1]. Nous nous attarderons sur les modules présentant un fort degré d'innovation, ainsi que sur les aspects fonctionnels de la plateforme qui offrent des processus d'observation, de manipulation et de validation efficients. Le prototype a pu ainsi être utilisé pour des études mandatées par l'ONU sur "l'immigration en période électorale" et par le Forum d'Avignon sur la "communication de la culture en France". Notre objectif est d'illustrer les possibilités offertes par la fouille de données multimodales pour observer et appréhender la production des média français et son évolution dans un contexte de mutation très agité.

Projet collaboratif de recherche des programmes CONTINT de l'ANR[2], l'Observatoire Transmedia (OTMedia) est une expérimentation grandeur réelle d'analyse de l'information diffusée en France par la Télévision, la Radio, le Web, l'Agence France Presse et Twitter. Le corpus, issu de la collecte sur la période juin 2011 à décembre 2013, est particulièrement riche en terme d'élections (les primaires socialistes, les élections présidentielles, les élections législatives, les élections américaines) mais comprend aussi l'affaire Merah, le développement des conflits syrien et malien, l'affaire Cahuzac… Le projet a pour objet l'étude et la traçabilité des évènements médiatiques grâce à recherche et la fouille de données transmedia. Les corpus médiatiques sont particulièrement intéressants pour des approches d'analyse logicielle car ils se caractérisent par de fortes redondances dans des volumes de données importants. Ces redondances se déclinent sur chaque modalité et présentent des liens de ressemblances assez diverses. En effet, les évènements médiatiques importants sont traités par tous les media et sur toutes les modalités ; ils se composent à la fois de reprises d'information totales (dépêche AFP ou autres) ou partielles ou de créations originales d'articles, billets ou reportages… Il s'agit donc de mettre en place une plateforme d'analyse des différents flux médiatiques pour détecter, circonscrire, suivre, mesurer, analyser et étudier la propagation des évènements médiatiques et leurs dérivés.

2 Un contexte de mutation agité : bref rappel historique

La presse écrite existe depuis le XVIIe siècle, les agences de presse se développent à partir du milieu du XIXème siècle, les media audiovisuels (radio et TV) font leur apparition au début du XXème siècle puis deviennent dominants dans la deuxième moitié du siècle. Des articulations entre ces acteurs médiatiques se sont donc établies pendant plusieurs décennies. Pour simplifier, tout s'organisait selon l'adage de Hubert Beuve-Méry, fondateur du quoti-

[1] Les lecteurs sont invités à consulter les références bibliographiques pour approfondir ces aspects.
[2] Ina (coordinateur), INRIA(Zenith), Syllabs, LIA, AFP, Univ Paris 3 (CIM), LATTS

dien Le Monde en 1944 et dont il restera le directeur jusqu'en 1969 : « La radio annonce l'événement, la télévision le montre, la presse l'explique. »

En 1981, est apparue la chaîne toutes informations de Ted Turner, Cable News Network (CNN), dont la diffusion souvent en direct des événements médiatiques a profondément bousculé le paysage des media, forçant les grandes agences à évoluer vers la diffusion de services d'information en vidéo (White 1997).

L'apparition des chaînes d'information en continu a ouvert une première brèche dans le système mondial de collecte de l'information médiatique, les grandes agences devant rivaliser, notamment avec CNN, pour donner l'information en premier. Dans le même temps, CNN et d'autres chaînes devenaient elles aussi une source d'informations pour ces agences.

L'apparition de l'Internet grand public, vers 1995, a ensuite contribué à agrandir cette brèche dans le système de communication de l'information hérité du XIX siècle. Il y a désormais interconnexion entre presse, radio, télévision, et Web, au point que certains parlent même d'écosystème de l'information plutôt que de système médiatique, pour mieux insister sur l'interdépendance entre tous ces media, interdépendance technique, en raison du continuum numérique, qui facilite le transfert d'une information d'un média à l'autre, doublé d'une interdépendance économique et sociale.

Les journaux, dont les scoops étaient autrefois rapportés par les agences, sont devenus, avec leurs sites Web, accessibles en quelques clics et quelques secondes, à l'autre bout du monde. Les radios, TV du monde entier sont disponibles sur le Web, la plupart du temps gratuitement, et rencontrent de nouveaux acteurs « pure players » (agrégateurs, sites participatifs, blogueurs).

Les internautes eux-mêmes sont les artisans d'une propagation de l'information d'un média à l'autre. Les extraits de journaux télévisés sont reproduits sur des plateformes de vidéos et, réciproquement, les témoignages amateurs, filmés via des appareils miniatures (caméscopes, smart-phones), alimentent les chaînes de télévision lors d'évènements impromptus. Les partis politiques, les bourses du monde entier, les gouvernements, les entreprises, les experts ont désormais leur site Web et communiquent via ce site, ou via des blogs ou des réseaux sociaux (de type Twitter ou Facebook) et communiquent de plus en plus en direct avec leurs sympathisants, administrés, clients... court-circuitant les media traditionnels.

Dans cette véritable révolution technologique, sociale et économique que subit le monde des media, quel est désormais le cheminement de l'information ? Quelle est encore l'influence des media dits traditionnels sur l'agenda médiatique, sur la consommation de l'information par les internautes ?

Si la « circulation circulaire de l'information » (Bourdieux 1996) entre journaux, stations de radio, et chaînes de télévision, avec souvent pour origine le recours aux dépêches d'agences ou aux communiqués de presse, caractérisait déjà l'époque précédente, le déploiement de l'Internet tend toutefois à amplifier ce phénomène. Tout d'abord en ajoutant un média de plus dans cet enchevêtrement informationnel. Ensuite, en facilitant techniquement la duplication des contenus sur plusieurs supports. Enfin, en modifiant l'échelle et la vitesse de cette circulation de l'information.

L'information est aujourd'hui l'objet de réappropriations multiples, par les professionnels du journalisme comme par les internautes amateurs, à tel point qu'il est parfois difficile d'en retrouver l'origine. Ceci amène évidemment des questions immédiates sur la fiabilité et la qualité de l'information : l'appropriation d'une information par les internautes, via les sites de réseaux sociaux ou les discussions dans la blogosphère, est-elle source d'enrichissement ou au contraire de dénaturation? Plus fondamentalement encore, si l'on considère que les media jouent un rôle démocratique dans l'information des citoyens, le fonctionnement de cet espace public « augmenté » par le numérique constitue-t-il une évolution favorable? Assiste-t-on à une démultiplication de contenus originaux, ou au contraire à une certaine redondance de l'information, reproduite sur chacun des media, même si dans des formats différents?

Dans leur "création destructrice" à la Schumpeter, quel rôle joue les nouveaux media dans le cheminement de l'information? Sont-ils des suiveurs, des copieurs ou au contraire des innovateurs? Renouvellent-ils le genre avec un positionnement plus libre, moins institutionnel, moins politiquement correct? Comment propagent-ils les informations dans cette "circulation circulaire" qui, bien que déjà présente à l'époque précédente, prend sur le net des allures de spirale sans fin? Alors que de plus en plus de publications sont contraintes de mettre la clef sous la porte, faute de lecteurs et de publicité, comprendre comment circule l'information dans notre société constitue un enjeu majeur sur le plan économique pour l'industrie des media ainsi que sur le plan sociétal pour le fonctionnement démocratique de notre société.

3 Enjeux et orientations

Un des enjeux de l'Observatoire Transmedia était de partir des besoins d'analyse exprimés par les chercheurs en SHS et les acteurs de l'information, en particulier l'AFP, et de collaborer tout au long des développements pour élaborer de nouveaux concepts, modèles et outils spécifiques à l'analyse du paysage informationnel. Cette collaboration transdomaines voulue et soutenue sur la durée du projet a impliqué une contrainte importante : la plateforme devait être utilisable au plus tôt dans l'agenda du projet pour avoir le temps d'évoluer avec les usages. Ainsi, le premier prototype a-t il été disponible 15 mois après le début du projet. La collaboration s'est articulée autour du périmètre de collecte, de la définition des critères d'analyse, des tests d'usages (évaluations des résultats et de l'ergonomie) et de l'analyse des biais du système. Enfin, le développement de certains modules de détection spécifiques, non envisagés au début du projet, ont été ajoutés au prototype pour élaborer des éléments de réponse à des questions clairement définies, comme par exemple la création du *taux de reprise* de l'information publiée.

Le périmètre de collecte des ressources a été établi au début du projet avec les partenaires SHS et l'Inathèque. Il a évolué en fonction des intérêts d'analyse exprimés au cours du projet. Si le projet cible l'actualité de manière générale, l'agenda médiatique, à savoir la présence de multiples élections dans la période considérée, a orienté la capture sur la thématique "politique". Pour les élections législatives par exemple, la collecte s'est étendue aux sites/blogs des députés et des sénateurs. Le corpus comprend les éditions de milieu de journée et du soir pour les chaines TV historiques (TF1 F2 F3), les éditions du soir (magazine ou journal) pour

F5, M6, Arte et Canal+, et les deux éditions principales pour les chaînes en continu I-TELE & BFM. Pour la radio, nous avons sélectionné d'une part les stations réalisant les plus fortes audiences ou présentant des plages d'information intéressantes : d'une part, les radios généralistes RTL, France Inter, Europe 1, RMC Infos, France Culture, et d'autre part, les radios d'information en continu : France Info, RFI, BFM et enfin Radio Classique, qui propose une matinale spécifiques sur l'actualité. Plus de 1800 sites, correspondant à 10 catégories, sont collectés à partir de leur flux RSS : environ 60 sites institutionnels, 260 sites de partis politiques, 650 sites de personnalités politiques, 70 sites de presse & agrégateurs ou portails de presse, 15 pure players, 14 radios, 17 TV, 19 syndicats, plus de 600 blogs politiques ou d'actualité générale et une classe « divers ». L'AFP étant partenaire du projet, l'intégralité des flux de dépêches est intégrée aux ressources. A la fin du projet, le 31 décembre 2013, le corpus comprend plus de 5 Millions de documents sources.

Les enjeux technologiques du projet sont liés au volume mais aussi à la diversité des sources d'information prises en considération. Il s'agit d'élaborer des référentiels de représentation homogènes des données, intégrant le traitement des modalités visuelles, sonores et textuelles. Dans un deuxième temps, la difficulté réside dans la mise en œuvre de différentes phases de fouille sur ces données automatiquement enrichies, potentiellement bruitées et incomplètes.

La plateforme distingue deux grandes phases de traitements : dans un premier temps, la collecte, l'analyse de contenu pour l'enrichissement des métadonnées associées aux documents sources, et l'indexation des différentes modalités, puis les modules de recherche interactive et de fouille de données.

4 Plateforme d'enrichissement et d'indexation

Les processus de recherche et de fouille reposent sur une étape de préparation des données. L'utilisateur final doit déterminer les critères d'observation des données et les observables qui en découlent. Une attention toute particulière doit être portée lors de la description des sources pour assurer que toutes les observables définies font bien l'objet d'un processus de description et peuvent être modélisées avec une notion de similarité ou de distance : il serait difficile de classer les images par couleur si elles n'ont pas été décrites préalablement avec un critère relatif à la couleur. Un travail en amont avec les utilisateurs experts a permis d'établir les critères et les modules de description et de détection nécessaires aux cas d'usages à satisfaire. Ainsi, la collecte et l'analyse des documents sources font-elles intervenir de nombreux composants logiciels qu'il faut coordonner pour assurer l'enrichissement du format pivot de métadonnées et son intégration dans la plateforme. La figure Fig1 illustre l'architecture du système et ses composants.

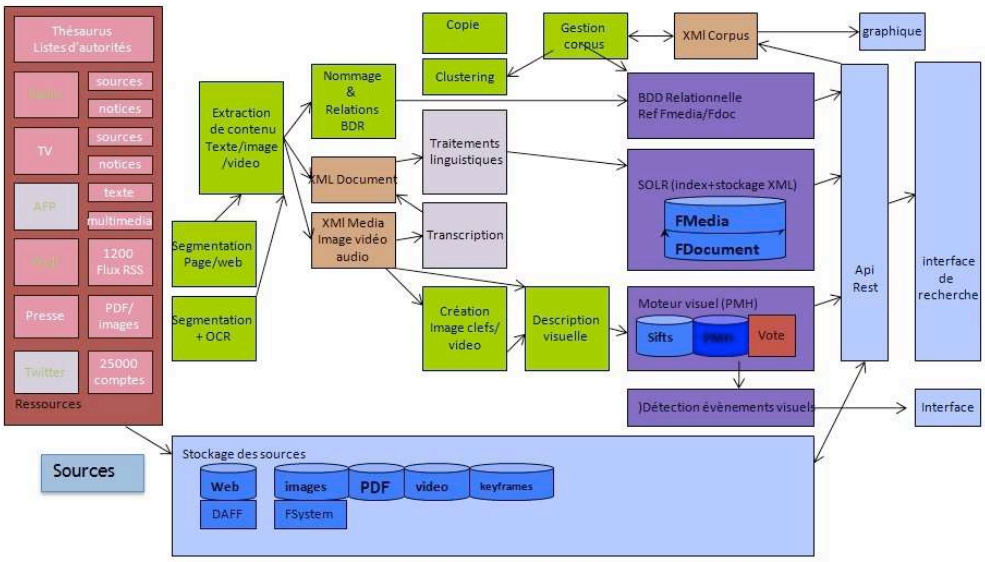


FIG. 1 *Architecture du prototype OTMedia*

Collecte

Trois systèmes de collecte et de stockage ont été mis en place en fonction du type de sources. Pour le web, la liste des flux RSS est tenue à jour dans un fichier Excel selon un formatage normalisé. Chaque ligne représente un flux, chaque modification du fichier donne lieu à une procédure de validation. L'outil de captation vérifie l'ensemble des flux de la liste une fois par heure, et relève les articles nouveaux. Chaque nouvelle URL d'article est ainsi envoyée dans la chaîne de captation DLWeb[3] pour que la page et l'ensemble de ses ressources internes (images, vidéos, HTML, etc.) soient captés puis stockés dans le format DAFF[4]. Les sources de la radio de la télévision et de l'AFP sont collectées sur des serveurs spécifiques deux fois par jour et stockées sur disque dur. En complément, le télétexte et les notices documentaires associées aux flux tv et radio captés sont collectées dans un délai de 3 semaines après la diffusion des programmes. Enfin, Twitter fait l'objet d'une capture spécifique développée par Syllabs et basée sur les travaux de (Cataldi & all 2010). Les tweets issus d'un noyau de 20000 tweetonautes spécialisés dans les news et sélectionnés par le CIM sont collectés. Un ensemble de procédures de filtrage et de normalisation du contenu est appliqué (sélection des tweets en français, suppression des items comme le smiley ou la ponctuation, extraction du hashtag…). Des graphes de cooccurrences de termes ou d'entités nommées sont construits sur des fenêtres temporelles et leur comparaison permet de faire émerger les nouveaux sujets, qui peuvent alors être utilisés comme requête pour compléter l'information. Nous noterons que les tweets ne font pas partie des entités (reportages, articles, billets de blog) manipulées par le système car leur granularité est plus fine. Ils sont stockés dans une

[3] Dépôt Légal du Web de l'Ina, service en charge de l'archivage du web français relatif aux media audiovisuels

[4] Format propriétaire du Dépôt Légal du Web de l'INA

base NOSQL et seront exploités uniquement pour statuer si twitter a diffusé une information avant les acteurs médiatiques traditionnels.

Données & métadonnées

La plateforme gère 4 types d'entités :

- les *documents sources*, qui correspondent à une dépêche AFP, une page web, un reportage TV ou radio, un billet de blog ou encore une déclaration officielle...
- les documents *multimédia,* fichiers images, audio ou vidéo contenus dans les *documents sources*
- les *fichiers de métadonnées* enrichies, qui décrivent les documents sources. On y retrouve principalement des informations sur l'origine des documents et des résultats d'analyse automatique. Les métadonnées sont uniformisées au sein d'un format propriétaire qui assure ainsi une représentation homogène des documents, quelle que soit leur provenance
- les *corpus*, ensembles de *documents sources*, qui peuvent être générés automatiquement ou manuellement.

Des *données externes* formant des connaissances nécessaires pour des traitements annexes ou pour les modules de fouille, peuvent être prises en compte par le système.

Les métadonnées relatives aux *documents sources* sont issues des données natives (date, flux, media…) et d'une suite de traitements informatiques permettant de segmenter le contenu, de l'analyser et d'en extraire les informations saillantes. Chaque type de flux est prétraité afin d'extraire un contenu homogène. Les pages web sont segmentées : le titre et les contenus textuels, images, vidéo ou audio sont extraits les sources tv et radio sont virtuellement découpées par reportage grâce aux timecodes présents dans les notices documentaires. Enfin, toutes les métadonnées disponibles (notices documentaires, métadonnées AFP, métadonnées des images ou vidéos sur le web, etc) sont intégrées au système.

Enrichissement automatique des metadonnées

Une phase d'enrichissement automatique des métadonnées, spécifique à chaque modalité, vient compléter le processus de création des métadonnées. Grâce à un langage qui permet de définir et de détecter des formes linguistiques spécifiques et l'usage de dictionnaires, Syllabs extrait les mots saillants, les entités nommées mais aussi les citations, avec leur auteur, lieu et date. Lors de l'étude des cas d'usages, les chercheurs et professionnels de l'information ont mis en avant l'intérêt de la notion de "reprise de l'information" qui établit la place, les relations et en définitive, le rôle des différents acteurs. Ces besoins ont mené au développement, par l'Ina, d'un module de détection de copie/plagiat et d'un module de détection de référencement (selon l'AFP, d'après les sources de Médiapart...) pour tracer les liens d'interdépendance entre les media.

Le laboratoire d'Informatique d'Avignon (LIA) a transcrit les données audio en procédant, pour chaque journée, à un enrichissement du vocabulaire. Une première passe de transcription est réalisée. Les segments audio correspondant à des mots de faible probabilité d'apparition (fiabilité faible de la transcription) sont retraités en incluant le vocabulaire spécifique extrait des sources écrites de cette journée. Si de « nouveaux mots » présentent une probabilité d'apparition plus forte, alors ils sont intégrés au vocabulaire. Enfin une dernière passe de transcription est réalisée. La phase d'enrichissement du vocabulaire est particuliè-

rement intéressante pour les entités nommées. En effet, les personnes et les lieux peuvent prendre des formes très éloignées du français, amenant à des transcriptions très fantaisistes.

L'indexation textuelle est réalisée avec SOLR, plateforme de recherche fulltext Open Source, développée sous la bannière de l'*Apache Software Foundation*[5]. SOLR permet de créer des solutions de recherche répondant à la plupart des besoins courants de recherche d'information, sans avoir besoin de développements spécifiques.

La description visuelle consiste à générer un ensemble de vecteurs correspondant aux points visuellement saillants de l'image. Ces vecteurs sont constitués de dérivées multi-échelles locales de l'intensité des pixels (Lowe 1999). Plus une image sera finement décrite, plus l'accès aux détails de cette image sera envisageable. Pour atteindre nos objectifs, à savoir la recherche d'une instance d'objet de taille réduite dans une image, la description visuelle pourra mettre en œuvre plus de 500 points saillants par image. Les vidéos sont segmentées en images de contenu proche. Ainsi, un plan panoramique ou un zoom seront-ils représentés par une série d'images dès lors que la différence de contenu entre deux images successives dépasse un seuil donné. Ces images sont ensuite traitées comme les images fixes. Les structures d'index forment le noyau des moteurs de recherche en contrôlant leur efficacité. Pour la partie visuelle, déroulons le processus pour en percevoir les enjeux : la description d'un corpus de 1000 heures de vidéo correspond environ à 90 Millions d'images. Si on décrit chaque image avec 500 descripteurs, le moteur de recherche doit trouver le voisinage d'un descripteur parmi quelques 50 Milliards de descripteurs ! A l'aide d'un processus non dédié et sur une seule machine, il mettrait … quelques mois pour retourner la réponse à une requête simple. Le rôle du processus d'indexation est d'organiser les données et d'associer ce rangement à des processus d'accès dédiés pour accélérer la recherche de voisinage. Si les moteurs de recherche textuels ont déjà été largement explorés et exploités, c'est qu'ils présentent, outre leur importance fondamentale pour les usages, une spécificité en terme de répartition des vecteurs dans l'espace. Les documents textuels sont décrits par des vecteurs de mots, mais le langage n'étant pas une association aléatoire de termes, la répartition des vecteurs dans l'espace de description suit une distribution particulière, permettant de traiter la recherche de manière pseudo-locale. Cette propriété conduit à contourner en partie la problématique posée par les gros volumes. L'index visuel permettant d'effectuer une recherche interactive dans ces milliards de descripteurs est un cœur de technologie issu de la recherche et développé conjointement depuis près de 10 ans par l'INRIA et l'INA (Joly and all 2008). Cette technologie de moteur de recherche vectoriel repose sur l'idée d'accès localisés probabilistes multiples à une table de hachage grâce à un apprentissage probabiliste. Cette méthode, qui permet en outre de réduire considérablement l'espace mémoire nécessaire au stockage des clefs, est actuellement une des méthodes les plus efficaces pour la recherche approximative d'items par similarité dans des espaces de grande dimension. Un autre avantage conséquent de ce moteur de recherche réside dans sa souplesse d'usage. En effet, le système peut prendre en considération plusieurs types de mesures de similarité (L1, L2,

[5] Créé initialement par des ingénieurs du site CNET Networks, le moteur de recherche Apache Lucene, reconnu comme un des meilleurs logiciels dans son domaine, a été rendu open source et donné à la *Apache Software Foundation* en 2006. SOLR ajoute la dimension plateforme à Lucene : il comprend un serveur d'indexation et de recherche de type REST entièrement configurable.

Hamming…), plusieurs types d'entrées (vecteurs binaires, réels, données non vectorielles…), plusieurs familles de fonctions de hachage (projections aléatoires, à base de noyaux, optimisées comme PCA ou LDA) ou encore plusieurs types de requêtes (plus proches voisins, items dans un rayon donné…). Enfin, un paramétrage permet de régler la qualité des résultats, sachant que plus les résultats seront exhaustifs, plus le temps de calcul sera important (progression non linéaire). Les images ou les parties d'images requêtes sont décrites de la même façon que les images de la base et les vecteurs associés à leurs points saillants constituent les requêtes dans l'index. Les items résultats sont obtenus par un vote sur les images comprenant des points saillants proches et présentant des caractéristiques géométriques similaires entre elles.

Enfin, une base relationnelle SQL est nécessaire pour gérer plus simplement les informations nécessaires aux modules d'affichage et de fouille, à savoir le référencement des documents, les relations père/fils entre *documents sources* et *documents media*, la gestion des corpus et des données externes. Pour satisfaire aux contraintes de développement décentralisé et évolutif propres aux projets de recherche, la plateforme OTMedia propose un framework Java dédié. Des webs services transverses, qui s'appuient sur des technologies éprouvées (Jaxb et Hibernate), fournissent les outils supports pour l'indexation et l'accès aux données (nommage, accès aux sources, accès aux métadonnées, accès hiérarchique aux composants d'un document).

5 Le prototype OTMedia: vers la fouille interactive

La fouille de données, data mining, ou encore extraction de connaissances à partir de données *a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques*[6]. Dans le cadre d'OTMedia, la fouille de données est utilisée dans un premier temps pour faire émerger des groupes, des tendances, des structures ou des mouvements, qui permettent de comprendre les phénomènes à une échelle macro alors qu'ils interviennent dans des ensembles trop grands pour être appréhendés par une approche humaine. Dans le contexte professionnel de l'INA, pour des raisons de maintien de qualité de la base, nous sommes amenés à favoriser des processus semi-automatiques laissant une place importante à l'intervention humaine. Le contexte d'OTMedia, lié la présence d'utilisateurs experts et donc la nécessité de produire un outil fiable, nous a conduit à entériner cette démarche. Notre choix de développement s'est donc orienté sur un premier prototype de recherche, transmodal, interactif et composé de fonctionnalités relativement simples à manipuler et réalisables pour les partenaires technologiques du projet à 15 mois de son début. L'idée était de permettre aux chercheurs en SHS d'obtenir, avec un peu de manipulation et d'efforts de validation, rapidement des "vues fiables" sur des données de masse transmedia inaccessibles auparavant. Ces choix, d'interactivité, de visualisation et de manipulation ont été renforcés par l'attention donnée à la validation des résultats. En effet, la phase de validation implique la construction de vérités de terrain sur les données, tâche facilitée par les manipulations de corpus mise en œuvre par le prototype.

[6] Définition wikipedia

Interfaces de recherche, de consultation et de manipulation de corpus

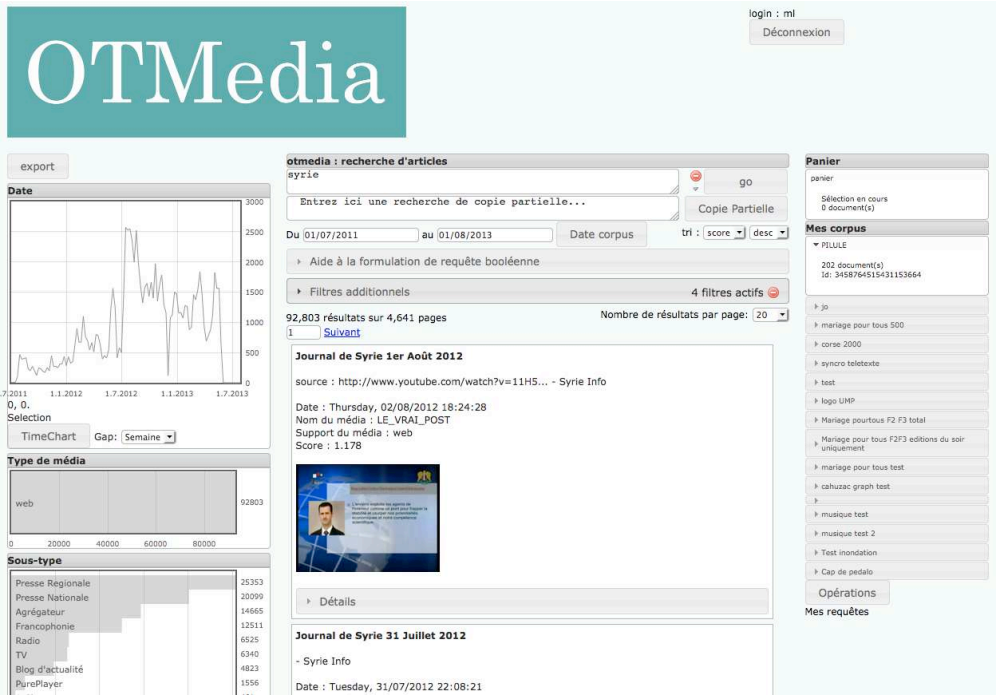


FIG. 2 – *Interface du moteur de recherche OTMedia.*

Le prototype se présente comme un moteur de recherche dédié : l'utilisateur compose une requête simple, à l'aide de la barre de recherche standard, ou complexe, grâce aux widgets d'*aide à la formulation de requête* et de filtrage *filtres additionnels* (cf fig 2) qui lui permettent, par exemple, d'exclure des combinaisons de mots ou de sélectionner des groupes de media ou de support spécifiques. Les documents résultats sont affichés en liste, chaque item de la liste étant cliquable pour accéder au *document source* et au *fichier de metadonnées enrichies* correspondant (cf fig 3), afin que l'utilisateur puisse, à tout instant, valider le contenu de ce qui lui est présenté. Outre la répartition chronologique, des *tableaux résumés* des résultats sont disponibles sur la gauche de l'interface. Ils permettent à l'utilisateur d'avoir un retour sur l'ensemble des résultats affichés : le nombre de résultats pour chaque support, pour les 20 media les plus représentés, le nombre d'occurrences des mots les plus saillants, des personnalités, des acronymes ou des lieux les plus cités. Ces tableaux interactifs permettent d'obtenir une première vue d'ensemble des résultats, de détecter des anomalies ou d'affiner les requêtes. Sur la droite sont affichés les corpus créés par l'utilisateur. Des fonctionnalités de création, de suppression d'items, de fusion, d'intersection et de visualisation sont disponibles pour permettre à l'utilisateur d'explorer en détail des parties d'intérêt du corpus général.

FIG. 3 – *Exemple de consultation d'une source télévisuelle de France 2:la vidéo est disponible, le mot requête* ***Syrie*** *est surligné en jaune pour la partie textuelle et des curseurs marquent les apparitions du mot Syrie dans la bande audio pour se positionner directement sur les passages intéressants.*

La recherche visuelle

Les recherches visuelles s'effectuent via une deuxième interface, activée en sélectionnant une image dans l'interface générale. Une version agrandie de l'image est présentée à l'utilisateur, qui requête soit l'intégralité, soit une partie sélectionnée à l'aide de la souris (partie grisée sur la figure 4). De par le choix d'une description basée sur un assemblage de similarités locales insensibles à la couleur, le moteur retrouve des *copies* d'images, mais aussi et surtout des images *similaires* ou *partiellement similaires*. Nous qualifierons de *similaires* des images présentant des contenus visuels proches mais pouvant contenir des déformations, des occultations ou des changements de couleur. Ces propriétés conduisent alors à de multiples usages de la détection visuelle. L'association d'un ensemble de *copies* visuelles avec les documents sources dont elles sont issues permet d'étudier la diffusion d'une image spécifique dans les media : quand et par qui a–t'elle été utilisée? La sélection de la palme située à droite des orateurs à l'assemblée nationale rassemble les images de députés au perchoir, assez nombreuses au demeurant (fig 4)! Enfin la sélection de petits objets comme les logos, nous permet de regrouper toutes les images d'une même manifestation culturelle ou sportive disposant d'un "habillage" (Mostra de Venise, UsOpen...cf fig 9), les illustrations des campagnes des partis politiques (front de gauche, UMP...), et peut mener à des utilisations plus commerciales.

FIG. 4 – *A gauche, l'image sélectionnée initialement. Au milieu, notez la sélection en gris de la palme à la gauche du député, à droite une partie des résultats retournés par la requête issue de la sélection de la palme.*

Courbes temporelles comparatives

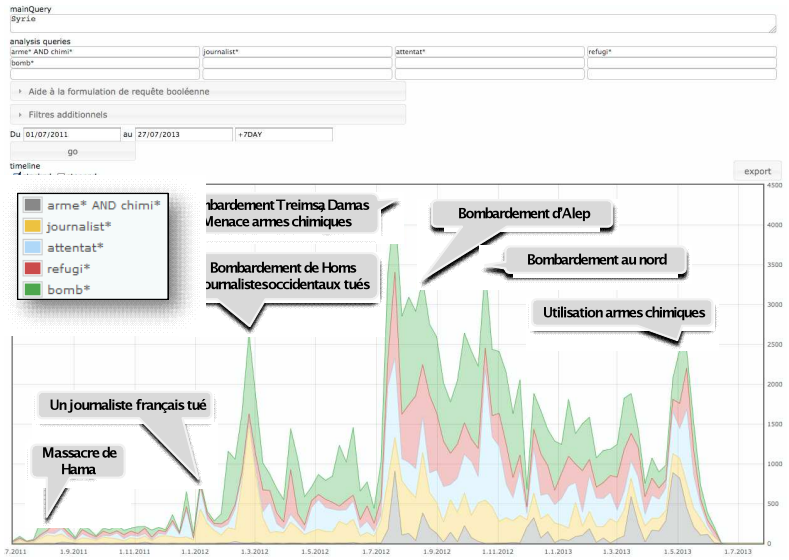
La troisième interface produit des chronologies comparatives d'occurrences, de cooccurrences de termes ou de champs structurants. Concrètement l'utilisateur spécifie via l'interface une liste de 9 items au maximum (au delà les courbes ne sont plus discernables), termes, noms de personnes, acteurs médiatiques ou thématiques de son choix. Après validation de cette requête, les courbes affichent, pour chaque item, le nombre de documents qui le contiennent. Par exemple, l'utilisateur pourra comparer, sur une période donnée, combien de fois sont cités les noms des candidats à l'élection présidentielle. L'utilisateur peut aussi spécifier une requête cible. Sur l'exemple précédent, la requête *éducation* affichera le nombre de fois où le nom de chaque candidat est associé au terme *éducation* tout au long de la campagne. Dans un troisième temps, des filtres sur les champs structurants permettent d'observer les mêmes courbes pour un acteur médiatique spécifique, voir le comparer à d'autres acteurs ou à un sous ensemble des media. On peut alors potentiellement répondre à la question : est-ce qu'un acteur médiatique cite plus souvent la question de l'éducation traitée par tel candidat que les autres acteurs médiatiques?

Cette interface permet aussi de visualiser des chronologies d'évènements sur une période donnée en choisissant, comme item de requête, des identifiants de corpus *évènements* automatiquement ou manuellement générés.

Un exemple sommaire d'analyse chronologique : le conflit Syrien.

Les figures 5A et 5B sont relatives au conflit syrien de juillet 2011 à juillet 2012. La collecte ayant subi deux augmentations de périmètres sur la période étudiée, nous ne commenterons pas les chiffres bruts mais dégagerons des tendances relatives. La dimension internationale du conflit (cf fig 4b), avec une position très forte de l'ONU est confirmée par l'amplitude des cooccurrences de Syrie avec ONU, puis USA, Russie, et Chine. On notera l'apparition de la "ligue arabe" en automne 2011, avec une augmentation de son implication jusqu'au printemps 2012 suivie d'un retrait de plus en plus prononcé depuis. De la même façon, les courbes relatives au Conseil National Syrien illustrent bien son évolution: c'est un acteur qui

apparait en octobre 2011, et prend part aux négociations internationales pendant près de 6 mois, puis sa représentativité étant mise en cause, son influence décroit et il intègre une "coalition nationale" plus large en octobre 2012. Les pics de la figure 4a, marquent les principaux évènements du conflit avec les attaques par l'armée de Ham (juillet 2011), de Homs (février 2012), de villages du centre du pays (juillet 2012) et les réponses de l'opposition sous forme d'attentats. Ce schéma nous permet d'observer aussi que les média français commencent à utiliser les mots d'armes chimiques dès le 23 juillet 2012 lorsque Bachar el Assad reconnait posséder des armes chimiques et menace de les utiliser en cas d'agression étrangère. Depuis cette date, le sujet est de plus en plus présent, jusqu'à l'envoi d'une mission d'observation de l'ONU qui confirmera l'usage des armes chimiques en septembre 2013...



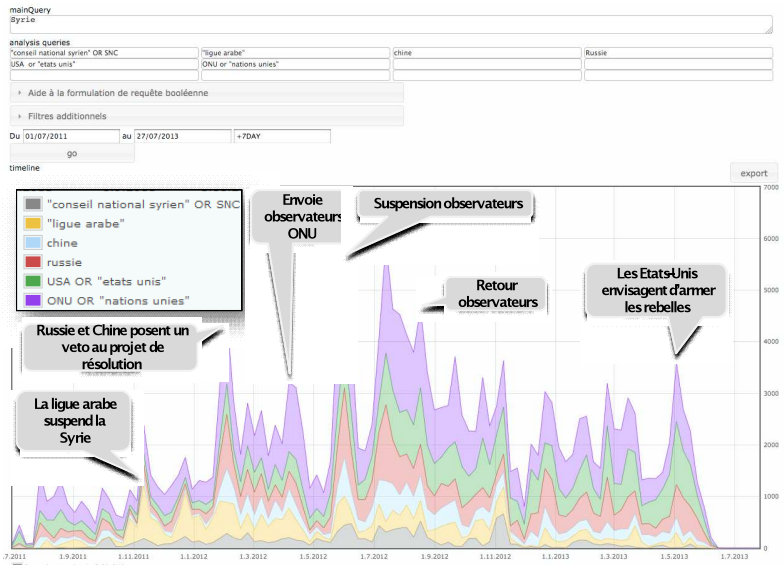


FIG. 5A & 5B – *la requête Syrie avec deux vocabulaires développant deux points de vue sur le conflit. La figure 3a projette l'amplitude du traitement du conflit sur un vocabulaire lié à l'action alors que la figure 3b prend en compte les grands acteurs diplomatiques du conflit.*

Enfin, la chronologie détaillée d'un corpus est visualisable grâce à une projection des documents sur une vue temporelle (cf fig 5). Les chercheurs en SHS Éric Lagneau, Sylvain Parasie et Franck Rebillard[7] se sont intéressés à la propagation de l'information via l'étude de deux événements de la campagne présidentielle 2012 : l'apostrophe de « capitaine de pédalo » lancée par Jean-Luc Mélenchon à François Hollande et l'affaire du logement parisien de Jean-Pierre Chevènement. « Capitaine de pédalo » apparaît le 12 novembre 2011 sur le site de Jean-Luc Mélenchon, avant d'être repris par le JDD, par les agences de presse AFP et Reuters, par RTL puis la presse la radio et la télévision. Les blogs ne sont pas en reste sur cette affaire puisque le "blog de Jocelyne", très réactif pendant la période électorale, est dans les premiers media à réagir à cette invective et à la polémique qu'elle suscite. Au total, 339 résultats sont trouvés dans la base (à noter que la source du JDD n'a pas été collectée dans le corpus mais était citée par les dépêches d'agences). L'affaire de l'appartement de Jean-Pierre Chevènement montre une propagation en deux temps puisque l'affaire avait été dévoilée par l'AFP en juin 2011, et c'est l'Express qui la remet à l'ordre du jour à l'automne 2011.

[7] respectivement à l'AFP, au LATTS (Ecole des Mines) et au CIM (Université Paris 3)

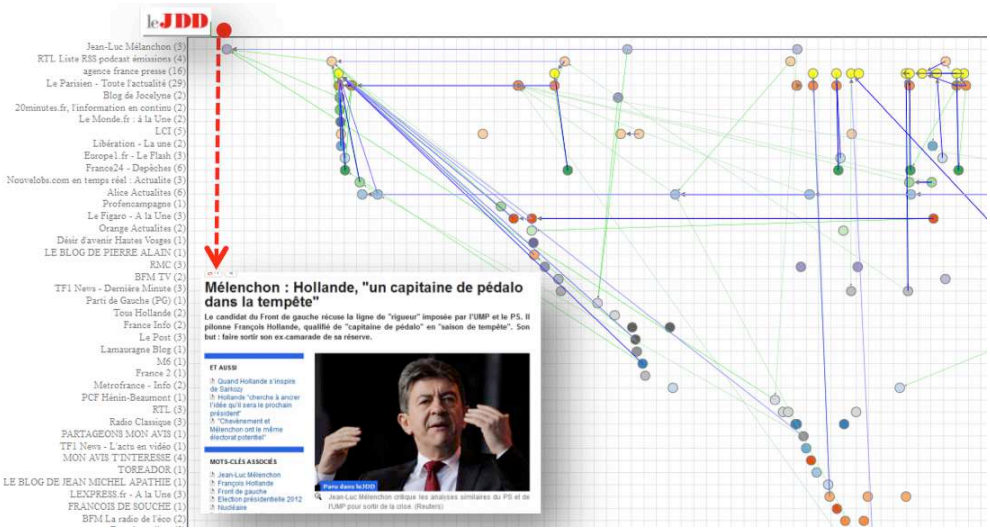


FIG. 6 — *Vue Chronologique détaillée "Capitaine de Pédalo"*

Processus interactif d'analyse & de validation

Toutes les interfaces du système communiquent et sont accessibles de manière à enchainer les tâches nécessaires à une analyse. La requête courante peut servir ainsi de requête globale pour la vue comparative ; la sélection d'une période de temps et d'une courbe sur la vue comparative effectue la requête équivalente sur l'interface de recherche textuelle, permettant ainsi de consulter les documents relatifs à un pic. Tous les corpus, qu'ils soient automatiquement ou manuellement générés à partir des interfaces de recherche textuelle ou visuelle sont consultables, visualisables et manipulables.

Le sujet de l'étude réalisée pour l'ONU-AOC était l'analyse de la thématique *immigration* en période électorale. Pour cette étude, intervenue avant la finalisation du détecteur d'événement, nous avons procédé à une fouille interactive via l'interface de manière à neutraliser les biais du prototype (cf section 7). Nous nous sommes basés sur les ressources de l'AFP pour établir une liste de vocabulaire pertinent à interroger et cibler ainsi les évènements majeurs. Les *tableaux* de gauche dans l'interface textuelle sont une aide pour ces tâches car ils offrent un retour sur les résultats courants, permettant de vérifier la pertinence du lot de documents, de découvrir du vocabulaire ou de détecter des anomalies (cf fig 7 & 8). Dans cette optique, tous les tableaux sont cliquables afin de pouvoir facilement restreindre ou élargir des requêtes. Nous avons ensuite étendu ces corpus à l'intégralité de la base. Nous avons pu ainsi montrer que pour l'élection présidentielle de 2012, l'immigration constituait bien un sujet de campagne, puisque les media en parlaient plus que sur les autres périodes; mais que, contrairement aux deux précédentes élections présidentielles, les évènements rapportés étaient plutôt d'ordre politique que sur la sécurité. Ils concernaient le vote par le sénat de la loi sur le vote des étrangers aux élections locales, les discussions sur l'immigration économique, les problèmes de l'immigration au niveau européen…

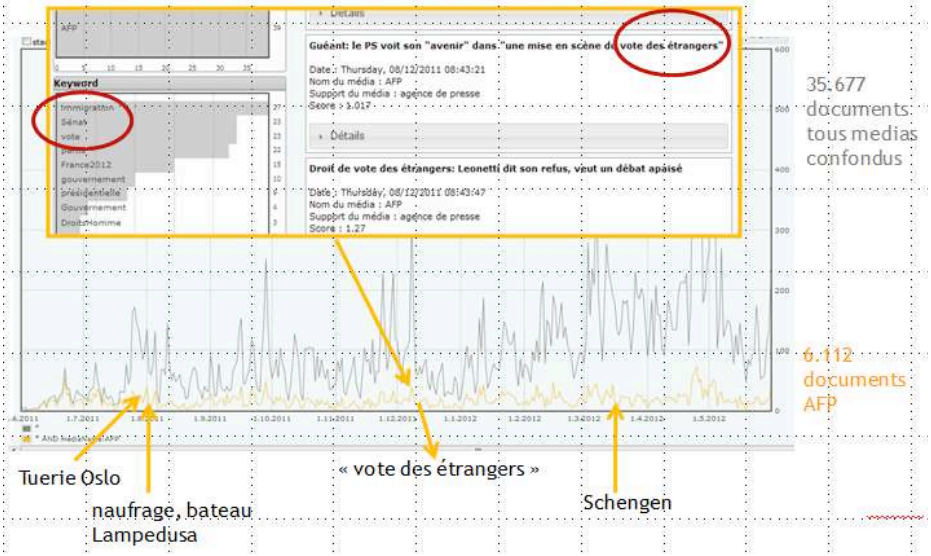


FIG. 7 – Découverte des thématiques traités sur l'immigration dans les medias par l'analyse des évènements et du vocabulaire sur les fils de dépêches AFP

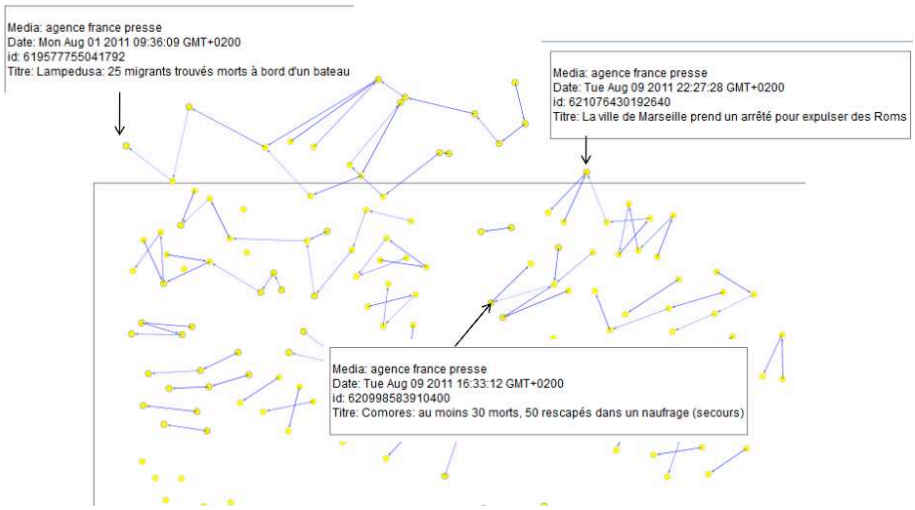


FIG. 8 – visualisation des fils de dépêches AFP correspondant aux sélections de la figure 7

6 Emergence d'objets : détection d'évènements textuels et d'évènement visuels

Une des tâches fondamentales de la fouille de donnée est de regrouper et labelliser des ensembles d'items "proches" afin de faire émerger des objets de "plus haut niveau" synthétisant la masse des items. L'idée sous jacente est de réduire le volume des données en les classifiant et les hiérarchisant pour tenter d'en percevoir le contenu. En effet, si le moteur de recherche répond à la question, la base contient elle des éléments proches de ma requête? La fouille tente de répondre à la question : que contient cette base ? Pour percevoir les enjeux liés à ce changement de paradigme, examinons les différentes étapes de ce processus d'émergence.

Dans le contexte du projet, nous avons étudié deux types d'évènements spécifiques correspondant principalement à des regroupements d'objets similaires sur les modalités textuelles et visuelles. Une des caractéristiques des documents liés à l'actualité est sa répartition temporelle qui suit une *loi de Poisson*. Ces évènements sont détectés sur des fenêtres temporelles glissantes afin de minimiser le bruit et garantir des volumes traitables avec les moyens de calculs à disposition.

Evènement visuel

La première étape consiste à générer des liens de proximités entre items. La création d'un lien de proximité entre deux objets s'effectue en réalisant une requête sur le moteur puis en modélisant un lien de similarité entre les k premiers résultats ou les résultats dans un rayon donné. Notons que si la base contient N objets, il faudra réaliser N requêtes... ou établir des stratégies. Pour cette tâche encore, les cas textuels et visuels diffèrent: un texte dispose d'une représentation en vecteur de mots saillants adaptée quelle que soit la base de documents. C'est donc le choix ou l'adaptation de k ou du rayon en fonction du contexte qui, bien qu'ayant fait l'objet de nombreux travaux de recherche reste une problématique délicate. Pour l'image, il faut trouver l'ensemble des mots visuels de chaque image signifiants par rapport à la base considérée. Détecter les objets visuels les plus fréquents de la base supposerait donc de requêter toutes les parties d'images des images du corpus (soit un nombre infini), pour être sûr de trouver celles qui sont effectivement *fréquentes*. Les travaux de recherche dans ce domaine visent à établir des stratégies pour minimiser autant que possible le nombre de requêtes nécessaires pour créer des liens de proximités signifiants (Letessier & all 2012). La bonne nouvelle est que plus les objets sont fréquents, plus on a de chance de les sélectionner lorsqu'on choisit une partie d'image au hasard. Pour donner un ordre d'idées, la détection des objets les plus fréquents dans un corpus de 5000 images mis à disposition par l'université d'Oxford pour servir de benchmark à la communauté scientifique internationale a nécessité plus de 200 millions de requêtes. Ce constat met en évidence l'importance d'un moteur visuel performant pour aborder les problématiques de fouille. En remarque finale, nous ajouterons que le nombre de requêtes nécessaires à l'émergence croit beaucoup plus vite que le nombre d'items de la base. Lors du projet, les volumes de l'ordre de 20000 items ont été traités. Lorsque les liens de proximités sont générés, des algorithmes d'agrégation (clustering) classiques comme Louvain ou MCL peuvent être utilisés pour établir les groupes d'items proches. L'item représentatif ou l'étiquetage des groupes sont réalisées par des processus manuels,

semi manuels ou automatiques classiques (centralité dans le graphe de similarité, max d'occurrence de termes...).

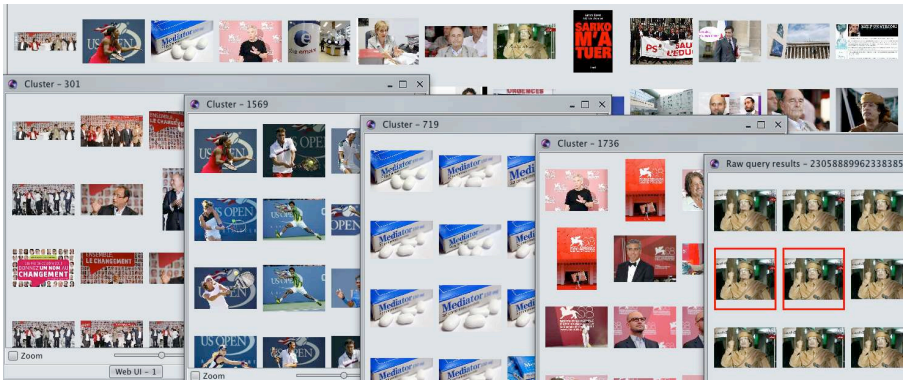


FIG. 9 – *les évènements médiatiques la semaine du 30 août au 7 septembre 2012: la campagne des primaires socialistes, l'USOpen, l'affaire du Mediator, la Mostra de Venise, et l'intervention de Khadafi à la télévision Libyenne*

L'évènement visuel du jour, de la semaine et du mois repère avec les techniques précédemment évoquées les images les plus fréquentes sur la période considérée. Néanmoins, il suffit d'une campagne de publicité agressive ou d'un habillage bien utilisé sur un media actif pour perturber les résultats. Une contrainte a donc été rajoutée au niveau de la création des liens de proximité: les documents dont sont issues les images résultats doivent provenir de différents supports ou acteurs media (cf fig 9). Néanmoins le mois de décembre 2011 présentait en résultat la publicité Benetton (cf fig 10). Effectivement cette publicité ayant fait l'objet de recours de la part du Vatican a été utilisée en illustration 49 fois, devenant elle même un évènement médiatique!

FIG. 10 – *l'évènement de la semaine de 14 au 21 novembre 2012: la publicité Benetton devient un évènement médiatique...*

Evènement textuel

La détection d'évènements médiatiques à partir des documents textuels[8] s'effectue en plusieurs phases. Face à la particularité de notre corpus, nous faisons l'hypothèse qu'un évènement médiatique sera nécessairement traité à un moment ou un autre par l'AFP. La première phase consiste à repérer quelles sont les dépêches AFP qui sont le plus reprises. Elles nous servent de germe pour définir les périmètres des principaux évènements médiatiques d'une journée. Ensuite, à partir d'une analyse des mots saillants de ces agrégats, nous complétons les évènements en y ajoutant les documents provenant des autres media. Enfin dans une dernière passe, les groupes sont fusionnés dans le temps. Les résultats sont en cours d'évaluations. Nous attacherons notamment une importance particulière à l'évaluation de nos détections en tenant compte de la disparité des éléments textuels à notre disposition : articles de presse, transcriptions audio, télétexte, notes documentaires. A notre connaissance, c'est le seul corpus existant qui présente une telle diversité. Enfin, des améliorations sont prévues pour augmenter la qualité des détections aussi bien en terme de périmètre des évènements que de granularité de détection.

Autres outils

Pour faciliter la perception et la validation des évènements par l'utilisateur, une interface de fouille visuelle a été développée avec le Labri (Renoust & all, 2013), permettant d'évaluer la qualité des groupes d'articles « proches ». Cette interface permet à l'utilisateur d'identifier et de tracer les éléments de vocabulaire partagés par les articles et d'établir un retour sur la cohésion des groupes formés. Elle permet aussi d'analyser comment un media traite d'un sujet de par la topologie du réseau de vocabulaire associé. La figure 11 montre le vocabulaire associé au traitement du décès de Michael Jackson par TF1, France 2, France 3 et M6. TF1 et France 2 ont beaucoup relayé l'évènement sous toutes ses formes, ils commentent autant sa vie (enfants, scandales, chanson...) que le décès ou les circonstances du décès: le graphe des termes a une topologie "en étoile", caractéristique des évènements *multifacettes*. Les graphes de vocabulaire de M6 et France 3 sont très compacts, illustrant un traitement factuel et très homogène de l'affaire.

[8] On fait généralement référence à ce domaine de recherche sous l'appellation *Topic Detection and Tracking* (Allan & all 2002) pour un aperçu des approches.

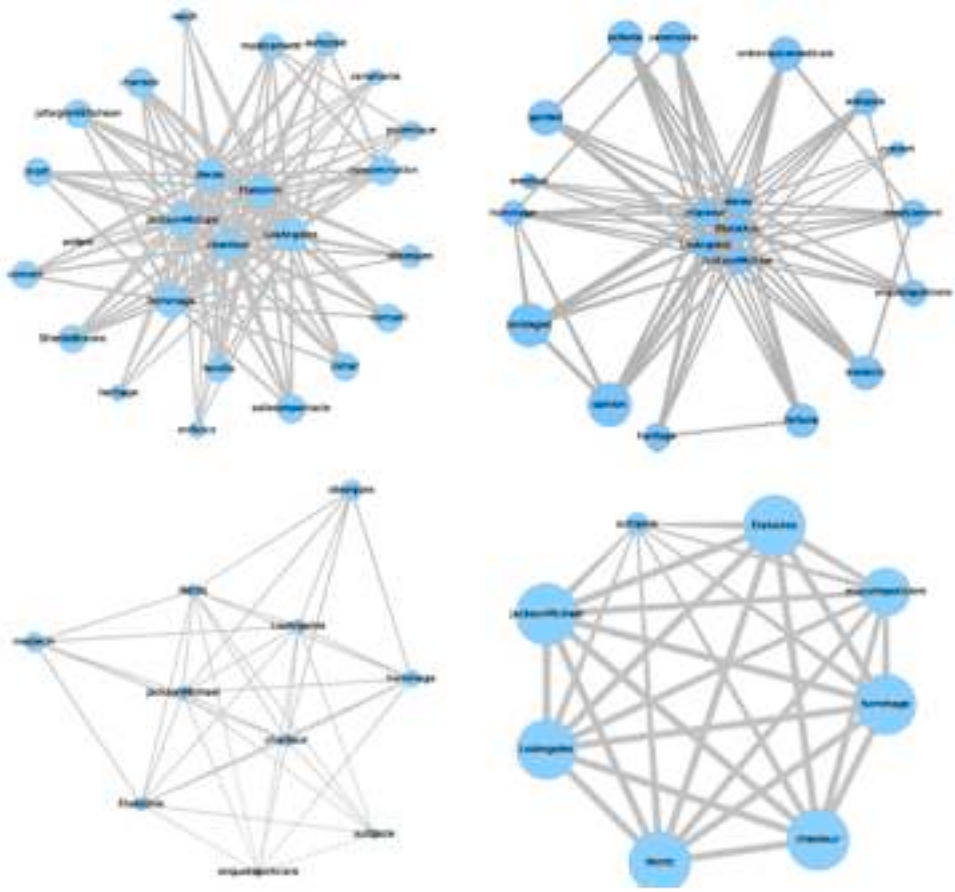


FIG. 11 – Les graphes de vocabulaires associés à la mort de Mickael Jackson par TF1, France 2, France 3 et M6 (de gauche à droite et de bas en haut). La taille des termes reflète leur occurence.

Des outils complémentaires ont été développés pour répondre à des questions spécifiques. Ainsi, l'AFP était intéressée par les taux de reprise de ses dépêches sur les différents media sur l'année 2012. La figure 12 montre le pourcentage d'articles en fonction du taux de copie des dépêches pour 5 media en ligne. Le taux de copie est la proportion d'un article qui est directement issu d'un autre article et n'ayant subi aucune modification. Elle montre que pour ces media, le taux de reprise illustre parfaitement leur positionnement dans le paysage. Ainsi, pour les agrégateurs, on constate que 70 pourcents de la diffusion d'Orange provient du flux de l'AFP, puisque la courbe est plate (le taux de copies de 0,8 plutôt que 1 est dû à l'habillage ou aux erreurs de segmentation) alors que le taux de Yahoo! est proche de zéro, ce qui s'explique par le fait que Yahoo! s'est désabonné de l'AFP en 2012. Rue 89 produit des contenus originaux, avec très peu de reprise de très petites tailles, ce qui correspondrait plutôt à la présence de citations. Environ 20 pourcents de la production de 20 minutes représentent des dépêches AFP, alors que le Huffington Post, avec une courbe de pente descendante,

emprunte aux dépêches AFP une partie de son contenu mais complète le plus souvent l'information dans des articles plus longs

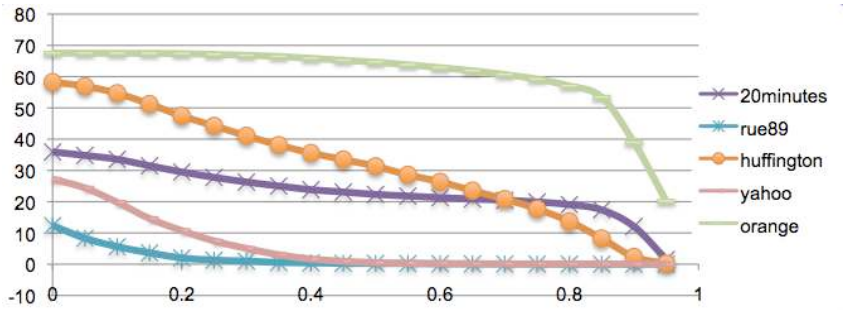


FIG. 12 – *Pourcentages de la production en fonction du taux de reprise des dépêches AFP pour 5 media en ligne...*

A partir d'un export des résultats du projet, Dario Companio[9] et Franck Rébillard ont étudié, grâce au logiciel TLab, les phases diachroniques de l'affaire Merah. Cette étude montre l'intérêt d'intégrer au projet des outils de fouille comme l'analyse factorielle basée sur le vocabulaire. TLab prend entrée les mots saillants et les groupe sous forme de mots de la même racine. Le vocabulaire est réduit de 130000 mots à 5570 formes. Puis une sélection semi automatique permet de sélectionner les 750 mots les plus pertinents. Les matrices de cooccurrence de ces mots sont créées et approximées grâce à l'analyse factorielle. Les deux dimensions les plus caractéristiques des données issues de l'analyse factorielle forment les axes d'une projection bidimensionnelle dans laquelle les mots les plus co-occurrents se retrouvent proches. Cette représentation permet de suivre l'évolution du vocabulaire au cours du temps et ainsi, les différentes phases de l'évènement. 5 thématiques ont été associées au vocabulaire et projetées sur un axe temporel. La thématique "tuerie de Toulouse" est majoritaire les 19 et 20 mars et est présente jusqu'au 23 mars. L' "Assaut du raid" et la "mort de Merah" augmentent et déclinent sur 3 jours, les 21, 22 et 23 Mars. La "discussion publique sur le contexte sociofamilial" s'étend du 23 au 26 mars et réapparait les 29 et 30. Enfin, l'"inhumation de Merah" a occupé majoritairement l'actualité du 28 au 29 alors que le "traitement politique" est toujours présent dans une petite proportion sur l'intégralité de la période. La projection sur les différents media de ces thématiques permet de donner une idée de l'angle éditorial de chaque support. Globalement, la radio rapporte de manière importante le "contexte sociofamillial", la télévision privilégie l'"assault du raid", alors que la répartition du web se rapproche de l'AFP, et présente un traitement plus équilibré, accordant la plus grande place à la "tuerie" et la plus petite au "traitement politique" de l'affaire.

7 Evaluation, bilan et perspectives

La collaboration trans-domaines est extrêmement intéressante et riche, même si elle suscite parfois quelques incompréhensions de part et d'autre. Dans un premier temps, les concepts complexes des SHS sont rarement directement modélisables par des ensembles de

[9] CIM

critères ou mesures bien définis que peuvent manipuler les algorithmes. Ainsi, le concept d'"évènement médiatique" est-il resté un sujet de discussion entre les partenaires pendant une grande partie du projet. La définition philosophique d'évènement médiatique et celle retenue pour le projet "documents se rapportant à un même fait, à son évolution ou aux commentaires qui en découlent", sont assez distantes! Néanmoins, la prise en considération des multiples dimensions d'analyse nécessaires aux sciences humaines a été très riche car elle a mené à des enchaînements de traitements linguistiques et visuels plus complets que ceux prévus au départ du projet. La manipulation du prototype par des utilisateurs concernés a permis de mettre en place une démarche de conception centrée utilisateur tout au long projet[10]. Cette boucle a permis une évolution parallèle et itérative du prototype et des usages. L'évaluation porte en premier lieu sur l'utilité du système mais aussi sur son utilisabilité[11]. Des recommandations ont été recueillies auprès des utilisateurs pour chaque version du prototype. C'est ainsi que plusieurs fonctionnalités ont été ajoutées au niveau de l'étude des besoins, pendant ou après les tests. La gestion de corpus de résultats, l'export de données, le suivi de citations ou la détection de copies partielles textuelles en sont des exemples. Par ailleurs, l'analyse de la validité des résultats a permis d'opérer une amélioration en qualité de certains modules du système. Entre autres, la segmentation des pages web a été modifiée pour améliorer la suppression des informations non utiles, des données supplémentaires ont été ajoutées aux flux à capter pour catégoriser au mieux les résultats ; certains blogs, de par leurs pratiques éditoriales non conventionnelles, créaient trop de bruit et ont été supprimés. Enfin, les recommandations des utilisateurs sur la manipulation du prototype ont aussi été prises en compte pour en améliorer l'utilisabilité. Certaines données sont devenues interactives, les requêtes sont surlignées dans les documents, les interfaces ont été reliées entre elles pour enchaîner logiquement les opérations liées à une tâche d'analyse… Les tests utilisateurs ont permis d'étudier l'équilibre entre l'automatisation technologique et le contrôle qu'il faut laisser à l'utilisateur.

En dernier lieu, l'utilisation du prototype par des utilisateurs experts a permis de mettre en évidence les deux types de biais du système OTMedia : les biais dus aux traitements technologiques et ceux liés aux pratiques éditoriales des media. Les plus grands biais du prototype apparaissent en effet pour les contenus textuels du web. Pour certains titres ou blogs, la segmentation n'est pas totalement fiable et supprime ou augmente le contenu des pages avec, par exemple, les liens sur les articles du jour ou les articles connexes. D'autre part, certains acteurs republient plusieurs fois la même page avec quelques mots de différences. Ces deux cas de figure ont pour effet l'absence ou la surreprésentation de certains acteurs dans le paysage global. Certaines dates de production sont fantaisistes (2025...), d'autres sont étonnantes: ainsi certains titres diffuseraient les dépêches intégrales de l'AFP avant même l'AFP! La désambigüisation, la fusion et la hiérarchisation des entités nommées est une fonction d'amélioration de la qualité absolument prioritaire pour l'évolution du prototype : il s'agit d'associer automatiquement "Angela Merkel", "Merkel", "la chancelière allemande" à une même personne ou Damas à la Syrie.

---

[10] Cette démarche repose sur la norme ISO 13407 qui définit la mise en œuvre d'une boucle itérative d'analyse, conception et évaluation.

[11] Norme ISO 9241-11 : L'utilité est la capacité d'un système à réaliser une tâche pour laquelle il a été conçu et repose sur la qualité des résultats. L'utilisabilité est la capacité du système à être facilement utilisé par une personne. Cette notion englobe plusieurs critères comme la performance, la satisfaction et la facilité de réalisation.

Un des aspects fondamental du projet quant à l'usage de ces technologies de fouille a été leur validation dans des cadres bien maîtrisés afin de mesurer les biais générés par les outils. En effet, les systèmes d'analyse peuvent générer des biais à tous les niveaux, de la description à l'étape finale de visualisation, et ainsi fausser l'interprétation des résultats. C'est pourquoi en sus de l'innovation technologique, c'est toute la méthodologie d'usage, en relation avec les pratiques, qui a et doit encore faire l'objet de recherches et d'expérimentations quant à ces approches quantitatives encore assez nouvelles. OTMedia, grâce à la collaboration transversale entre recherche technologique et usages experts, a été et continuera sans doutes à être une opportunité pour étudier comment minimiser/faciliter/accompagner le travail de validation de l'utilisateur par l'interaction.

Les perspectives autour d'un Observatoire des Media Français sont vastes, tant pour la recherche scientifique, la recherche SHS, que pour les media eux mêmes. Deux projets internationaux offrent des approches intéressantes pour l'analyse des média. Europe Media Monitoring (EMM 2001), à l'initiative de l'Europe, collecte et analyse en temps réel plus de 10000 sources textuelles multilingues en 60 langues. NewsRover (NewsRover 2011) de l'université de Columbia propose une analyse transmodale su un choix de media américains, à l'usage des journalistes. Les outils d'analyse et de fouille ou d'aide à la consultation des sources radio et audiovisuelles, tels que la segmentation et la reconnaissance de locuteurs, la reconnaissance de visages, l'extraction de textes dans l'image, l'agrégation avec feedback, les logiciels d'analyse factorielle et statistiques, la mise en œuvre de mesures existantes ou à découvrir sur les graphes de proximité ou de cooccurrence..., sont autant d'exemples qui trouveraient leur usage dans ce cadre. Enfin, si l'on considère les media comme "la vitrine que veut bien se donner la société", l'étude des thèmes ou des faits politiques, sociaux, scientifiques ou culturels, en relation avec leur visibilité médiatique, ouvre des perspectives d'analyse sociétale importantes. Les politiques *open data* développées ces dernières années et le développement d'outils d'analyse de gros volumes de données rendent ces développements d'envergure possibles, si l'on s'en donne les moyens.

Summary

The world of media is strongly impacted by the digital revolution: means of production, publishing, and broadcasting broaden and users' practices evolve, changing on the way established economic rules. The TransMedia Observatory is a framework of analysis. Its aim is to make the relationship between the Internet, radio and television broadcasts and press agency more intelligible. The conception, evolution and validation of the framework benefits of a strong collaboration between scientific research and social science. The first human studies with the tool give partial answers to important questions for citizen : Who produced the information? does broadcast multiplication ensure nevertheless information plurality?

Références

Allan, J., (2002), *Introduction to topic detection and tracking*. Topic detection and tracking, Springer US.

Bourdieu P., (1996), *Sur la télévision*, Liber-Raisons d'agir, Paris,

Cataldi, M., Di Caro, L., & Schifanella, C., (2010). *Emerging topic detection on Twitter based on temporal and social terms evaluation*. MDMKDD'10 - 10th International Workshop on Multimedia Data Mining (pp. 1-10). New York, New York, USA: ACM Press.

Joly, A.,Buisson, O., (2008), *A Posteriori Multi-Probe Locality Sensitive Hashing*. In ACM Multimedia , pages 209–218, Vancouver, Canada.

Letessier, P., Buisson, O., Joly, A., (2012), *Scalable Mining of Small Visual Objects*, In Proceedings of the 20th ACM international conference on Multimedia, Nara, Japa:

Lowe. D., (1999), *Object recognition from local scale-invariant features*. In ICCV, page 1150, Kerkyra, IEEE.

Renoust B., Melançon G., and Viaud ML., (2013), *Measuring group cohesion in document collections*, The 2013 IEEE/WIC /ACM International Conference on Web Intelligence (WI), Atlanta, USA.

White P., (1997), *Le village CNN, la crise des agences de presse*, Presses de l'Université de Montréal.

Pour en savoir plus:

OTMedia (2012) **http://www.otmedia.fr/,** + demo sur youtube

EMM (2001): **http://emm.newsbrief.eu/** Observatoire Européen de medias textuels multilingues

NewsRover (2012): http://www.ee.columbia.edu/ln/dvmm/newsrover/ projet américain d'analyse de masse des media à l'usage des journalistes