IMI Call for Projects - Proposal HES-ESH

Titre du projet

Original Journalistic Information Label and Traceability: an approach to label content and enable information traceability in order to valorize journalistic data and fight against Fake News

Résumé

Dans un environnement digital où chaque individu peut devenir créateur d'information, il devient primordial de pouvoir identifier et authentifier le contenu original produit par les journalistes afin de le différencier, le valoriser et propager ainsi une information de qualité, traçable, mais aussi différenciable des *fake news*.

Mots clés

Information Label, Information Traceability, Security Management, Standards, Fake News

Description complète du projet

Contexte général

Les médias d'information sont confrontés à de profondes mutations dues au développement des technologies numériques, aux nouvelles consommations digitales de l'information et aux nouveaux médias concurrents qui en sont issus. La digitalisation a mis sous tension les industries médiatiques. Elle a modifié la façon de produire des contenus, de les diffuser et de les consommer.
La digitalisation a permis aux institutions et aux marques d'être en relation directe avec leurs audiences. Les outils digitaux, comme les réseaux sociaux, ont également permis à tout un chacun de pouvoir être créateur d'information. Cet élément important a accéléré la diffusion d'informations et a créé un environnement multi-sources.
La façon dont le consommateur se forge une opinion sur le monde a changé. Il devrait trier et filtrer l'information, mais il est en réalité confronté à une bulle d'informations. Il en résulte la dissémination de fausses nouvelles (*fake news*). Une telle désinformation a des effets extrêmement négatifs sur les individus et la société.

Label qualité suisse

En 2017, la Commission fédérale des médias (COFEM) a partagé dans un rapport[1] sur l'avenir des médias, son intérêt à mettre sur pied un label de qualité qui permettrait de reconnaître les contenus journalistiques satisfaisant aux normes de qualité minimales et de les distinguer des offres alternatives telles que les blogs ou les stratégies numériques de communication commerciale (marketing de contenu, native advertising, branded content, etc.).

[1] https://www.emek.admin.ch/fr/actuel/apercu/

Ce rapport souligne que “les utilisateurs ont toujours plus de difficultés à distinguer les contenus véritablement journalistiques des contenus qui ne satisfont pas aux exigences minimales d'un journalisme professionnel (par exemple la vérification des sources, la pertinence, la distance, l'exactitude, la classification et la transparence)”.

Identification, authentification et traçabilité de l’information

Comme c’est déjà le cas dans beaucoup de domaine, la mise en œuvre d’une identification et traçabilité de l’information permet de valider les contenus et leurs sources. L’attribution d’un identifiant étant réalisée dans un système sécurisé qui repose sur l’authentification des auteurs. Nous savons alors précisément “qui a dit quoi”, à quel endroit trouver cette information (un article de journal par exemple), mais aussi comment y faire référence de manière sûre, sur le long terme, et ainsi expliciter la source de ce que l’on commente, partage ou prétend.

La traçabilité repose sur deux composantes essentielles:

- un **identifiant universel** et **pérenne** de l’objet en question (un article par exemple), soumis à un système d’authentification
- les **métadonnées**, liées à cet identifiant, et donnant les informations de contexte sur cet objet (l’auteur, identifié lui-aussi, la date, etc.)

Dans l’idéal l’identifiant et les métadonnées sont accessibles sur le Web, et ceci sur le long terme.

Des mécanismes d’identification et de traçabilité sont déjà implémentés dans l’industrie et le commerce, mais aussi pour l’information textuelle dans le monde des librairies ou des publications scientifiques par exemple. Différentes solutions techniques ont fait leurs preuves pour l’attribution d’identifiants universels et la gestion de métadonnées:

- GS1 pour ce qui est de la supply chain[2] (code-barre, QR code, etc.)
- L’ISBN pour l’édition de livres
- Digital Object Identifier (DOIs)[3] utilisé dans divers domaines tels les articles scientifiques, les oeuvres du cinéma et de la télévision (EIDR[4]), etc.
- Identifiers.org, “un système de résolution bien établi qui permet de référencer des données pour la communauté scientifique, en mettant l’accent sur le domaine Life Sciences.”
- L’ISSN[5] utilisé pour identifier les journaux, magazines et périodiques de toutes sortes, imprimés ou électroniques, ainsi que son complément Serial Item and Contribution Identifier (SICI[6]) pour les données spécifiques à un article. En Suisse le centre ISSN[7] est hébergé à la Bibliothèque nationale (BN)

[2] https://www.gs1.ch/fr/home/th%C3%A8mes/tra%C3%A7abilit%C3%A9-avec-les-standards-de-gs1
[3] https://www.doi.org/
[4] https://eidr.org/
[5] http://www.issn.org/understanding-the-issn/what-is-an-issn/
[6] https://en.wikipedia.org/wiki/Serial_Item_and_Contribution_Identifier
[7] https://www.nb.admin.ch/snl/fr/home/informations-professionnels/issn.html#-1253990407

- L'utilisation des technologies Web classiques (URLs, déréférencement, etc.) qui sont au cœur du Linked Data[8]

Des métadonnées essentielles pour le traitement informatique

Pour faire face à la masse d'information, il serait fort utile si les machines pouvaient assister le lecteur. Malheureusement nous nous retrouvons devant un problème classique en informatique où il s'agit de gérer des données que la machine ne comprend pas: le texte en langage naturel. Malgré les avancées dans le domaine, il est de nos jour en effet toujours très difficile pour un programme de manipuler de l'information non structurée, et deux approches opposées, mais complémentaires, sont au cœur de la recherche actuelle: soit aider la machine à comprendre le texte naturel (Traitement automatique du langage naturel[9], apprentissage automatique[10], Fact Checking, etc.), soit mettre à disposition de la machine des données structurées qu'elle est capable de manipuler et permettre ainsi la mise en œuvre de nouvelles solutions logiciel. Cette deuxième approche est promue depuis plus de 15 ans par le W3C lui-même et porte le nom de Web des données[11] (aussi connu sous le nom de "Web 3.0", "Web sémantique", "Linked Data").

En ce qui concerne les métadonnées, le New York Time et le International Press Telecommunications Council (IPTC) ont d'ailleurs contribué à une solution, et ceci en 2012 déjà. Dans cette annonce[12] ils décrivent clairement la problématique ainsi que leur solution qui est actuellement totalement intégrée dans une mise en application du Linked data qui connait un grand succès: schema.org[13]. Il est à noter que l'approche schema.org, initiée et soutenue par les grands moteurs de recherche (Google, Microsoft, Yahoo, Yandex), permet de faire d'une pierre deux coup en fournissant des métadonnées exploitables par les machines, tout en améliorant la visibilité du contenu en matière d'optimisation pour les moteurs de recherche (SEO). Des recommandations spécifiques pour le domaine des news sont d'ailleurs expliquées[14].

Le marché

Le marché des médias en Suisse romande a la particularité d'avoir plusieurs sociétés anonymes actives sur une zone de diffusion privilégiée, allant d'un canton ; Vaud, Genève, Valais, Fribourg, Neuchâtel ou Jura, à une région. Cette régionalisation du marché reste importante tant au niveau du contenu que des audiences. Le public a la particularité de consommer du contenu spécialisé au niveau international tout comme du contenu local.
À noter que notre partenaire, le groupe ESH Médias bénéficie d'un lectorat principalement régional mais désire aborder la réflexion de façon globale en répondant aux besoins de toutes les strates de presse, régional, national et international. En considérant la spécificité du marché, notre projet de recherche aborde un approche participative inter-médias afin d'aboutir à des résultats bénéficiant à l'ensemble de la branche (par ex. standards).

---

[8] https://www.w3.org/DesignIssues/LinkedData.html
[9] https://fr.wikipedia.org/wiki/Traitement_automatique_du_langage_naturel
[10] https://fr.wikipedia.org/wiki/Apprentissage_automatique
[11] https://fr.wikipedia.org/wiki/Web_des_donn%C3%A9es
[12] https://open.nytimes.com/rnews-is-here-and-this-is-what-it-means-ea25f13417d7
[13] https://schema.org/
[14] https://schema.org/docs/news.html

Description du projet (objectifs, méthodes et livrables)

Ce projet de recherche appliquée vise dans un premier temps à mener différentes réflexions autour de la mise en avant de contenus originaux journalistiques qui repose sur la création d'un label permettant de promouvoir du contenu de qualité et lutter contre la désinformation. Dans un deuxième temps une solution technique sera élaborée pour gérer la certification du contenu et l'obtention de ce label (Figure 1). Cette solution sera ensuite implémentée dans un système "Proof-of-Concept" (POC) qui permettra de démontrer la publication en ligne d'un article labellisé permettant la traçabilité de l'information soit par le lecteur soit de manière automatisée (par logiciel informatique). Le POC sera finalement expérimenté et évalué pour fournir un retour complet sur l'ensemble du travail réalisé.

Le projet aborde une approche appliquée et collaborative afin de développer un POC fonctionnel adapté aux besoins d'ESH médias et applicable aux autres partenaires de l'IMI. Les préconisations seront partagées aux autres membres de l'IMI.

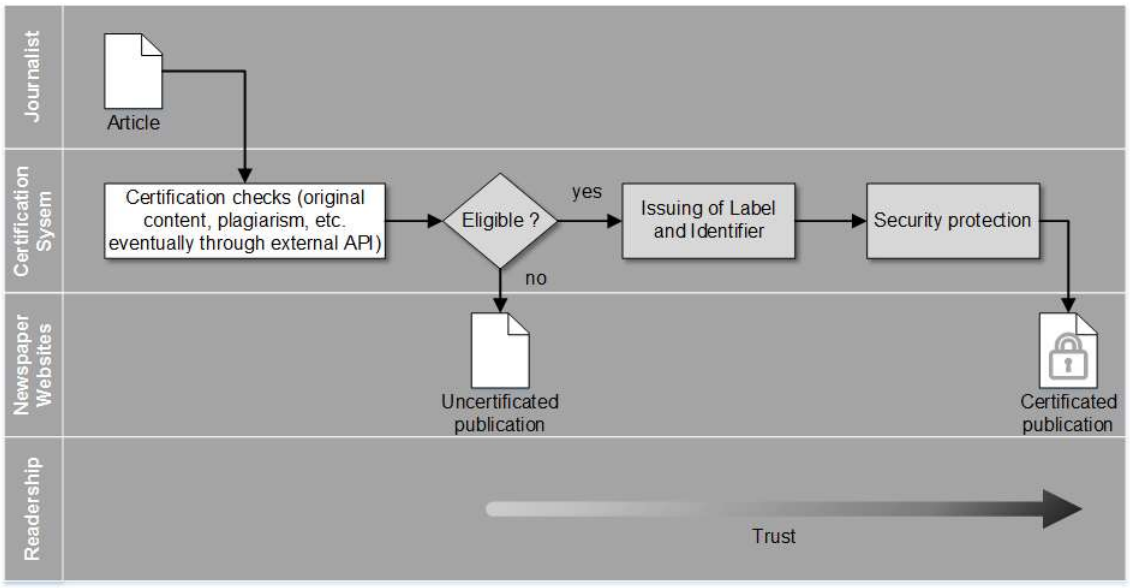

Figure 1 - Processus envisagé de certification et publication labellisée

Objectif général

Le projet a pour objectif de comprendre empiriquement l'impact de la mise en place d'un label sur le lectorat suisse, d'identifier le potentiel économique lié et de mettre en place un système d'authentification et de traçabilité sécurisé du contenu original grâce aux métadonnées générées par la publication de contenu journalistique en ligne.

Ce projet vise à fournir des réponses aux questions de recherche suivantes :

1. Est-ce que la mise en place d'un label promouvant la qualité d'un contenu comme original augmente le taux de lecture chez le lectorat ? Quel type/forme de label impacte au mieux le lectorat ?
2. Quels sont les méthodes et systèmes les plus optimaux afin de pouvoir certifier et authentifier l'originalité d'un contenu journalistique ?

3. Quelle est l'approche la plus optimale pour le déploiement d'un système de certification inter-médias au niveau régional, national et international ?

WP1 - Perception et impact du contenu journalistique labellisé sur le lectorat

Objectifs :

- Mieux comprendre les mécanismes de consommation d'information et de confiance du lectorat suisse romand
- Identifier les facteurs de différenciation entre du contenu original et non-original
- Définir le format de label impactant positivement le lectorat (user centered design)

Méthode :
Dans ce premier volet, nous souhaitons mener une approche exploratoire qualitative à travers des entretiens semi-directifs. L'échantillon sera constitué de 20 individus en Suisse romande. La représentativité de l'échantillon sera réalisée par critères (âge, ville/campagne). Nous souhaitons également nous appuyer sur le réseau des partenaires IMI afin de pouvoir bénéficier d'une homogénéité du lectorat romand. Les résultats de cette première étape, nous permettront d'identifier quel est le format de label impactant de façon la plus optimale le lectorat suisse romand.

Livrables :

- Résultats sous forme de synthèse des entretiens réalisés
- Recommandations

WP2 - Systèmes de certification : exigences et architecture

Objectifs :

- Analyser l'existant: comprendre les processus humains et technologiques de publication des médias partenaires
- Réaliser l'état de l'art des méthodes de certification, applicables à des articles de journaux publiés en ligne
- Réaliser l'état de l'art des systèmes d'attribution d'identifiants uniques et de métadonnées, applicables à des articles de journaux publiés en ligne
- Déterminer les exigences et l'architecture pour le système de certification des articles de journaux publiés en ligne, en se basant sur le résultat du WP1, ainsi que l'état de l'art et l'analyse de l'existant de ce WP 2

Méthode :
Nous acquerrons une connaissance du domaine en discutant avec les spécialistes et en analysant les solutions techniques mises en œuvre autour de la publication d'articles et de métadonnées sur leurs sites Web. A partir de là, et en nous basant sur notre expertise en matière de cybersécurité et Web des données, nous étudierons comment appliquer les solutions techniques existantes (certification, signature électronique, linked data, schema.org, etc.) en vue d'identifier, authentifier et permettre le traçage du contenu des médias.

Sur cette base nous pourrons alors définir comment implémenter une certification du contenu, selon les attentes du lectorat (WP1), et qui soit garante d'un travail journalistique professionnel (vérification des sources, plagiat, etc.). Du point de vue technique nous

tiendrons compte des contraintes imposées par les solutions de publication utilisées actuellement par les différents médias.

Livrable :
Rapports décrivant:

- Analyse et évaluation des méthodes de certification, incluant l'attribution d'identifiants et gestion des métadonnées
- Analyse et description des processus de publication en ligne
- Proposition d'architecture pour le système de certification, compatible avec les processus existants de publication

WP3 - Évaluation des menaces et gestion de la sécurité

Objectifs :

- Réaliser une analyse de risque/menace
- Déterminer la sécurité à mettre en œuvre: à quel niveau et de quelle manière

Méthode :
Une analyse de risque/menace nous permettra de déterminer les exigences au niveau de la sécurité. En déterminant les attaques possibles, les moyens de l'adversaire, ainsi que les conséquences d'une attaque nous pourrons déterminer quelle technologie mettre en place dans ce contexte.

Livrable :

- Rapport sur l'analyse de risque et la gestion de la sécurité

WP4 - Conception et développement d'un système de labellisation (POC)

Objectifs :

- Développement d'un POC du système de labellisation
- Implémenter le POC dans une simulation la plus réelle possible du processus de publication en ligne d'un média

Méthode:
Nous proposerons une solution technique que nous mettrons en œuvre dans un POC fonctionnel d'identification et traçabilité de l'information en vue de la certification. La solution sera élaborée selon les méthodes agiles de développement, en collaboration avec les principaux intéressés, à savoir les médias et leurs journalistes qui produisent le contenu, mais aussi les lecteurs qui le consomment.

Livrables :

- Prototype du système de labellisation (exécutable et code source)
- Documentation technique

WP5 - Démonstration, expérimentation et évaluation

Objectifs :

- Démontrer le système aux médias partenaires
- Expérimenter le système (par les médias partenaires)
- Evaluer le système

Méthode :
Le POC sera démontré aux médias partenaires qui pourront alors le tester dans la publication d'articles sur ce système de test. Nous l'évaluerons ensemble afin de compléter le rapport et fournir un retour complet sur le travail réalisé.

Livrable :
- Rapport d'évaluation du système de certification

WP6 - Gestion du projet et valorisation

Objectifs :
- Optimiser la gestion de projet pour assurer des résultats de qualité à destination des partenaires de l'IMI
- Coordonner la valorisation du projet pour amener une réflexion nationale

Méthode :
À travers une gestion de projet agile, nous souhaitons développer un projet en respectant le budget, les délais et la vision de l'IMI. Nous nous assurons de bien aligner les différentes initiatives stratégiques afin d'atteindre les objectifs de départ, en minimisant les risques et en évitant les dépassements de coûts. La valorisation du projet a également un lien direct sur la réalisation et le succès du projet. Nous souhaitons mettre en avant les résultats auprès des membres de l'IMI et valoriser le projet au niveau national.

Livrables :
- Une publication scientifique
- Un workshop interne IMI
- Communications médias

Résultats attendus

Les principaux résultats sont:
- Proposition d'un label indépendant qui permette la certification de contenu à l'échelle nationale
- Démonstration d'un système de certification permettant de gérer techniquement ce label

Nos résultats permettront d'identifier le potentiel avec l'instauration d'un label certifiant le contenu journalistique professionnel. Cela permettra aux médias de gagner en attractivité, d'accroître le nombre d'annonceurs et de fournir des informations certifiées rapidement.

Le label reposera sur plusieurs briques technologiques dont nous démontrerons la mise en œuvre:
- L'attribution d'identifiants universel et pérenne aux articles de journaux
- L'authentification des sources
- L'accès en ligne à moyen et long terme aux articles et leurs métadonnées, à travers leur identifiant
- La publication en ligne d'articles incluant le label

De plus, le système de certification mise en œuvre permettra de faciliter la citation d'articles de journaux labellisés, ainsi que d'améliorer l'optimisation pour les moteurs de recherche (SEO):

- Faciliter la citation des sources par référence sur l'identifiant, en permettant par exemple qu'un tweet face une référence directe à un article labellisé
- SEO: Les métadonnées seront implémentées pour correspondre aux dernières recommandations des moteurs de recherche en matière d'indexation

Il résultera également de ce projet un POC fonctionnel de labellisation de l'information dans le milieu des médias.

Description de la dimension pluridisciplinaire

Le projet se veut pluridisciplinaire et s'appuie sur une collaboration entre la HES-SO Valais-Wallis, le groupe de presse ESH médias et les membres partenaires de l'IMI.

La plus-value pour la recherche

Alors que l'identification, l'authentification et la traçabilité de l'information a fait ses preuves dans bon nombre de domaines, de la supply chain aux publications scientifiques, son application dans le monde du journalisme et des articles de presse en est à son balbutiement.

Nous pourrons étudier la problématique et fournir une solution technique pour cette mise en œuvre qui reposera sur des briques technologiques ayant déjà fait leurs preuves, notamment la gestion d'identifiants universels, la signature électronique ou encore schema.org et le Linked data.

L'innovation pour les médias

Nous relevons qu'il n'existe pas de moyen fiable à l'échelle de la Suisse (voire à un niveau européen, voire universel) de pouvoir distinguer une information vérifiée et ayant fait l'objet d'un travail journalistique original, de tout autre contenu. Les marques médias ne sont pas suffisantes dans cette optique, nous recherchons un moyen indépendant des intérêts privés ou publics de certifier un article.

De fait, ce projet est innovant dans la mesure où il s'agirait ainsi de la première certification de contenu à une échelle nationale.

Portée sociétale du projet

La prolifération des fake news et la désinformation sont des risques avérés pour les démocraties (cf dernières élections présidentielles américaines ou mouvement des gilets jaunes en France). A notre échelle, on perçoit localement une inquiétude, une prise de conscience de la population romande/suisse face à ces risques. Nous constatons déjà que le phénomène de la désinformation se retrouve également sur les réseaux sociaux en Suisse.

Valoriser les contenus permet de renforcer le modèle économique d'un média d'information et ainsi d'assurer le bon fonctionnement d'une démocratie. Dans la continuité, cela doit permettre au public d'avoir de nouveau confiance dans les médias.

Bénéfices concrets envisagés pour les partenaires médias de l'IMI

- Un tel projet permet à nos médias de rendre facilement identifiable un contenu vérifié et original.
- Ces contenus seraient distingués et par conséquent renforceraient la valorisation du travail journalistique. A terme cette certification et la communication faite autour doivent nous permettre de regagner la confiance du public et légitimer la nécessité de payer pour de l'information de qualité.
- En proposant une approche open source, transparente et éthique, nous visons un taux d'acceptation du projet le plus haut possible, limitant les possibles remises en question.
- Tous les membres de l'IMI doivent pouvoir utiliser cette certification.
- Cette certification serait accessible à tous les médias d'information numérique répondant aux critères d'éligibilité.
- La durée de vie de la certification dépendra de son adoption par l'ensemble des parties prenantes. Nous visons toutefois une durée de vie la plus longue possible.