MAUD EHRMANN [1]

EXPLORER LA PRESSE NUMÉRISÉE : LE PROJET IMPRESSO

« Impresso – Media Monitoring of the Past » est un projet de recherche interdisciplinaire dans lequel une équipe d'historiens, de linguistes informaticiens et de designers collabore à la mise en données d'un corpus d'archives de presse numérisées. Les principaux objectifs du projet sont d'améliorer les outils d'extraction d'information pour les textes historiques, d'indexer sémantiquement des journaux historiques, et d'intégrer les enrichissements obtenus dans les pratiques de recherche des historiens au moyen d'une interface nouvellement développée.

Les journaux historiques sont un reflet des sociétés passées. Publiés régulièrement depuis des siècles, ils ont enregistré les guerres comme les événements mineurs, ont rendu compte des questions internationales, nationales et locales, et documenté la vie au jour le jour [2]. Ils ont, en un mot, suivi la grande et la petite histoire, constituant au fil des siècles et selon l'expression de l'éditeur du *Washington Post* en 1963, *« the first rough draft of history »* (la première ébauche de l'histoire). En tant que sources historiques, les journaux sont également le reflet des environnements politiques, sociaux et économiques dans lesquels ils furent produits, et

1 Le projet Impresso est le fruit du travail d'une équipe composée d'Estelle Bunout, Simon Clematide, Marten Düring, Andreas Fickers, Daniele Guido, Roman Kalyakin, Frédéric Kaplan, Matteo Romanello, Paul Schroeder, Philipp Ströbel, Thijs van Beek, Martin Volk et Lars Wieneke. Maud Ehrmann se fait ici la porte-parole de cette équipe et de ce travail collaboratif.

2 Nous remercions chaleureusement nos partenaires des archives, bibliothèques et journaux pour avoir partagé notre enthousiasme et mis à disposition leurs collections de journaux numérisés ainsi que les historiens (UNIL et Infoclio) pour leurs conseils et leur soutien pendant les ateliers et tout au long du projet. Des remerciements s'adressent aussi aux nombreux chercheurs associés qui ont accepté d'évaluer les prototypes d'interface d'Impresso à plusieurs reprises ; David Smith pour ses conseils concernant la détection de réutilisation des textes ; Benoit Seguin pour ses contributions sur la partie visuelle du moteur de recherche et le système de recommandation. Enfin, le Fonds national suisse (FNS) pour son soutien (projet CR-SII5_173719).

sont aujourd'hui porteurs d'une information dense et multiple qui les rendent inestimables pour les chercheurs [3].

Des sources riches donc, voire prolifiques, dont l'exploitation se heurte à une difficulté récurrente, à savoir l'exploration fastidieuse de grandes quantités de documents. Le dépouillement est bien sûr fonction de la question de recherche, mais il faut bien souvent déplier de nombreux numéros, tourner des dizaines de pages et parcourir des centaines d'articles afin de trouver, sans garantie de succès, les indices ou l'information espérés. Depuis quelques années, cette difficulté d'exploitation a cependant commencé à s'atténuer. Longtemps conservés sur des rayonnages d'étagères ou sur des microfilms, les archives de presse font en effet l'objet d'une numérisation massive et des millions de fac-similés, ainsi que leur contenu lisible par machine grâce à la reconnaissance optique de caractères (OCR, *Optical Character Recognition*), deviennent accessibles via divers portails en ligne. Cette évolution est le fruit d'efforts soutenus de la part de bibliothèques et d'archives pour mettre en œuvre la numérisation, améliorer les technologies d'OCR, et généraliser l'accès au texte intégral des archives de presse. Pour la Suisse, il s'est notamment agi du projet Presse Suisse en Ligne lancé en 2011. Orchestrée par la Bibliothèque Nationale et soutenue par la Conférence suisse des bibliothèques cantonales, cette initiative a réuni de nombreux partenaires cantonaux et donna lieu, en 2018, à la création de l'actuelle [e-newspaperarchives.ch] [4]. Parmi les initiatives helvétiques, il importe également de mentionner la plateforme Scriptorium dédiée à la presse vaudoise et créée en 2012 par la Bibliothèque cantonale et universitaire de Lausanne [5], et le projet pilote autour des archives du journal *Le Temps* [6]. À l'échelle européenne, au-delà de nombreuses campagnes de numérisation nationales [7], les projets européens Impact et Europeana Newspapers ont considérablement contribué à la numérisation et « OCRisation » de collections européennes au début des années 2010, pour certaines aujourd'hui

3 Paul Aron, Micheline Cambron, Gianni Haver, Marie-Ève Thérenty, François Vallotton, « Les Jeux olympiques de Berlin de 1936 dans la presse internationale. Présentation générale », in *Belphégor : Littérature populaire et culture médiatique,* 15, 1, 2017, pp. 1-5 ; Alain Clavien, *Grandeurs et misères de la presse politique*, Lausanne : Antipodes, 2010 ; Dominique Kalifa, *La civilisation du journal : histoire culturelle et littéraire de la presse française au XIX^e^ siècle*, Paris : Nouveau monde, 2011.

4 [https://www.e-newspaperarchives.ch].

5 [https://scriptorium.bcu-lausanne.ch] ; voir également la contribution de Silvio Corsini dans le présent numéro.

6 Voir letempsarchives.ch et Yannick Rochat *et al.*, « Navigating through 200 Years of Historical Newspapers », in *Proceedings of the 13th International Conference on Digital Preservation*, Berne : IPRES, 2016.

7 Aurélien Brossé, « Dans les collections de presse de la Bibliothèque nationale de France », in Claire Aslangul, Bérénice Zunino (éds), *La presse et ses images/Die Presse und ihre Bilder*, Berne : Peter Lang, (à paraître).

disponibles via Europeana [8]. Enfin, à une échelle plus globale, signalons également les campagnes de numérisation et d'acquisition de texte à très large échelle en Grande-Bretagne, aux États-Unis et en Australie [9]. Ce rapide survol des principales initiatives de numérisation laisse entrevoir l'ampleur du « tournant numérique » *(digital turn)* affectant les archives de presse ; un tournant décisif à bien des égards.

La numérisation représente de toute évidence une avancée majeure en termes de préservation et d'accès aux documents, et a un impact certain sur les pratiques de recherche [10]. Néanmoins, la recherche par mots-clés sur les textes « OCRisés » ne donne que rarement entière satisfaction et conduit bien souvent à d'innombrables faux positifs (documents retournés non pertinents au regard de la requête) et d'indétectables faux négatifs (documents non retournés pertinents au regard de la requête). Bien que plus facilement accessibles, les archives de journaux numérisés sont tout autant volumineuses que leurs équivalents papier et « feuilleter la presse ancienne par gigaoctet » n'est pas entreprise aisée [11]. Bien plus, la presse numérisée offre surtout la possibilité de déployer des techniques issues des domaines du traitement automatique du langage (TAL), de la vision par ordinateur et de la visualisation de données avec le potentiel de faciliter la recherche, la navigation, et la découverte du contenu informationnel de ces archives. Comment mettre en œuvre de tels traitements pour, au-delà de la recherche par mot-clé, permettre l'indexation sémantique et l'exploration de grandes collections de journaux historiques ? Ce fut là l'un des objectifs principaux poursuivis par le projet « Impresso – Media Monitoring of the Past », dont nous présentons ici une vue d'ensemble.

LE PROJET « IMPRESSO – MEDIA MONITORING OF THE PAST »

CADRE INSTITUTIONNEL ET OBJECTIFS

Impresso [12] est un projet de recherche dans lequel une équipe composée de linguistes informaticiens, d'historiens, de designers et de développeurs travaille à l'indexation sémantique de journaux historiques suisses et luxembourgeois, et à l'intégration des enrichissements obtenus dans une interface d'exploration adaptée aux pratiques de

8 [http://www.impact-project.eu et http://www.europeana-newspapers.eu].

9 *Chronicling America* (US) and *Trove* (Australia).

10 Bob Nicholson, « The Digital Turn », in *Media History*, 19, 1, 2013, pp. 59-73 ; Ian Milligan, « Illusionary Order : Online Databases, Optical Character Recognition, and Canadian History, 1997-2010 », in *The Canadian Historical Review*, 94, 4, 2013, pp. 540-569.

11 Claire-Lise Gaillard, « Feuilleter la presse ancienne par giga octets », in *Digitised Newspapers – A New Eldorado for Historians ?* », in Estelle Bunout, Maud Ehrmann, Frédéric Clavert (éds), *Studies in Digital History and Hermeneutics*, Berlin : De Gruyter, (à paraître).

12 [https://impresso-project.ch].

recherche historique. Financé par le Fonds national suisse, le projet est mené par le Digital Humanities Laboratory de l'EPFL, le Luxembourg Centre for Digital and Contemporary History de l'Université de Luxembourg, et l'Institute for Computational Linguistics de l'Université de Zurich [13], et est soutenu par un réseau de partenaires experts composé d'historiens, de bibliothèques, d'archives et de journaux [14].

Le projet s'articule autour de trois objectifs : l'amélioration et l'application de techniques d'extraction d'information issues du TAL afin de transformer des sources historiques au contenu textuel bruité et non structuré en données sémantiquement indexées ; le design et l'implémentation d'une interface permettant de rechercher, d'explorer et de visualiser les sources et leurs enrichissements sémantiques ; enfin, l'évaluation active et continue des outils produits avec un cas d'étude historique – la résistance à l'Europe –, ainsi qu'une réflexion sur l'utilisation des outils numériques dans les sciences historiques avec la considération des aspects méthodologiques, épistémologiques et pédagogiques.

Ce faisant, Impresso aborde les défis posés par les vastes collections de journaux numérisés, à savoir :

1) Des silos de journaux : les collections de journaux sont loin de toutes être numérisées, et lorsqu'elles le sont leurs modalités d'accès sont très hétérogènes en raison de restrictions légales et de contraintes liées aux politiques de numérisations ;

2) Des données volumineuses et désordonnées : les archives de presse numérisées sont composées de différents types de données (images, sorties d'OCR et métadonnées), lesquelles sont le plus souvent sous des formats très hétérogènes malgré l'existence de standards. Ces données sont par ailleurs fréquemment incomplètes et comportent de nombreuses incohérences (doublons, absence de contenu, indication de langue incorrecte, pages dans le mauvais ordre, etc.) ;

3) Textes historiques bruités : la qualité variable des sorties OCR, la segmentation des articles souvent défectueuse et le manque de ressources linguistiques appropriées affectent grandement la robustesse des algorithmes de traitement de textes et d'images [15]. Des processus de correction et d'évaluation de la qualité de l'OCR, et de normalisation de langues historiques sont nécessaires.

13 [https://www.epfl.ch/labs/dhlab/] ; [https://www.c2dh.uni.lu/ et https://www.cl.uzh.ch/en.html].

14 La section d'histoire de l'Université de Lausanne et infoclio.ch, le service suisse d'information pour les historiens ; les Bibliothèques nationales de Suisse et du Luxembourg, les Archives d'État du Valais, les Archives économiques suisses ; les journaux suisses *Le Temps* et *Neue Zürcher Zeitung*.

15 Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, *et al.*, « Assessing the Impact of OCR Quality on Downstream NLP Tasks », in *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, La Vallette : ICAART, 2020, pp. 484-496.

4) Exploration de contenu : à ce jour, peu d'interfaces de journaux favorisent la recherche et la découverte de contenu pertinent au sein de vastes volumes de données (sources et enrichissements), et presque tout est à inventer [16]. En outre, parce qu'une interface « contraint » ce que les chercheurs peuvent apprendre sur les sources et les contenus et façonne leurs flux de travail via outils et fonctionnalités, sa conception doit se fonder sur une évaluation attentive des besoins et pratiques des chercheurs en sciences humaines ;

5) Culture numérique : dans un contexte de recherche historique, l'évaluation critique des biais inhérents aux outils d'exploration, aux sources numérisées et aux annotations qui en sont extraites est primordiale pour une utilisation éclairée des données.

PRINCIPES DIRECTEURS DE L'INTERFACE

L'un des principaux résultats des efforts du projet Impresso est une interface d'exploration [17] de près d'une centaine de titres suisses et luxembourgeois [18] offrant des possibilités de recherche et d'exploration novatrices fondées sur de nombreux enrichissements sémantiques. Le public cible de cette application – dont la page d'accueil est illustrée en figure 1 – correspond principalement aux chercheurs universitaires en histoire et dans les disciplines connexes qui, n'ayant pas nécessairement d'expertise ou d'expérience avec les outils d'enrichissement automatique, sont néanmoins curieux de découvrir de nouvelles méthodes et de développer de nouvelles compétences.

Au-delà de la pure application de techniques automatiques (souvent fondées sur l'apprentissage automatique ou *machine learning*) aux sources, notre objectif est surtout de comprendre comment articuler la machine et le travail humain. En ce sens, l'application Impresso est avant tout un outil de médiation avec des sources historiques enrichies. D'un point de vue historique, son objectif principal est moins de permettre la découverte de motifs statistiques que de faciliter une exploration critique des sources et des données dérivées via des processus itératifs de recherche, de comparaison et de découverte. Pour ce faire, nous nous sommes appuyés sur trois principes directeurs interdépendants : la générosité, la transparence et la co-conception [19].

La nécessité d'attribuer plus d'importance à l'interaction avec les collections numérisées a été soulignée par Mitchell Whitehall avec sa proposition d'« interfaces

16 Maud Ehrmann, Estelle Bunout et Marten Düring, « Historical Newspaper User Interfaces : A Review », in *IFLA WLIC 2019 – Libraries : dialogue for change*, Athènes : IFLA, 2019. En ligne : [https://infoscience.epfl.ch/record/270246?ln=en].

17 Application : [https://impresso-project.ch/app]. Présentation vidéo : [https://go.epfl.ch/impresso-clip].

18 76 titres en avril 2021, 93 pour la prochaine échéance.

19 Pour plus de détails, voir les publications relatives sur [https://impresso-project.ch].

Figure 1 : Page d'accueil de l'application Impresso [https://impresso-project.ch/app].

générauses » [20]. Nous reprenons cette notion et utilisons le terme de *générosité* pour décrire des techniques exploratoires qui, au-delà des fonctionnalités de recherche par mot-clé, de filtrage et de navigation, « ouvrent » les collections et aident les utilisateurs à découvrir un contenu pertinent qu'ils n'auraient pu anticiper trouver, ou même savoir comment chercher. En un mot, il s'agit ici d'aider à découvrir.

L'intégration d'outils de traitement automatique des sources dans les pratiques de recherche historique soulève des défis épistémologiques et méthodologiques importants : les outils et les données extraites sont loin d'être neutres et il convient de favoriser une utilisation réflexive et critique des sources numérisées, des outils et des interfaces [21]. Aussi avons-nous particulièrement investi dans la documentation et le matériel didactique, avec trois priorités : informations sur la provenance et la qualité des sources numérisées, sur les méthodes d'enrichissement, et sur les méthodes

20 Mitchell Whitelaw, « Generous Interfaces for Digital Cultural Collections », in *Digital Humanities Quarterly*, 9, 1, 2015.

21 Andreas Fickers, « Towards A New Digital Historicism ? Doing History In The Age Of Abundance », in *VIEW Journal of European Television History and Culture*, 1, 1, 2012, pp. 19-26 ; Ian Milligan, *History in the Age of Abundance ?* Montréal : McGill-Queen's University Press, 2019 ; Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen, « Toward a Model for Digital Tool Criticism : Reflection as Integrative Practice », in *Digital Scholarship in the Humanities*, 34, 2, 2019, pp. 368-385.

d'exploration et de visualisation. Cette *transparence* a pour objectif d'aider à comprendre et à s'approprier les données et les outils [22].

Enfin, la *co-conception* décrit notre pratique d'interaction entre experts de différentes disciplines, et ce durant l'intégralité du processus de développement, de la recherche d'idées initiale à la résolution de problèmes en passant par l'évaluation et la prise de décision. Cette coopération continue, notamment entre les linguistes informaticiens, les designers, les historiens et les développeurs, a été cruciale pour le développement de fonctionnalités et de visualisations appropriées.

Au final, nous proposons de définir la « fouille de texte critique » *(critical text mining)* comme le résultat de la considération conjointe de ces principes et pratiques, où les techniques de fouille de textes et de visualisation de données sont intégrées de manière transparente dans une interface co-conçue permettant d'explorer un corpus et des données documentés et d'articuler au mieux le *machine understanding* avec les pratiques générales de recherche historique. Nous avons tenté de mettre en œuvre une telle entreprise, et la section suivante donne un bref aperçu de la « fabrique » Impresso.

APERÇU DE LA « FABRIQUE » D'UNE INTERFACE D'EXPLORATION DE JOURNAUX

ACQUISITION DES SOURCES ET ÉVALUATION DES BESOINS

Le point de départ est double avec, d'une part, la constitution du corpus et, d'autre part, la collecte et évaluation des besoins des utilisateurs. L'acquisition des sources numérisées est un parcours semé de difficultés et qui revêt différentes facettes (institutionnelle, politique, juridique, technique). Sans s'appesantir sur ce processus coûteux en temps – mais gratifiant lorsque les données arrivent sur les serveurs – il importe de mentionner une autre source de difficulté, à savoir la multiplicité des formats, tant des sources que des métadonnées. Notre approche est ici fondée sur la définition d'un format canonique et d'une librairie (programme informatique) modulaire. C'est à ce stade que des informations sur les lacunes de certaines archives sont récoltées afin de documenter au plus juste le corpus final disponible via l'interface [23].

Au regard de la collecte et évaluation des besoins des utilisateurs, un problème récurrent dans les projets interdisciplinaires comme Impresso est, d'une part, la difficulté

22 Via notamment : a) une page FAQ [https://impresso-project.ch/app/faq] b) des *« i-buttons »* documentant les fonctionnalités et processus sous-jacents c) un guide de découverte disponible sur la page d'accueil, et d) des vidéos explicatives sur la chaine *YouTube* du projet : [https://go.epfl.ch/impresso-youtube].

23 Sur le processus de prétraitement des données : Matteo Romanello, Maud Ehrmann, Simon Clematide et Daniele Guido, « The Impresso System Architecture in a Nutshell », in *EuropeanaTech Insights*, 2020. En ligne : [https://infoscience.epfl.ch/record/28359].

des historiens à formaliser et exprimer des besoins sans connaître au préalable et de manière exacte la technologie disponible (problème de l'œuf et de la poule) et, d'autre part, la difficulté des informaticiens à prédire exactement si et comment une technologie pourra soutenir le travail des historiens, et ce malgré leur excellente connaissance du potentiel de cette dernière. En application du principe de co-conception, des échanges continus ont eu lieu au sein de l'équipe principale, du consortium du projet, et de la communauté. De nombreux besoins ont émergé, parmi lesquels : évaluation de la qualité de l'OCR, possibilité de créer des sous-ensembles du corpus, importance des opérations de recherche, sauvegarde des requêtes, besoin de transparence des données, et nécessité de pouvoir comparer et exporter des articles, collections et/ou données. Au final, il est apparu que l'application Impresso devait permettre aux historiens de trouver les articles qu'ils recherchent, explorer librement le contenu du corpus disponible, et prendre conscience de ce qu'il est possible (ou non) d'y trouver. Ces besoins peuvent être regroupés en cinq activités principales, à savoir la recherche, la découverte, la collecte, la comparaison et l'évaluation critique des sources.

EXTRACTION D'INFORMATION

Plusieurs techniques d'extraction d'information sont appliquées au corpus Impresso. Du point de vue du TAL et de la vision assistée par ordinateur, chacune constitue un champ de recherche à part entière, et leur application à des textes historiques bruités (par exemple avec de nombreuses erreurs dues à l'OCR), multilingues et hétérogènes – qui plus est ici à très large échelle – constitue un véritable défi.

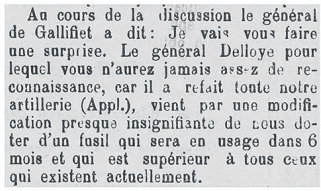
Traitement lexical et indexation textuelle – De nombreux processus sont mis en œuvre afin de préparer le matériel textuel, de la reconnaissance de la langue des articles à la mesure de la qualité de l'OCR, en passant par la normalisation des variantes morphologiques [24] et l'amélioration de l'OCR [25]. L'indexation du matériel textuel, permettant la recherche standard par mots-clés couplée à des facettes sur les métadonnées, est réalisée selon un modèle de sources unifié servant de base à tout le reste.

Traitement des entités nommées – Les unités référentielles telles que les noms de personne, de lieu, et d'organisation sous-tendent la sémantique des textes et guident

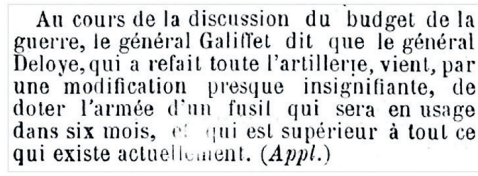
24 Peter Makarov, Simon Clematide, « Semi-Supervised Contextual Historical Text Normalization », in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7284-7295.

25 Phillip Ströbel, Simon Clematide, Martin Volk, « How Much Data Do You Need ? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR », in *Proceedings of the 12th Language Resources and Evaluation Conference*, *op. cit*, pp. 3551-3559.

Facsimiles



Au cours de la discussion le général de Gallifiet a dit: Je vais vous faire une surprise. Le général Delloye pour lequel vous n'aurez jamais ass-z de reconnaissance, car il a refait toute notre artillerie (Appl.), vient par une modification presque insignifiante de nous doter d'un fusil qui sera en usage dans 6 mois et qui est supérieur à tous ceux qui existent actuellement.



Au cours de la discussion du budget de la guerre, le général Galiffet dit que le général Deloye, qui a refait toute l'artillerie, vient, par une modification presque insignifiante, de doter l'armée d'un fusil qui sera en usage dans six mois, et qui est supérieur à tout ce qui existe actuellement. (*Appl.*)

OCR

An cours de la discussion le général de Gallifiet a dit: Je vab voua faire une surprise. Le général Delloye pour lequel vous n'aurez jamais ass-z de reconnaissance, car il a refait tonte notre artillerie (Appl.), vient par une modification presque insignifiante de nous doter d'un fusil qui sera en usage dans 6 mois et qui est supérieur à tous Ctux qui existent actuellement.

Au cours de la discussion du budgel de la guerre, le général Gaiilïet dit que le général Deloye, qui a refait toute l'artillerie, vient, par une modification presque insignifiante, de doter l'armée d'un fusil qui sera en usage dans six mois, e ! qui esl supérieur à toul ce qui existe actuel lt-ment. {Appl.)

Figure 2 : Exemple de détection de text reuse entre *L'Indépendance Luxembourgeoise* du 21 février 1900, p. 3 (gauche) et *L'Impartial* du 22 février 1900, p. 3 (droite), avec une déclaration du général Galliffet à la Chambre de Paris à propos de l'armement. Malgré des formulations différentes et des erreurs d'OCR, le passage repris est détecté (également dans deux autres titres). (URL du TR cluster correspondant [juin 2021] : [https://impresso-project.ch/app/text-reuse-clusters/?sq=&clusterId=tr-nobp-all-v01-c60129690381&q=galiffet&page=1].)

leur interprétation. Leur reconnaissance et désambiguïsation automatiques, bien que mises en difficulté face aux textes historiques [26], facilitent grandement la recherche d'information dans des collections textuelles.

Topic Modeling – Le *topic modeling* permet de déterminer les thèmes ou *topics* présents dans des documents. Un nombre déterminé de *topics* est extrait du corpus via un apprentissage probabiliste, et chaque article se voit attribuer un ou plusieurs topics. Chaque topic est défini par une liste de mots représentatifs, comme par exemple le topic « cours, école, enseignement, professeur, année, classe, formation, instruction,… » faisant référence à l'éducation publique.

Text Reuse – Le *text reuse* (ou réutilisation de textes, ci-après TR) correspond à la réitération significative d'un (segment de) texte, généralement au-delà de la simple

26 Maud Ehrmann, Ahmed Hamdi, Elvys L. Pontes, *et al.*, « Named Entity Recognition and Classification on Historical Documents : A Survey », in *ACM Computing Surveys (CSUR)*, [https://arxiv.org/abs/2109.11406].

répétition du langage courant. Dans le contexte de la presse, la détection automatique de TR permet de mettre en évidence l'existence de plusieurs copies d'un même article – identiques ou légèrement différentes – ainsi que les répétitions d'extraits d'articles ou de pages (annonces, publicités, etc.). Cela peut par exemple faciliter l'étude des communiqués émanant d'agences de presse, dévoiler des manipulations de contenu, ou révéler des orientations éditoriales différentes. La figure 2 offre un exemple de sortie de l'outil de TR, avec la détection d'une reprise dans divers titres d'une déclaration du Général Galliffet.

Similarité visuelle – En complément des processus d'analyse de texte – focus principal du projet –, une technique de calcul de similarité d'image est utilisée pour faciliter l'exploration des images présentes dans le corpus.

Système de recommandation – Enfin, nous avons également développé un système expérimental de recommandation de contenu fonctionnant à partir des multiples couches d'annotations sémantiques décrites ci-avant. À partir de collections personnelles d'articles, les utilisateurs peuvent amorcer une recherche de contenu similaire, y compris en sélectionnant et ajustant le poids de telle ou telle dimension [27].

Ces techniques d'extraction d'information produisent de multiples couches d'annotations sémantiques, lesquelles sont combinées et accessibles via une interface unique [28]. Cette union de sources et de nombreux enrichissements ne se fait pas au hasard ou de manière opportuniste, mais repose sur un design réfléchi et collaboratif.

DESIGN DE L'APPLICATION

Le design de l'application s'est effectué à plusieurs mains sur la base de prototypes concrets et testables développés en plusieurs phases. Le défi majeur de la conception de l'interface fut *d'intégrer de manière cohérente les différents types de données enrichies*, autrement dit de déterminer comment les utilisateurs peuvent interagir avec les documents, les images et les enrichissements, utiliser ces derniers pour faciliter leur processus de recherche, ou encore passer d'un composant de l'interface à l'autre sans avoir le sentiment de se perdre. À cet égard, la conception de l'interface a été guidée par deux objectifs principaux : offrir un maximum de liberté à l'utilisateur dans la création de requêtes via tous les composants de l'interface tout en gardant un niveau de complexité acceptable, et permettre une intégration souple et transparente des modes de lectures distante et attentive *(distant and close reading)*.

27 Pour ces processus, voir les courtes vidéos explicatives : [https://go.epfl.ch/impresso-youtube-forumz].

28 Maud Ehrmann, Matteo Romanello, Simon Clematide, *et al.*, « Language Resources for Historical Newspapers : The Impresso Collection », in *Proceedings of the 12th Language Resources and Evaluation Conference*, *op. cit*, pp. 958-968.

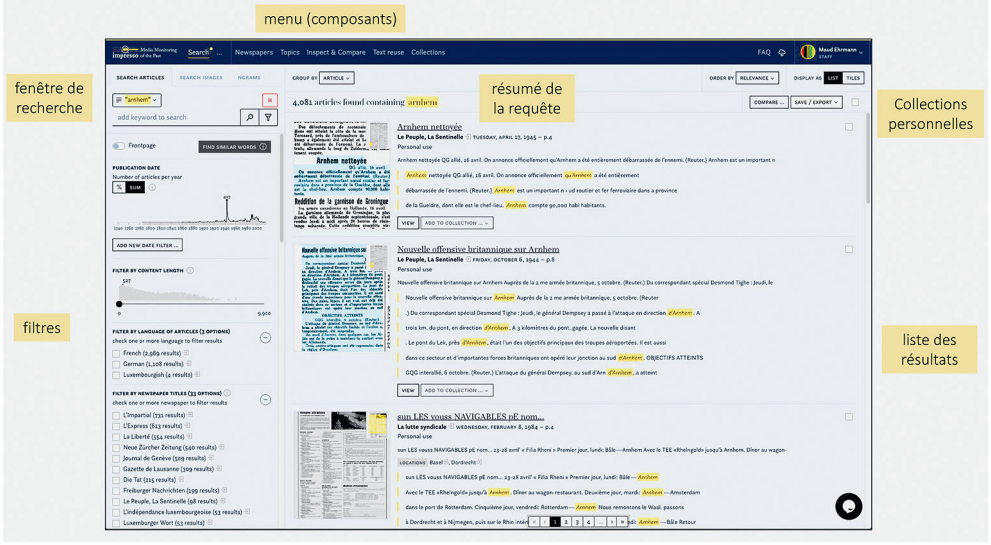


Figure 3 : Vue de la page « search » avec la requête « arnhem ».

QUELQUES COMPOSANTS DE L'INTERFACE

Pour finir, nous décrivons très brièvement deux composants de l'interface.

Recherche. Exigence essentielle des historiens, le puissant composant de recherche combine tout à la fois la recherche, des capacités de filtre fondées sur les enrichissements, de multiples opérateurs et des fonctionnalités « généreuses », et rend obsolète une fonction de recherche avancée séparée.

Prenons pour exemple une recherche sur la couverture de la bataille d'Arnhem dans la presse de 1944 à nos jours. Cette étude débuterait par une recherche à partir du mot-clé *arnhem*. Cette requête renvoie environ 4000 articles, et un coup d'œil rapide à leur distribution confirme un pic d'occurrences en 1944, suivi d'une présence irrégulière les années suivantes. L'image ci-dessus offre une vue d'ensemble de la page de recherche avec cette requête. Soulignons ici deux possibilités permettant d'affiner cette recherche. Tout d'abord le module de suggestion de mots-clés, fondés sur des *words embeddings* propres à capturer des similarités orthographiques et sémantiques, propose à l'utilisateur les mots similaires suivants : *amhem* (faute d'OCR sur « rn »), *arnheim* (nom allemand), ou encore *arnehm* (typo originale ?). La figure 4 illustre ces suggestions, dont la prise en compte permet d'intégrer des articles mentionnant le mot-clé mal transcrit (avec alors environ 6800 résultats). Ensuite, une inspection rapide des filtres de recherche et notamment des « topics » liés à ces résultats permet de se rendre compte que certains articles ont trait à l'équipe de football d'Arnhem, sujet peu pertinent au regard de l'étude.

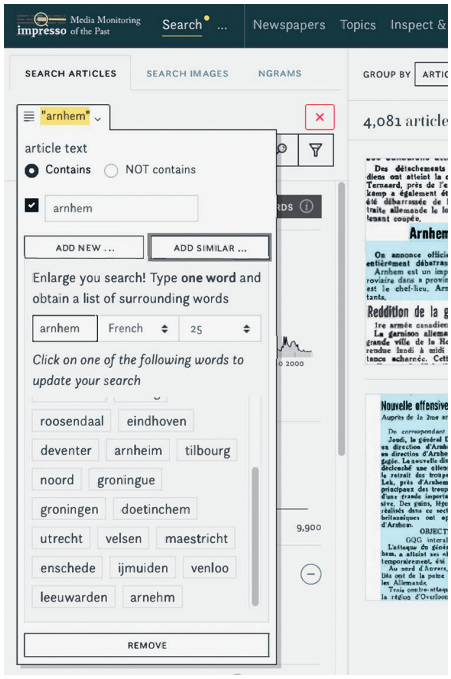


Figure 4. Suggestion de mots-clés pour la requête « arnhem ».

L'utilisateur peut ici exclure les topics liés au sport (e.g. « match, équipe, ligue, club… ») ou sélectionner les topics relatifs à la guerre. Le set d'articles redescend à environ 1500, est garanti « non sport » et couvrant les erreurs d'OCR principales.

Viewer. Un composant indispensable de l'interface est bien sûr la « visionneuse » des sources. Sans donner plus de détails sur cet élément classique, mentionnons tout de même quelques spécificités : une vue « page » donnant à voir les annotations les plus saillantes en tant que *marginalia*, et une vue « article », donnant à voir les annotations et les passages faisant partie d'un groupe de *text reuse*, et offrant, entre autres, la possibilité de sauver un article dans une collection personnelle. Les figures 7 et 8 illustrent ces vues.

Nous laissons au lecteur l'opportunité d'explorer plus avant l'interface en ligne, où il pourra également découvrir l'existence des composants suivants :

- *Inspect & Compare*, permettant de comparer de manière approfondie deux requêtes ou collections personnelles ;
- Un explorateur de titres, donnant des informations sur les métadonnées, la distribution temporelle, ou encore les données manquantes des titres composant le corpus ;
- Un explorateur de fréquence de mots et de statistiques sur les données ;
- Un moteur de recherche visuel ;
- Une importante documentation (« i-buttons » et FAQ).

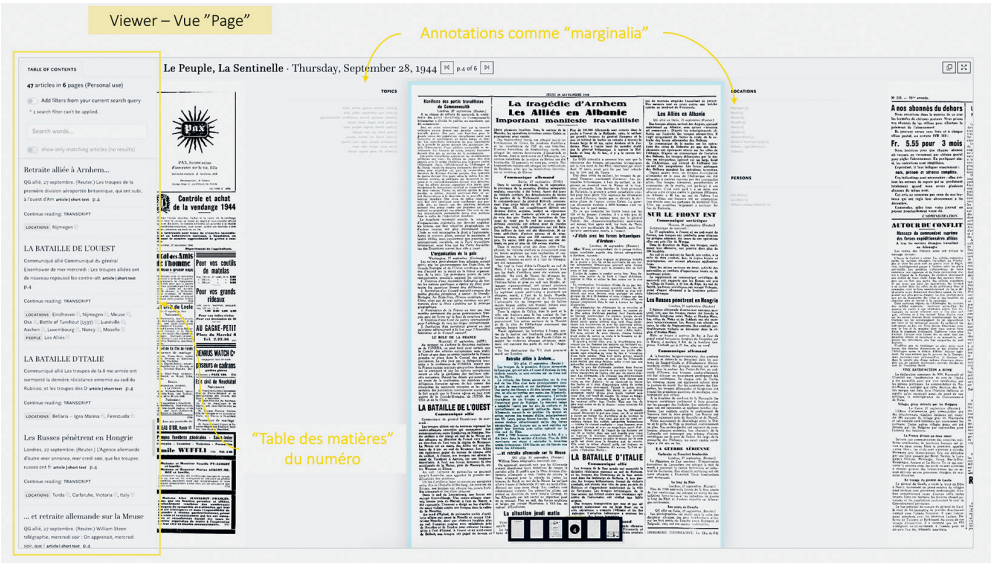


Figure 5. Vue « page » de la visionneuse.

Figure 6 : Vue « article » de la visionneuse.

POUR CONCLURE

Comment apprécier la valeur ajoutée de l'interface Impresso pour la recherche, la collecte, la comparaison et la découverte d'information, par opposition aux pratiques courantes fondées sur la recherche par mots-clés et la lecture attentive *(close reading)* ? La réponse à cette question ne peut être que partielle à ce stade, mais il est possible d'apporter quelques éléments de réponse. Premièrement, il nous semble opportun sur ce point de distinguer deux types d'information : des informations « explicites » d'une part (par exemple suggestion de mots-clés, *marginalia* de la visionneuse d'articles, filtres par langue ou type de contenu), qui sont déjà utilisées dans les pratiques existantes et pourraient, avec beaucoup d'effort, être obtenues manuellement ; et des informations « implicites » d'autre part (topics, TR, similarité d'images, capacité de comparer), qu'il serait prohibitif, voire impossible, de compiler manuellement. Dans le corpus global, ces dernières révèlent différents types de relations entre des éléments spécifiques et des éléments connexes, et ne seraient pas accessibles dans un cadre de recherche « traditionnel ». Deuxièmement, ce qui rend l'application particulièrement intéressante pour la recherche historique est le fait que ces informations, via les composants de l'interface, peuvent *interagir* les unes avec les autres. En effet, le processus de construction itérative de requêtes combinant divers éléments permet aux chercheurs d'observer et d'analyser les sources sous de multiples perspectives. Au-delà de leur dimension informative, ces dernières peuvent donc également conduire à reformuler ou nuancer la ou les questions de recherche initiales sur la base de nouvelles connaissances, et peuvent elles-mêmes être transformées en objets de recherche tangibles sous la forme de collections.

Comme tout outil de travail (numérique ou non), l'interface Impresso a des limites : les sources disponibles au travers de l'interface correspondent à un ensemble limité ; la qualité des données (OCR ou performances des outils de TAL) est encore imparfaite ; l'interface elle-même peut être perçue comme trop complexe (nous avons volontairement choisi de ne pas réduire ou cacher la complexité des données et des outils) ; et enfin, la persistance des sources numérisées et de leurs enrichissements n'est pas garantie dans l'absolu et demeure une question ouverte. Si certaines de ces limites peuvent être dépassées à plus ou moins court terme, d'autres participent d'une transformation des méthodes de travail de plus longue haleine. Dans tous les cas, y compris avec des sources et enrichissements parfaits, une interface n'est qu'une étape parmi d'autres dans le travail de l'historien, qui commence bien en amont et se poursuit après.

Au final, tant au niveau du traitement des données que du design et des pratiques de recherche, il apparait que les enrichissements TAL peuvent être combinés de manière surprenante et efficace avec les fonctionnalités traditionnelles de recherche de

documents, ce qui n'était pas évident au début du projet. L'interface Impresso offre ainsi une exploration multidimensionnelle de larges collections de journaux numérisés et de données afférentes extraites automatiquement. Sur la base de nombreuses techniques d'analyse de textes et d'images, cette application permet, d'une part, de faciliter les processus de recherche, de découverte et de comparaison de sources pour toute question historique s'appuyant tout ou partie sur les archives de presse et, d'autre part, d'intégrer la complexité des sources numérisées et outils dans le travail des historiens et d'en faire, progressivement, des opportunités tangibles pour les sciences humaines et sociales.