Indexer des parcours thématiques pour valoriser les collections de presse numérisée

Viviane Clavier
Viviane.Clavier@u-grenoble3.fr

Université Stendhal, Laboratoire Gresec, Grenoble3

Résumé : Notre étude se situe dans le cadre d'une réflexion sur la valorisation des collections de presse numérisée du XIXème siècle et sur les modes d'accès à ce patrimoine, témoignage inestimable de notre passé. Nous nous intéressons aux parcours thématiques, dispositifs qui peuvent permettre au grand public de découvrir les collections. L'élaboration de parcours nécessite une indexation thématique du texte intégral. Notre propos est de définir ce qu'est un thème dans la presse, notion qui fait intervenir le typage des unités rédactionnelles ainsi que des marqueurs linguistiques qui rendent compte du positionnement éditorial du journal sur les événements.

Mots-clés : collections de presse numérisée, valorisation du patrimoine, indexation thématique

1 Introduction

Fin 2006, le journal hebdomadaire lyonnais le *Progrès Illustré (1890-1905)* est numérisé et mis en ligne par la bibliothèque municipale de Lyon sur un site expérimental, considéré comme pilote en France[1]. Une partie de la collection, notamment les *Causeries*, a fait l'objet d'une reconnaissance optique de caractères (ou océrisation). Début 2010, la collection numérisée a migré sur le site officiel de la bibliothèque municipale de Lyon : le *Progrès Illustré* figure à présent parmi de nombreux titres de presse régionale lyonnaise du XIXème siècle[2]. Représentatif d'une période qualifiée de « l'âge d'or de la presse » (Bellanger, 1972 : 22), le *Progrès Illustré* était le supplément littéraire du *Progrès* et paraissait le dimanche. La singularité et la richesse de cette collection régionale a été décrite par Jean-Pierre Bacot (2005). Vaste fourre-tout qui s'inscrit dans la lignée des *magazines*, la presse illustrée mêle littérature populaire, gravures, faits-divers, chroniques de jardinage, actualités régionales.
Actuellement, la plupart des institutions en charge de mettre en ligne leurs collections patrimoniales s'orientent vers une numérisation de masse qui conduit à « digitaliser » des millions de pages. A l'inverse de ces palmarès volumétriques, la bibliothèque municipale de Lyon reste plus modeste dans ses objectifs et souhaite mettre l'accent sur la constitution et la valorisation d'un patrimoine numérique de presse à destination du grand public. Isabelle Westeel rappelle à ce titre que le *« patrimoine » au sens d'héritage commun et de propriété collective est le bien de tous les publics,*

[1] (Landron, 2010)
[2] http://collections.bm-lyon.fr/presseXIX/showObject?id=PER003&date=00000522

tous légitimes pour se l'approprier, il faut donc que les bibliothécaires pensent « usages et publics » dès la conception des projets de mise en ligne (Westeel, 2004). C'est à l'occasion de la mise en ligne du *Progrès Illustré* que des contours plus précis ont été donnés à la notion de valorisation dans le cadre du programme de recherche CaNu XIX[3]. La valorisation est à entendre à plusieurs niveaux : a) permettre au lecteur de construire ses propres documents à partir des sources ; b) permettre aux professionnels des bibliothèques de construire des parcours thématiques ; c) offrir au lecteur une reconstruction du contexte spatial et temporel dans lesquels ces textes et gravures ont été produits. C'est le deuxième objectif qui nous intéresse, *i.e.* la construction de parcours thématiques. Nous conviendrons d'appeler *parcours thématiques* un dispositif de mise en exposition d'une collection numérique suivant un ensemble de *sujets* prédéfinis. La construction de ces parcours repose sur le principe de l'indexation thématique. Les parcours sont susceptibles d'être utilisés par le grand public pour découvrir les collections, ils suivent le même objectif que les dossiers thématiques qui sont des documents de synthèse rédigés par les professionnels. Dossiers et parcours sont conçus pour éveiller la curiosité, donner accès aux contenus et favoriser le passage d'un mode découverte à un mode d'interrogation de la base. Les parcours ne sont donc pas une fin en soi, mais un moyen de s'approprier les connaissances nécessaires pour devenir autonome dans la consultation.

La conception de dossiers thématiques est une pratique largement répandue sur les sites web des bibliothèques numériques (Gallica) ou des agrégateurs de contenus numériques (Europeana) : ils permettent de faire vivre le site, de l'animer et également de mettre en valeur un patrimoine numérique suivant un choix de thèmes attractifs. Cependant, les dossiers supposent un travail humain important puisqu'il s'agit de dépouiller la base, d'indexer le contenu au fil de la page, de collecter, trier, organiser l'information et, *in fine,* de rédiger une synthèse. Les parcours sont conçus comme une alternative aux dossiers thématiques, ce qui permettrait de diversifier les modes d'accès aux collections et d'épargner les professionnels des étapes de mise en forme et de rédaction abouties des documents de synthèse. Il faut comprendre cette étude comme une réflexion plus générale sur l'indexation thématique des journaux du XIXème siècle qui s'inscrit dans une optique de mise en valeur de ce type de fonds.

Si la question de l'indexation thématique n'est pas nouvelle en soi, ce qui est plus original en revanche, c'est de l'appliquer à des collections de journaux. Dans la presse, la définition de la thématique ne peut, selon nous, faire l'impasse d'une analyse qui s'attache à décrire les discours « au prisme de la ligne éditoriale » pour reprendre les propositions de Roselyne Ringoot, qui stipule que « l'analyse éditoriale est en quelque sorte un concept textuel [et que] c'est l'analyse du journal qui permet de la dégager » (2004 : 88). Cet objectif conduit à redéfinir les contours de notions généralement convoquées dans le domaine documentaire tels que les *sujets* et les *thèmes* que nous avons croisés avec une notion mobilisée dans les médias, les *événements*. Le principal objet de cet article réside dans la définition de ces notions qui déterminent trois niveaux d'indexation. Au-delà de ces aspects, l'étude pose également la question des ressources à mobiliser aux différentes étapes du processus d'indexation : quelles sont les contributions respectives des corpus, des collections, des connaissances des spécialistes, des documents

[3] Projet de recherche sur les *corpus numériques,* CaNu XIX (Canards Numériques du 19ème siècle, resp. Geneviève Lallich-Boidin), financé par la région Rhône-Alpes http://cluster13.ens-lsh.fr/spip.php?article117

historiques et des attentes des usagers ? Que peuvent apporter les outils et méthodes d'indexation automatique, sachant que les textes océrisés de collections numériques présentent entre 15 et 20% d'erreurs ? Enfin, quelle est la place des langages contrôlés dans ce genre d'application ? Les langages contemporains sont-ils adaptés à la description de documents anciens, peuvent-ils répondre à des états de langue différents, celui des collections du XIXème siècle, celui des usagers du XXIème siècle ? Sont-ils tout simplement adaptés au grand public, plus familier de requêtes sur les moteurs de recherche que sur les bases de données ?
Après une présentation des objectifs liés à la valorisation du patrimoine numérique et une analyse du rôle que peuvent jouer les parcours thématiques, nous dresserons un bref état de l'art sur les thèmes pour en venir à notre propre définition de l'indexation. Une étude des sciences et des techniques dans les *Causeries* illustrera la proposition d'indexation à trois niveaux.

2 Le parcours thématique, un enjeu de valorisation du patrimoine numérique ?

2.1 Missions pour les bibliothèques numériques : conservation, diffusion et mise en visibilité des collections

Parmi les innombrables programmes de numérisation du patrimoine écrit, la presse ancienne connaît depuis quelques années un grand succès aussi bien en France qu'à l'étranger. Les programmes de numérisation sont pour l'essentiel dévolus aux bibliothèques, plus rarement aux organismes de presse. Pour ces derniers, la numérisation est liée à des objectifs de réédition éditoriale qui permettent de « donner une seconde vie » aux collections (par exemple, *Le Progrès, 150 ans d'actualités à la Une*, novembre 2009), alors que pour les bibliothèques, le processus de numérisation est lié à des objectifs de conservation et de diffusion (Mezzasalma, 2009).
Dans son rapport sur la « Numérisation du patrimoine écrit », Marc Teissier préconise toutefois une stratégie plus offensive destinée à rendre les bibliothèques numériques visibles sur le web. Il évoque trois actions susceptibles de favoriser l'accès à une bibliothèque numérique (ici Gallica) et d'en accroître la visibilité : a) la multiplication des accès, depuis la base, à des contenus variés (stratégie dite de « liens fins ») ; b) l'amélioration du signalement et du référencement ; c) et un meilleur accès pris en compte par les moteurs de recherche des métadonnées et de l'indexation de l'ensemble des contenus (indexation « plein texte ») (Teissier, 2010 : 28)
Ces incitations doivent conduire à indexer massivement le contenu des bibliothèques numériques, dans le but d'améliorer le référencement auprès des moteurs de recherche et de favoriser un accès en mode texte. Certains professionnels de l'information affichent cependant une certaine prudence à l'égard du mode de recherche plein texte et évoquent d'autres formes d'accès au contenu. Ainsi, Isabelle Westeel (2009) affirme que ce mode de recherche, tout en étant indispensable, ne peut se passer d'une découverte préalable des collections :

> *On touche là à la nécessaire réflexion sur l'accès aux documents, qui ne saurait calquer les catalogues des bibliothèques : il faut donner à voir une collection numérique avant d'offrir une recherche précise, les promenades libres ou guidées ont une signification.* (Westeel, 2009 : 30)

Cette réserve bien légitime concernant la mise en ligne d'un patrimoine numérique culturel sans réflexion préalable sur les modalités d'accès a été maintes fois soulignées. Récemment, Marie Després-Lonnet indique au sujet d'Europeana qu'il y a une confusion entre la capacité technique de « rendre accessible » des

données, par leur « mise en ligne » et l'accès aux savoirs, c'est-à-dire, la possibilité effective donnée à un individu de s'approprier de nouvelles connaissances (Miège, 1997, cité par Després-Lonnet, 2009 : 19). L'auteur précise en effet que la numérisation du patrimoine relève davantage d'une *textualisation* que d'une *digitalisation,* et que l'accès au patrimoine numérique suppose de franchir la barrière de la langue, « barrière autrement plus redoutable que celle qui consiste à pousser les portes d'un musée » (Després-Lonnet : 21).

Finalement, la valorisation du patrimoine numérique révèle dans sa mise en œuvre et ses missions des objectifs difficilement conciliables. Pour les bibliothèques, la conservation et la diffusion imposent une politique de numérisation systématique et massive afin de préserver les collections. Cependant, compte tenu des coûts liés à la numérisation[4], la reconnaissance optique de caractères ne pourra s'appliquer aux collections dans les mêmes proportions. Or, l'océrisation est l'étape indispensable pour procéder à l'indexation du texte intégral, traitement qui pourrait assurer la visibilité des collections auprès des moteurs de recherche et l'accès au grand public. La réflexion sur les parcours thématiques se situe donc dans un contexte technique, institutionnel et politique relativement complexe. Il semble qu'elle trouve davantage d'écho auprès d'institutions bénéficiant d'une marge de manœuvre plus importante au niveau de leur politique documentaire. Les bibliothèques municipales, en raison de leur relative autonomie vis-à-vis des grands programmes nationaux et européens, de leur proximité avec leurs publics et de leur attachement aux collections ont sans doute plus de latitude pour valoriser leur patrimoine numérique.

2.2 Attentes des usagers : l'accès au document en mode texte, l'importance relative des thèmes

Les études consacrées aux usages des bibliothèques numériques sont encore relativement rares (Gallica[5], Europeana[6]). L'enquête réalisée sur Gallica 2 montre que les usagers plébiscitent les bibliothèques numériques dès lors que l'on peut accéder en mode texte aux collections, le feuilletage d'un document en mode image étant perçu comme fastidieux. Certains usagers reconnaissent toutefois que l'accès en mode texte présente des difficultés.

> *« Le perfectionnement sur Gallica 2, c'est la mise en mode texte. On sait ce qu'il y a dans l'ouvrage en quelques minutes. Dans Gallica 1, ça prend du temps de feuilleter, de trouver la table des matières. Mais il faut savoir rechercher. Il y a de l'imprécision dans le texte qu'on recherche.* » (Matharan et al., 2008 : 7)

Par ailleurs, cette enquête, qui s'attache aux pratiques de recherche d'information sur l'internet, indique que l'accès est facilité lorsque l'organisation de l'information est présentée de manière thématique. Si l'on peut supposer que cette assertion s'applique également à des collections fermées, aucune étude, à notre connaissance, ne peut encore le confirmer.

> *Une question ouverte permettait aux répondants d'indiquer quels sites ils fréquentaient (sur Internet) et auxquels ils participaient le plus. 240 réponses ont en tout été récoltées. Une logique thématique nette en émerge : on voit que les répondants mentionnent avant tout des sites traitant d'un thème précis, spécialistes d'un sujet (ibid : 7)*

---

[4] Une page numérisée coûterait environ 1€ auquel il faudrait rajouter 1€ supplémentaire pour la conservation. (intervention orale de Pascal Sanz, Directeur de Département Droit, économie et politique de la BnF, aux journées d'études « Regards croisés sur la mise en ligne et la valorisation de la presse XIX-XXI. » Lyon, les 6 et 7 mai 2010).

[5] (Matharan et al., 2008)

[6] (Lesquins, 2007) ; (Bouvier-Ajam, 2007)

Plus récemment, une enquête sur les usages du *Progrès Illustré* conduite par les participantes au Projet CaNu XIX (Paganelli et Mounier, 2010) montre que les lecteurs se répartissent globalement en deux catégories : les spécialistes consultent le support papier, le grand public, le support numérique. Le profil type du lecteur de la version en ligne du *Progrès Illustré* est lettré, souvent à la retraite, peu préoccupé par la pertinence des résultats fournis, parce qu'il ne cherche rien de précis. Ce qui le gêne en revanche, ce sont les problèmes d'affichage, les caractères étant souvent trop petits. Il s'intéresse essentiellement aux dates, aux lieux, aux événements, aux personnes, en dernier lieu aux sujets. Cette enquête ne révèle pas d'engouement particulier pour les parcours ou les dossiers thématiques.
Si l'on résume ces différents points de vue, il semble y avoir un consensus fort autour de l'accès en mode texte. En revanche, si la dimension thématique semble appréciée des internautes, elle ne saurait suffire pour le *grand public lettré*[7], lequel attend des collections de presse, des informations en lien avec l'actualité régionale de l'époque, des lieux et des dates. Concernant les parcours thématiques, on peut pour l'instant supposer qu'ils s'adresseront à un public de non-spécialistes, i.e qui ne sont ni amateurs éclairés ni professionnels, ce que l'on convient d'appeler par méconnaissance, « le grand public ».

2.3 Modes d'accès aux collections de presse : l'accès thématique encore marginal

Récemment, Agnieszka Smolczewska-Tona et Geneviève Lallich-Boidin (2008) ont présenté un état de l'art des différents modes de recherche dans les collections numériques de presse : des dispositifs les plus élémentaires, (accès par titre et par date de publication), aux plus élaborées (recherche par mots-clés), certaines interfaces offrent parfois la possibilité d'affiner les critères de recherche. Les auteurs mentionnent que les accès thématiques (*Colorado Historic Newspaper)* ou par sujets (*Brooklyn Daily Eagle Online)* commencent à se développer.
Ce mode d'accès, qui permet de naviguer dans une collection suivant une logique thématique ou par sujet, est considéré comme beaucoup plus performant que le mode de recherche par mot-clé (Abdullah et Gibb, 2009 cité par Da Sylva, 2009). Pierre Zweigenbaum et Benoit Habert (2004) indiquent que le « foisonnement de données textuelles et d'outils » conduit la plupart du temps à une désorientation des usagers, et qu'il est nécessaire de fournir des « boussoles sémantiques » pour naviguer dans les documents. Plusieurs travaux montrent que la navigation dans une structure pré-établie serait une aide considérable pour les usagers. Il existe plusieurs dénominations pour qualifier ces outils : « outil de butinage » pour Da Sylva (2009), il permettrait de guider l'usager et favoriserait « une appropriation graduelle d'un contenu, même si l'utilisateur n'a aucune connaissance préalable de celui-ci. » ; il serait également une « aide à la lecture » qui permettrait l'évaluation de la pertinence de documents. « Système de visualisation de l'information » pour Davis (2006)[8], il permettrait de regrouper les documents semblables, de cerner la pertinence des documents retrouvés et de combiner la recherche par mots-clés.
Si l'on observe le site de la bibliothèque municipale de Lyon, on peut observer deux procédés pour faire découvrir les collections. La rubrique intitulée « Notre sélection d'articles et d'illustrations » renvoie au premier procédé. Il permet à l'usager de choisir des documents dans une liste : les critères de la sélection ne sont cependant pas explicites. Les documents sont issus de différents titres de presse offerts dans la base, il n'y a pas de critères thématiques qui président à la

[7] (Paganelli et Mounier, 2010)
[8] cité par Da Sylva (2009 : 264)

constitution de ces listes. Ce mode de découverte permet de présenter la collection la plus célèbre du fonds (Le *Journal de Guignol*), des anecdotes (*Gazette de la mode*), des événements historiques (*Le canal de Panama*). Le second procédé réside dans la constitution de « dossiers thématiques ». En page d'accueil du site, sept dossiers sont présentés dans un espace dédié, distinct de l'espace kiosque et de l'interface de recherche, comparable à un espace d'exposition, au sens muséal du terme. Les dossiers reposent sur le principe de la synthèse et donnent lieu à un nouveau document placé sous la responsabilité d'un auteur. Les titres des dossiers mentionnent différents sujets : la bicyclette, l'anarchisme, les grandes affaires criminelles, la mode etc. En revanche, la facture des documents varie d'un dossier à l'autre. Par exemple, le dossier *Notre fin de siècle appartient à la bicyclette* s'appuie exclusivement sur des sources extraites du *Progrès Illustré* et des données historiques (naissance du code de la route, nom de l'inventeur du vélocipède). Il s'agit de présenter les points de vue d'éditorialistes du *Progrès Illustré* sur la bicyclette, de rapporter des citations d'hommes célèbres, de retracer des événements sportifs relatés par le journal et d'illustrer le dossier par des gravures sur le sujet. Le dossier *Le Progrès Illustré, témoin de son époque* suit également une logique chronologique et retrace les événements sociaux (grèves), politiques (exécution de l'anarchiste Ravachol), les affaires de corruption qui ont traversé les années 1891-1895. Encore différents se révèlent les dossiers *Surveillance et répression de la presse anarchiste* et *Elégante, suggestive, excentrique, la mode dans tous ses états*. En effet, les sources citées ne se limitent pas au *Progrès Illustré*, mais à d'autres titres du fonds de presse numérisée de la Bibliothèque (*L'émeute, L'étendard révolutionnaire, La mode illustrée*), voire à des sources extérieures comme le roman de Zola *Au bonheur des dames*, une base de données image sur les textiles, des ouvrages sur l'histoire de la mode. S'agissant du dossier sur la répression de la presse, le document s'apparente davantage à un travail de recherche croisant des sources extérieures, que sur une compilation d'extraits tirés de journaux.

En conclusion, bien qu'attractifs, les dossiers sont peu nombreux et couvrent peu de thèmes. Leur réalisation est par ailleurs largement dépendante des connaissances des indexeurs et se révèle très lourde en termes de traitement. Par ailleurs, les dossiers posent d'une part, la question de la nature des connaissances, des types de sources qui président à leur construction et d'autre part, la question de ce qu'est un thème dans la presse.

3 Définir, extraire et représenter des thèmes dans la presse

3.1 Le thème : une notion polysémique

La notion de thème est abordée par les disciplines littéraires et linguistiques et dans le cadre des pratiques documentaires.

Du côté des sciences du texte, il existe une littérature volumineuse consacrée au *thème* et à la *thématique*. Nombreux sont les travaux qui soulignent l'emploi difficile de cette notion au point même que certains auteurs s'interrogent sur la pertinence du concept.[9] L'une des raisons qui fait obstacle à une définition synthétique du *thème* réside dans l'extrême hétérogénéité des perspectives d'analyse. Le *thème* a été défini dans le cadre de la phrase, du texte et du discours. Les unités thématiques peuvent être diversement approchées. Elles peuvent être identifiées syntaxiquement grâce à des marqueurs de thématisation (Porhiel, 2005). Elles peuvent être assimilées à des unités lexicales structurées en champs

[9] voir Porhiel (2005) pour un état de l'art sur le thème.

sémasiologiques ou onomasiologiques (Trudel, 2009). Elles peuvent consister en un « agrégat des thèmes des phrases qui composent un paragraphe ou un texte » (Goustos, 1997)[10] ou encore ne correspondre à aucun constituant dans un énoncé, mais à un « topic », c'est-à-dire une relation « d'à-propos »[11]. Enfin, François Rastier (1999) indique qu'un thème peut renvoyer « à une structure stable de traits sémantiques, récurrente dans un corpus, et susceptible de lexicalisations diverses. » Il précise en outre que « selon les discours et les genres, les normes de lexicalisation des thèmes varient ». Pour Evelyne Martin le thème connaît plusieurs dénominations : il peut être l'équivalent de *motif* (au sens littéraire du terme) de *leitmotiv* (dans une acception plus musicale) et manifeste le plus souvent un principe de récurrence. (Martin, 1995 : 15-16).

Dans le domaine documentaire, la notion de *thème* est mobilisée pour décrire des outils d'accès au contenu, les index thématiques, ou pour concevoir des documents de synthèses, les dossiers thématiques. La terminologie est flottante entre *sujet* et *thème*, les index thématiques permettant d'accéder aux documents *qui parlent de la même chose*, i.e. qui traitent du même *sujet*. Dans le contexte numérique, Muriel Amar (2004) souligne que de nouveaux enjeux se profilent pour les index thématiques. L'auteur indique que les index doivent à présent être intégrés aux documents primaires afin de servir d'outils de recherche et de lecture « *sous réserve qu'ils relèvent du statut linguistique adéquat (unités nominales référentielles) »*. Par ailleurs, les index doivent permettre *de manipuler non plus l'intégralité d'un document mais aussi des segments pouvant, le cas échéant, être combinés pour produire de nouveaux documents, sous réserve que soient introduites des connaissances contextuelles, externes au document.* (Amar, 2004 : 62-63). La satisfaction de ces objectifs suppose une refonte profonde des objectifs de l'indexation documentaire. L'auteur constate en effet que la construction d'index thématiques rencontre des obstacles que l'indexation contrôlée ne peut résoudre dans le contexte numérique. D'une part, la construction d'un thème est une opération fondamentalement discursive qui nécessite des connaissances extérieures au document, auxquels «l'utilisateur » n'a pas accès. Il en résulte une interprétation difficilement « reconstituable ». D'autre part, les formulations utilisées dans les langages contrôlés sont de nature lexicale et référentielle (seuls les groupes nominaux sont des descripteurs), alors que les thèmes ne sont pas des unités référentielles mais discursives. Ce constat *désespérant* conduit l'auteur à poser la question de savoir comment concilier la thématisation et la référenciation dans les objectifs de l'indexation professionnelle. Elle prône une indexation qui permettrait un accès direct au texte intégral, qui ne s'attacherait pas à une indexation lexicale mais discursive. Elle nomme « indexation discursive » le processus qui consiste à donner à l'utilisateur non pas des mots pour dire les thèmes, mais des documents, regroupés thématiquement, qui sont les contextes « qui rendent intelligibles et interprétables les thèmes des documents » (ibid. 65). Pour l'auteur, indexer consiste alors à permettre la construction des unités d'interprétation que propose le texte, et non les nommer.

Nous adhérons à ce dernier point de vue qui réaffirme l'inadéquation d'une approche lexicologique pour atteindre les thèmes et qui pointe sur l'impossibilité de dénommer des thèmes au sein de catégories référentielles. Bref, ces arguments

[10] Cité par (Porhiel, 2005)

[11] Marandin J.-M., « Thème, topic de discours » *Dictionnaire de sémantique* http://www.semantiquegdr.net/dico/index.php/Th%C3%A8me_%28topic%29_de_discours

montrent que les approches documentaires fondées sur le recours à des langages contrôlés pour décrire les thèmes sont inappropriées. En revanche, mettre à jour des contextes intra-discursifs ou des documents extérieurs pour favoriser l'interprétation des thèmes nous semble une notion prometteuse. Nous allons à présent nous intéresser à la notion de thème dans la presse.

3.2 Les thèmes dans la presse

Deux grandes familles de travaux abordent les thèmes dans la presse : la sociologie des médias et la linguistique textuelle. Dans le champ de la sociologie des médias, ce sont les travaux d'inspiration linguistique ou sémiotique sur les documents textuels ou audiovisuels qui mobilisent la notion. La linguistique textuelle, quant à elle, s'attache à théoriser le texte et à hiérarchiser ses composants dans le cadre de grammaires. Le thème participe de l'organisation textuelle, la progression thématique étant notamment liée à la cohésion du discours (isotopie) et, aux marqueurs de connexité (anaphores). Bien que poursuivant des objectifs radicalement différents, ces deux familles de travaux se « re-connaissent ».

Ainsi, Jean-Michel Adam et Gilles Lugrin cherchant à typer les unités rédactionnelles et catégorielles de la presse contemporaine évoquent les *thèmes,* pour désigner *« des objets de discours inséparables des familles d'événements»* (Adam et Lugrin, 2000 : 13). Ils s'appuient sur les travaux de Maurice Mouillaud et Jean-François Têtu (1989) pour qualifier les « familles événementielles » de catégories référentielles apparaissant au sein des rubriques. Les *nouvelles politiques, les catastrophes, les conflits sociaux* sont, pour ces spécialistes des médias, des familles d'événements, notion qui à son tour, fait l'objet d'une littérature titanesque. Ce qu'il faut retenir des événements, c'est que la qualification d'événement dans les médias n'est pas du ressort de la linguistique mais procède d'une reconfiguration de la réalité « *déformée* » par *« l'industrialisation des métiers de la presse, le développement des technologies modernes de communication et/ou les intérêts économiques et financiers des groupes qui les fabriquent* » (Arquembourg, 2006 : 14). Il existe des typologies d'événements dressées dans le cadre de la norme de métadonnées IPTC[12]. Michael Palmer présente et commente des exemples de ces catégories référentielles qui permettent de « ventiler l'actualité » (Palmer, 2006 : 53) Par exemple, la violence présente 21 catégories (*guerres et conflits, actes de terrorisme, rébellions*, etc.). D'un point de vue documentaire, ce genre de typologie peut présenter un intérêt pour enrichir les métadonnées dans les corpus de presse contemporaine. Pour la presse du XIXème en revanche, la difficulté essentielle consiste à typer des notions qui ont pu apparaître comme des événements en leur temps, mais qui, aux yeux de l'histoire n'en étaient pas… et inversement.

L'analyse d'un thème dans la presse nécessite la prise en compte de plusieurs niveaux qui résultent de « l'éditorialisation » des discours. Roselyne Ringoot rappelle en effet que « l'analyse d'un thème informatif nécessite un travail de diagnostic éditorial qui contextualise le traitement d'une information en fonction de la politique éditoriale d'un journal. » (Ringoot, 2004 : 88). Parmi les différentes dimensions qui interviennent pour le typage d'un thème, il y a notamment les éléments qui participent de la morphologie du journal, les rubriques qui permettent d'établir l'identité énonciative des journalistes, le péritexte (titres et

[12] L'IPTC (International Press and Telecommunications Council) est une organisation internationale créée en 1965 pour développer et promouvoir des standards d'échange de données à destination de la presse.
http://www.iptc.org/cms/site/index.html?channel=CH0086

intertitres), les genres, et, pour reprendre la proposition de Maurice Mouillaud et Jean-François Têtu, les « familles événementielles ». Afin de donner une consistance langagière à la notion d'événements, nous pouvons dans un premier temps ramener un événement à un énoncé décomposable en un ensemble de catégories sémantiques et aspectuelles[13], telles que des dates, des lieux, des personnes et des verbes. Le typage des événements se ramènerait à un exercice de catégorisation et de hiérarchisation de classes d'arguments et de prédicats. Pour Maurice Mouillaud et Jean-François Têtu, un thème dans la presse du XIXème serait par exemple, *le patriotisme, le courage, la barbarie, l'anarchisme, la colonisation* –[14] qui révèlent des récurrences sémantiques au sein du discours. On voit donc bien qu'un même événement peut se prêter à une multitude de discours, de cadrages et de thématisations.

Revenons à la presse illustrée. Cette dernière fait feu de tout bois comme l'indique le rédacteur en chef du *Journal Illustré « Le Journal illustré est le journal de tous, comme il est le journal de partout. […] Notre mosaïque illustrée n'a pas d'école. Nos dessins s'inspirent de toutes choses interprétées par tous les crayons. Notre texte est rédigé par des plumes de différentes couleurs. Qui nous en voudrait ? » (Bacot, 2005 : 117).* Les éléments qui participent de la morphologie du *Progrès illustré* sont différents de ceux de la presse actuelle. Les rubriques ne sont pas encore celles que l'on connaît. Ainsi le *Progrès Illustré,* qui se donne pour mission de mettre l'art et la littérature à la portée de tous, présente des titres de rubriques qui ressemblent plutôt à des intitulés de genres (*Feuilleton, Poésie, Roman d'aventure),* à des activités ludiques (*Récréation*), à des moments de divertissement (*Jeux d'esprit, Mots pour rire*), à des rendez-vous familiers (*Causerie*). Les unités rédactionnelles ne sont d'ailleurs pas systématiquement « rubriquées », certains écrits littéraires comportant uniquement un titre *(Le Cuirassier Blanc, Le papillon).* La *Causerie* ressemble pourtant bien à une rubrique au sens contemporain du terme[15, elle] donne au journal son identité populaire et badine. En revanche, le genre de la *Causerie* se rapproche de plusieurs genres actuels, parfois comparable à la tribune, au portrait, à la chronique ou au fait-divers. Les causeries sont rarement sous-titrées, parfois datées (75 causeries sur 389). Elles sont systématiquement signées : les auteurs des causeries sont au nombre de 10. Certains n'ont écrit qu'une seule causerie (Caribet, Arsène Alexandre). Jacques Mauprat est l'auteur de la plupart des causeries (353), Paul Clairfont vient en second. Certains auteurs ont pu être identifiés : ce sont des hommes de lettres, critiques d'art, biographes, romanciers, historiens, chroniqueurs. L'origine littéraire des auteurs donne à voir dans les causeries un curieux mélange d'une presse populaire, qui cultive « l'art de dire » comme dans le journalisme issu de l'Ancien Régime, c'est-à-dire une presse de lettrés qui maintient la primauté des « littérateurs » sur les « informateurs »[16]

En résumé, l'identification des thèmes doit être réalisée dans un cadre contraint qui prend en compte plusieurs niveaux de description qui contribuent à cerner la ligne éditoriale : a) le journal, identifiable par son titre, son numéro, sa date ; b) la rubrique associée au titre et sous-titre éventuels, la date, l'auteur ; c) le genre à défaut de rubrique et d) les familles d'événements, identifiées par des dates, des lieux, des personnes et des verbes. Ces niveaux de structuration peuvent être

[13] L'événement est en linguistique une catégorie aspectuelle des verbes.

[14] Exemples tirés de Têtu (1997)

[15] « La rubrique a plusieurs fonctions, parmi lesquelles celles de classification et de hiérarchisation des informations ; mais elle permet également de donner au journal une identité qui lui est propre. » (Herman et Lugrin, 1999 : 72)

[16] (Ferenczi, 1996 : 21)

décrits par des métadonnées XML, encodées avec un schéma de DTD en TEI[17], recommandations largement utilisées pour décrire les données textuelles en sciences humaines et sociales. Il reste à présent à s'interroger sur les méthodes à utiliser pour extraire les thèmes.

2.3 Les méthodes d'extraction de thèmes

Dans le contexte documentaire, les thèmes sont généralement issus d'un seul document, l'indexation thématique consistant à épuiser tout le contenu. Les thèmes peuvent être extraits manuellement ou automatiquement. Les méthodes automatiques représentent actuellement la seule solution pour traiter de grandes quantités de données.

Parmi les méthodes automatiques, Frédérik Bilhaut rappelle que deux familles de méthodes coexistent (2006 : 28). Les méthodes numériques se fondent sur la notion de cohésion lexicale, c'est-à-dire la répétition de mots comme indicateur d'homogénéité thématique. Dénommée *text-tiling* par Hearst, son inventeur, la méthode conduit à une segmentation linéaire du texte en unités continues qui ne se superposent pas, chaque segment étant décrit par un vocabulaire spécifique. Trouver des thèmes consiste alors à identifier les frontières qui révèlent des changements lexicaux dans les segments. La seconde famille de méthodes est linguistique et consiste à exploiter différents marqueurs linguistiques et positionnels porteurs d'indications de la structure thématique ; d'autres approches consistent encore à découper le texte en unité thématiques et rhématiques.

Le problème majeur qui se pose dans la méthode du *text-tiling*, réside dans le procédé de segmentation d'unités non superposables. En effet, dans le cas d'articles de presse, les thèmes peuvent se répéter dans les numéros, un thème pouvant faire l'objet de l'actualité pendant plusieurs semaines, voire plusieurs mois ; ils peuvent être abordés dans plusieurs titres de journaux, et, à l'intérieur d'un même numéro, dans plusieurs rubriques. En ce qui concerne les méthodes linguistiques en partie « automatisables », Frédérik Bilhaut fait remarquer que les marqueurs de thématisation semblent difficiles à caractériser a priori, et que la démarche relève davantage d'une démarche d'observation de corpus que de repérage automatique de thèmes (Bilhaut, 2006 : 100). Quant à la méthode de typage des unités thématiques et rhématiques, l'auteur reconnaît l'intérêt de la démarche, met cependant en doute le degré de généralité des ressources et évoque également la charge de travail nécessaire à leur constitution (*ibid.* : 101).

Récemment, trois auteurs du laboratoire LIRIS ont présenté une méthode de catégorisation de l'information d'actualité dans des dépêches de presse collectées par flux RSS (Laitang et al, 2009). La catégorisation s'appuie sur la sélection et la pondération de « termes » représentatifs d'une thématique (par ex. *le sport, la politique*) et d'un sujet, notion qui tient compte de l'apparition et de la disparition dans le temps de nouvelles (par ex. *élection américaine, tremblement de terre*). Ce qui est très intéressant dans cette démarche, c'est la double prise en compte des dimensions temporelle et thématique pour regrouper les informations d'actualité. La spécificité événementielle des corpus de presse est également mobilisée pour indexer les contenus, puisque les auteurs ont fait le choix de retenir les catégorisations de dépêches de presse (IPTC) évoquées *supra* comme base de référence thématique statique. Ce qui en revanche, semble difficilement transposable à notre travail c'est tout d'abord, le recours à des ressources sémantiques contemporaines pour enrichir les termes (ontologies, thesaurus, bases de données lexicales). En outre, les calculs de proximité entre les termes s'appuient

[17] http://www.tei-c.org/index.xml

sur les corpus lemmatisés, ce qui semble, là encore difficilement applicable aux articles de presse océrisés.
Cette dernière approche présente très clairement l'ensemble des étapes et méthodologies pour extraire, indexer et catégoriser des thèmes récurrents afin de déterminer les schémas de propagation des actualités sur internet. Notre objectif est différent, puisque nous voulons à partir d'un angle d'analyse précis, indexer des thèmes et les structurer en parcours. Dans ce qui suit, nous présentons notre définition du thème et nos choix d'indexation.

4 Les sciences et techniques dans le *Progrès Illustré* : étude de cas

Nous proposons de définir trois niveaux d'indexation qui interviennent pour l'identification d'une thématique. Le première niveau consiste à construire un index des *sujets* ; le deuxième, un index des *événements*, et le troisième, un contexte de localisation des *thème*s. Le niveau 2 est un sous-ensemble du niveau 1 : il révèle les sujets qui font l'objet d'un événement ; et le niveau 3 permet de contextualiser les niveaux 1 et 2 : il révèle le positionnement éditorial du journal dans le traitement médiatique d'un sujet, et favorise une interprétation de la thématique. Nous allons à présent détailler chacun de ces niveaux en évoquant le statut de l'unité textuelle, la nature des unités d'indexation et, le cas échéant, le langage d'indexation retenu.

4.1 Indexation par sujets

Le choix d'un *sujet* ne repose pas sur une analyse des fréquences « des mots » des textes. Les critères de choix sont extérieurs au corpus : ils peuvent résulter de la connaissance qu'ont les indexeurs des usagers du fonds, de l'analyse des traces de requêtes, de spécificités jugées amusantes ou curieuses de la collection (les anecdotes, les caricatures), de préconisations de spécialistes sur l'apport de ces documents pour l'histoire, la littérature, ou la presse, de choix muséographiques (faire une exposition), etc. La liste des *sujets* n'est donc pas définie *a priori*, elle se construit au fil du temps, en fonction de choix de valorisation. Cette démarche est également celle qui préside à la construction des dossiers thématiques.
Les dénominations des *sujets* peuvent être choisies dans un langage contrôlé, si possible interopérable, ce qui permettrait un moissonnage de métadonnées entre les bibliothèques numériques. Dans l'idéal, le langage de classification devrait être contemporain des collections afin d'être en meilleure adéquation avec les vocabulaires. Les premières classifications Dewey pourraient convenir, mais elles sont en anglais. La question de l'interopérabilité se situe donc à deux niveaux : celui de la langue, et celui des états de langue, puisqu'il faudrait envisager une correspondance entre une classification du XIXème siècle et sa version contemporaine. Il convient d'associer à chaque classe, les mots-clés issus du texte, *i.e* les vocabulaires qui décrivent les *sujets*. Localisés à différents niveaux de la structure rédactionnelle, les mots-clés constitueront l'index des *sujets*. Ce sont des groupes nominaux (entités nommées, noms communs et leurs diverses expansions) qui apparaissent dans des contextes relevant d'un même sujet. Cette perspective onomasiologique ne fournit aucun indice linguistique et quantitatif sur le repérage de ces entités : c'est pourquoi l'indexation est généralement réalisée à la main au fil de la page. Ce qui nous paraît intéressant, c'est de se servir du vocabulaire pour favoriser un accès aux collections. Fournir des « mots » aux usagers leur permettrait d'interroger les collections en mode texte. Par conséquent, un dictionnaire figurant les entrées lemmatisées, auxquelles seraient associées les

familles de mots issues des collections, les parasynonymes, les diverses expansions du lemme nous paraît satisfaire cet objectif. Suivre un parcours de découverte consisterait donc à choisir un sujet parmi une liste proposée, afficher le dictionnaire correspondant à un sujet, cliquer sur une entrée de dictionnaire et lire les textes qui comportent ces entrées. Dans l'idéal, il faudrait afficher des rubriques et des illustrations légendées. L'affichage peut suivre une logique chronologique, celle du numéro du journal.

Pour valider cette approche, nous avons conduit une étude sur les sciences et les techniques dans le *Progrès Illustré* en raison de l'importance historique que représente ce sujet « pour comprendre l'origine d'une mutation globale de la société » (Gille, 1978 : 773). L'extraction du vocabulaire qui dénomme des objets scientifiques, des découvertes, et plus largement, des acteurs ou des institutions en lien avec les sciences et techniques a été réalisée avec Nooj, un outil de description linguistique de corpus[18]. Cette étude a permis de dégager des ensembles lexicaux par année et de les annoter.

Il est bien connu que cette démarche comporte des difficultés importantes dues notamment aux phénomènes de polysémie et d'homographie (Ehrlich, 1995). Toute entrée lexicale doit par conséquent être vérifiée en contexte à l'aide d'un concordancier. Un autre problème de taille liée à la perspective onomasiologique que nous avons retenue est d'apprécier la pertinence des classes lexicales retenues pour décrire les sciences et techniques. En effet, le vocabulaire ne présente pas les propriétés généralement dévolues au lexique scientifique et technique : ce n'est pas un sous-langage. Dans les causeries, il n'y a pas de termes de spécialité puisque cette presse n'est pas de la vulgarisation scientifique. Par conséquent, les critères d'appréciation ne peuvent pas s'appuyer sur des critères morpho-syntaxiques. En outre, on souhaite identifier des unités lexicales qui ont *un lien* avec les sciences et techniques, et qui ne sont pas des termes au sens strict. Par exemple, les inventeurs d'une technique, les chirurgiens célèbres, les instituts de science, etc. Pour valider nos classes, nous avons utilisé des connaissances extérieures (essentiellement des chronologies et des ouvrages sur l'histoire des sciences et des techniques) ainsi qu'une analyse en contexte, puis nous avons utilisé les grammaires d'états finis pour annoter le corpus. C'est le contexte qui nous indique que *ficelle* est la dénomination lyonnaise de *funiculaire*, v*apeur* celle de *tramway à vapeur* et une chronologie qui nous indique à quelles dates ces moyens de transport ont été inventés. Ce travail s'est révélé relativement fastidieux, même s'il aboutit *in fine* à une couverture exhaustive des 389 causeries océrisées : l'annotation, en revanche est très rapide grâce aux grammaires de Nooj.

4.2 Indexation par événements

La caractérisation des *événements* pose plusieurs problèmes qui ne pourront être résolus dans le cadre de ce travail. Nous devons considérer qu'un *événement* comporte plusieurs dimensions : linguistique, médiatique et historique. Les dimensions médiatique et historique ne sont pas repérables en langue : ce sont des sources extérieures qui permettent de leur attribuer un statut événementiel. Bien que l'indexation des événements soulève de nombreuses questions, la notion nous semble fondamentale pour caractériser la thématique, puisque c'est à travers les commentaires des événements que transparaît le positionnement éditorial du journal.

Le lien entre les *sujets* et les *événements* peut se faire dans le cadre d'une approche actantielle. Tel *sujet* peut figurer en position d'*actant* dans le cadre d'un procès qui

[18] http://www.nooj4nlp.net

décrit un *événement*. On se retrouve donc au niveau de la phrase, à l'intérieur des rubriques. Par conséquent repérer un *candidat événement* revient à croiser des fonctions dans une phrase et des catégories sémantiques telles que des toponymes, des entités nommées, des dates et des prédicats. Ces *candidats événements* peuvent être indexés à l'aide de deux sources extérieures : d'une part, en associant (ou pas) un *candidat événement* à une « famille événementielle », on valide ainsi la dimension médiatique de l'événement. On peut s'inspirer du standard IPTC pour définir les familles événementielles à l'aide de catégories plus adaptées à la presse du XIXème siècle. D'autre part, il s'agit de valider le *candidat événement* sur un plan historique. Tel assassinat peut, par exemple, relever d'une suite d'événements analysés comme relevant d'une seule et même affaire. Seul le regard critique de l'historien est en mesure de relier des événements épars et de lui donner sens. Cette approche consiste à documenter le corpus par des sources externes, ce qui relève d'une sémantique du lien.

La validation de la proposition a donné des résultats intéressants, en tout cas, pour la connaissance du corpus. Nous avons choisi parmi les *Causeries* qui traitent des sciences et techniques, celles qui évoquent un événement. En nous appuyant sur les repères linguistiques mentionnés, on recueille toutes sortes de récits anecdotiques, qui n'ont sans doute pas la prétention médiatique d'un événement. Jugeons-en : a) *un cas de transfusion sanguine … au sang de chèvre* (1891/02/22, n°10) ; b) *la découverte scientifique d'un physiologiste lyonnais … sur des microbes lumineux* (1891/07/12, n°30), c) *un astronome téméraire … qui s'écrase en aérostat sur une cheminée* (1893/11/19, n°153) d) *un procédé de vieillissement du vin par l'électricité … raconté comme une recette de cuisine (1891/12/13 n° 52)* On observe que dans ces événements, les repères temporels sont souvent très flous : *il y a quelques jours, la semaine dernière, je vous disais l'autre fois…* La structure de l'événement se rapproche plutôt du récit, pour ne pas dire parfois, de la fable.

4.3 Indexation par thèmes

Viennent enfin les thèmes que nous proposons d'appréhender comme les traces d'un positionnement éditorial sur l'actualité. Ces traces sont une des manifestations de l'*angle* journalistique, au sens où l'angle peut être appréhendé en tant que « formant » (Ringoot, 2004 : 108). Dans une étude récente, nous avons travaillé la notion de positionnement dans le cadre de la lecture et de l'annotation de documents scientifiques (Clavier et Paganelli, 2009). Nous avons observé que lorsque des lecteurs consultent une thèse, ils sont beaucoup plus attentifs au positionnement scientifique de l'auteur de la thèse qu'à la terminologie. Le positionnement scientifique – qui n'est pas une notion linguistique – a été défini par un ensemble de traces linguistiques qui révèlent une posture surplombante de l'énonciateur et qui se situent dans le métadiscours. Indexer ces marqueurs, révéler le contexte métadiscursif pour favoriser la compréhension et l'appropriation des connaissances, est la conclusion à laquelle nous sommes parvenues.

Nous formulons l'hypothèse qu'il existe également des marqueurs de positionnement dans la presse, et que la mise en évidence de ces traces permettrait une lecture-découverte située de l'actualité. Si l'on considère avec d'autres auteurs[19] que la ligne éditoriale peut être analysée sous l'angle de l'énonciation, cette approche permet d'exploiter des « traces repérables » qui sont à l'articulation de « l'énonciation textuelle » et de « l'énonciation éditoriale » (Ringoot, 2004 : 92). Si le positionnement permet de situer le point de vue éditorial, comment organiser les marqueurs en thématiques ? Dans les *Causeries*, les commentaires des

[19] Voir (Ringoot, 2004 : 92) qui cite divers auteurs dont Annelise Touboul.

chroniqueurs sur les événements sont des lieux privilégiés de l'expression individuelle et éditoriale. Les marqueurs de positionnement peuvent apparaître à plusieurs niveaux : lexical, phraséologique, dans les proverbes et les dictons. Par exemple, le caractère régional de la presse peut transparaître à travers une énonciation fortement ancrée dans des repères territoriaux : personnalités lyonnaises, lieux et institutions de Lyon et de ses environs, événements qui ponctuent la vie locale. Ici, c'est le lexique qui est mobilisé, ce qui transparaît dans les listes de fréquence. La thématique n'en est pas pour autant « régionaliste », sinon, on confond les *sujets* avec les *thèmes*. Ce sont les réseaux de cooccurrences (ou d'association) qui vont permettre de faire émerger une interprétation. Par exemple, Lyon rivalise toujours avec Paris, la thématique pourra être le chauvinisme. Les dénominations des thématiques nous importent moins que la mise en évidence du contexte d'interprétation.
Les marqueurs sont par ailleurs révélateurs de postures énonciatives. Par exemple, l'une des façons de révéler l'approche populaire du journal consiste à montrer que la voix du chroniqueur, dans l'analyse qu'il fait de l'actualité, se fond systématiquement dans celle du plus grand nombre, la *vox populi*. Elle se manifeste le plus souvent par l'ironie : les scientifiques sont de grands hommes, mais leurs découvertes sont bien piètres en regard des grands fléaux de l'humanité. Ainsi la thématique de la dérision peut-elle s'appliquer aux découvertes scientifiques et techniques, et par extension à leurs auteurs. Contrairement au positionnement scientifique, qui s'inscrit dans une dynamique de sur-énonciation, le positionnement éditorial de cette presse illustrée relèverait de la sous-énonciation.
Si la thématique ne peut être installée dans le code, il faut pourvoir circonscrire le contexte, i.e. délimiter les commentaires Dans les *Causeries*, les commentaires sont imbriqués dans le récit des événements, à la manière d'une conversation, mais sont repérables par le jeu des repères énonciatifs. Formellement, l'on passe du *il* (dans l'événement) au *on* doxique, au *je* (dans les commentaires), de l'assertion déclarative aux interrogatives et aux exclamatives, du mode indicatif au mode impératif. L'introduction de guillemets ne renvoie pas à des citations comme dans la presse contemporaine, mais à des commentaires qui sont des aphorismes, des proverbes, des dictons. La présence de traces repérables permet l'annotation des commentaires.

5 Conclusion

La création de parcours thématiques constitue l'un des moyens pour diversifier les modes d'accès aux collections, tout en répondant à des objectifs de valorisation et de mise en exposition du patrimoine de presse numérisée. De ce point de vue, les parcours et les dossiers thématiques partagent les mêmes objectifs. A la manière d'un musée qui expose ses collections, les parcours et les dossiers présentent une sélection d'*objets* textuels et iconiques, agencés suivant une certaine logique qui leur donne sens. Cette entrée dans les collections par la voie muséographique est censée favoriser l'accès du grand public à ces ressources, plus que ne le ferait un mode d'accès par affichage des numéros de journaux sous forme de calendrier par exemple. La mise en œuvre des dossiers et des parcours est cependant différente. Les dossiers sont conçus manuellement par des professionnels, ce qui offre une qualité incontestable mais se révèle long et fortement dépendant des connaissances des indexeurs. Inversement, l'idée qui préside à la construction de parcours thématiques est de proposer une méthodologie reproductible et systématique, destinée à sélectionner des *objets*, à les délimiter et à les représenter. Ces différentes

étapes sont envisagées dans le cadre d'une indexation thématique des collections de journaux.

L'indexation thématique a été abordée dans une perspective textuelle et cherche à résoudre la question de l'impossible adéquation entre une chaîne de caractères et un thème, situation qui condamne toute tentative de recours à des langages documentaires contrôlés pour décrire la thématique. Nous avons travaillé la notion de thème dans la presse en mobilisant trois notions : les *sujets*, les *événements* et les *thèmes*. Les *sujets* sont choisis dans le cadre de classifications qui dénomment des classes de connaissances. Les vocabulaires sont issus des collections et alimentent les classes ; leur localisation tient compte de la morphologie du journal. Les vocabulaires sont lemmatisés et organisés à la manière d'un dictionnaire regroupant les parasynonymes et les diverses expansions. Les *événements* sont des entités complexes modelées par un double processus synchronique (la médiatisation d'un événement) et diachronique (sa trace dans l'histoire). Les événements peuvent être regroupés en familles (les catastrophes, les assassinats, etc.). Nous faisons l'hypothèse, qui reste à valider, qu'un événement connaît un ancrage langagier et qu'on peut le ramener à certaines catégories sémantiques figurant dans un cadre actantiel. Les *sujets* instancient des places actantielles. Les *thèmes* sont construits à partir des traces linguistiques qui révèlent le positionnement éditorial sur un événement. Les traces de ce positionnement sont repérables dans les commentaires que font les chroniqueurs de l'actualité. Les commentaires sont donc vus comme le lieu d'expression de la subjectivité du chroniqueur et le lieu de manifestation de l'*angle* du journal. Ces deux énonciations, individuelle et éditoriale donnent à voir la thématique. Indexer signifie alors associer des types de commentaires à des familles d'événements.

La mise en œuvre technique des parcours a été abordée à la marge. Certes, nous avons utilisé un outil linguistique qui offre diverses fonctionnalités utiles à l'indexation : étiquetage, annotation, analyse en contexte. Mais nous n'avons pas exploité le dictionnaire de mots inconnus… bien fournis en noms propres, et regorgeant de mots erronés qui entraveraient toute analyse syntaxique. Du côté des méthodes de classification supervisée qui pourraient permettre de classer les *Causeries* par sujets, il nous semble que le principal problème réside dans la prise en compte des fréquences d'occurrences. En effet, les vocabulaires qui alimentent les classes, se situent dans les basses fréquences et ne peuvent donc constituer un critère de classement. Enfin, reste la question de l'identification des commentaires, qui, bien que présentant formellement des marqueurs, nécessite également une réflexion sur le typage de ces unités textuelles.

Si l'indexation de parcours thématiques dans les collections de presse soulève des problèmes de faisabilité et d'automatisation, elle replace néanmoins la question de l'accès au patrimoine numérique dans une perspective d'interprétation et de contextualisation de l'information.

6 Bibliographie

7 Adam J.-M. et Lugrin G., « L'hyperstructure : un mode priviliégié de présentation des événements scientifiques » in Cusin-Berche F. *Rencontres discursives entre science et politique. Spécificités linguistiques et constructions sémiotiques*, *Carnets du CEDISCOR, 6*, Presse de la Sorbonne Nouvelle, 2000, p. 133-149.

8 Adam J.-M., « Unités rédactionnelles et genres discursifs : cadre général pour une approche de la presse écrite », *Pratiques*, *94*, 1997, p. 3-18.

9 Amar M., « L'indexation aujourd'hui », *Les dossiers de l'ingénierie éducative*, *vol. 49*, 2004, p. 61-65.

10 Arquembourg J., « De l'événement international à l'événement global : émergence et manifestations d'une sensibilité mondiale », *Evénements mondiaux regards nationaux, Hermès 46*, CNRS Editions, 2006, p 13-21.

11 Bacot J.-P., La presse illustrée au XIXe siècle : une histoire oubliée, Médiatextes, Limoges : Pulim, 2005. 235 p.

12 Bellanger C., *Histoire générale de la presse française*. T.3. De 1871 à 1940. Paris, Presses universitaires de France, 1972, 687 p.

13 Bilhaut F.**,** *Analyse automatique de structures thématiques discursives - Application à la recherche d'information*, mémoire de thèse de doctorat en informatique, Université de Caen, 2006.

14 Bouvier-Ajam L.*, Europeana. Etude sur les usages et les attentes relatifs à l'interface de consultation de la future Bibliothèque numérique Européenne*, Rapport final, 2007, 53 p.

15 http://www.bnf.fr/documents/ourouk.pdf

16 Clavier V. Paganelli C., « Marqueurs de positionnement et parcours de lecture : un enjeu pour la consultation des thèses en ligne ? » Actes du Colloque *Changements technologiques, mutations organisationnelles et information professionnelle : pratiques, acteurs et documents,* organisé par le GRESEC, les 10-11 décembre 2009 à Echirolles.

17 Da Sylva, L., « Outil de butinage du contenu des documents de collections numériques », *Patrimoine 3.0 : actes du douzième Colloque international sur le document électronique*, 21-23 octobre 2009, Université de Montréal, Canada, sous la direction de Khaldoun Zreik, Paris : Europia productions, p. 263-273.

18 Desprès-Lonnet M., « L'écriture numérique du patrimoine, de l'inventaire à l'exposition : les parcours de la base Joconde », *Culture & musées, 14,* 2010, p. 19-38.

19 Ehrlich D., « Une méthode d'analyse thématique. L'exemple de l'ennui et de l'ambition », *in* Rastier F.(sld), 1995, p. 85-103.

20 Ferenczi T., *L'invention du journalisme en France : naissance de la presse moderne à la fin du XIX*e *siècle*. Paris, Payot, 1996. 275 p.

21 Gille B., « Les techniques de l'époque moderne », in *Histoire des techniques : technique et civilisations, technique et sciences,* Paris : Gallimard, 1978, pp 773-858. (Encyclopédie de la Pléiade, 41).

22 Herman T. et Lugrin G., « La hiérarchie des rubriques : un outil de description de la presse », *Communication et Langages, 122,* 1999, p. 72-85.

23 Laitang C., Egyed-Zsigmond E., Calabretto S., « Diversité de l'Information dans les Sites de Presse », *Patrimoine 3.0 : actes du douzième Colloque international sur le document électronique*, 21-23 octobre 2009, Université de Montréal, Canada, sous la direction de Khaldoun Zreik, Paris : Europia productions, 2009, p. 111-128.

24 Landron, P.Y., « Valoriser la presse illustrée du XIXème : l'exemple de la BM Lyon», Communication aux Journées d'études *Regards croisés sur la mise en ligne et la valorisation de la presse XIX-XXI*, co-organisées par les laboratoires ELICO (Lyon), GRESEC (Grenoble) et la Bibliothèque Municipale de Lyon dans le cadre du Cluster 13 « Culture, patrimoine et création » les 6 & 7 mai 2010 à la Bibliothèque Municipale de Lyon.

25 Lesquins N., *Europeana : rapport de bilan sur les usages et les attentes des utilisateurs*, Bibliothèque Nationale de France, 2007, 60 p.

26 http://www.bnf.fr/documents/europeana_2007.pdf

27 Martin E., « Thème d'étude, étude de thème » in RASTIER F. (sld.), 1995, p. 13-24.

28 http://www.revue-texto.net/Parutions/Analyse-thematique/Martin.pdf

29 Matharan J., Chaguiboff J., Alliot F., *Rapport d'étude sur les usages communautaires et collaboratifs, sur place et à distance, des ressources numérisées de la BnF,* Bibliothèque Nationale de France, 2008.

30 http://www.bnf.fr/documents/rapport_web_communaute.pdf

31 Mezzasalma P., « Conserver la presse », Dossier *La conservation et la numérisation de la presse*, in *Chroniques de la BNF*, n° 47, 2009, p. 5-7.

32 Paganelli C. et Mounier E., « La presse ancienne numérisée : modes d'accès et pratiques de recherche », Communication aux Journées d'Etudes *Regards croisés sur la mise en ligne et la valorisation de la presse XIX-XXI,* co-organisées par les laboratoires ELICO (Lyon), GRESEC (Grenoble) et la Bibliothèque Municipale de Lyon dans le cadre du Cluster 13 « Culture, patrimoine et création » les 6 & 7 mai 2010 à la Bibliothèque Municipale de Lyon.

33 Palmer M., « Nommer les nouvelles du monde », *Evénements mondiaux regards nationaux, Hermès 46,* 2006, p. 47-56.

34 Porhiel S., « Les marqueurs de thématisation : des thèmes phrastiques et textuels », *Travaux de linguistique, 2/51*, 2005, p. 59-88.

35 Rastier F. «La sémantique des thèmes - ou le voyage sentimental», *Texto !* 1999.

36 http://www.revue-texto.net/index.php?id=570.

37 Rastier F., « La sémantique des textes : concepts et applications », *Hermès*, *16,* 1996, p. 15-37.

38 http://www.revue-texto.net/Inedits/Rastier/Rastier_Concepts.html#A.

39 Rastier F. (sld.), *L'analyse thématique des données textuelles, l'exemple des sentiments*, Paris : Didier Erudition, 1995.

40 Ringoot R. et Robert-Demontrond P., *L'analyse de discours*, Editions Apogée, 2004, 222 p.

41 Smolczewska-Tona, A. et Lallich-Boidin, G., « De l'édition traditionnelle à l'édition numérique : le cas de la presse du XIXe siècle. » In Broudoux E., Chartron G. (dir.). *Traitements et pratiques documentaires : vers un changement de*

paradigme ? Actes de la deuxième conférence Document numérique et société, Paris : ADBS éditions, 2008, p. 299-316.

42 Teissier M. *La numérisation du patrimoine écrit*, Janvier 2010, *La documentation française*, http://www.ladocumentationfrancaise.fr/rapports-publics/104000016/index.shtml

43 Têtu J.-F., *Le journalisme mis en scène. In*: La presse selon le XIXe siècle. Université Paris III, Paris, France, 1997, pp. 137-154.

44 http://halshs.archives-ouvertes.fr/docs/00/39/73/90/HTML/

45 Trudel E., «Champ sémantique, champ sémantique lexical ou classe sémantique ?», *Texto!* 2009

46 http://www.revue-texto.net/index.php?id=2277.

47 Westeel I., « Le patrimoine passe au numérique », *Bulletin des Bibliothèques de France*, *1,* 2009, p. 28-35.

48 Westeel Isabelle, « Patrimoine et numérisation : la mise en contexte du document » [en ligne], in *Colloque EBSI/ENSSIB. Montréal. 13-15 octobre 2004.*

49 http://www.ebsi.umontreal.ca/rech/ebsi-enssib/pdf/westeel.pdf

50 Zweigenbaum P. et Habert B. « Accès mesurés aux sens », *Mots. Les langages du politique*, *74,* 2004*,* p. 93-106.

51 http://mots.revues.org/index4673.html