Crowdsourcing et bases de données
Quelques réflexions

François Vignale
Laboratoire 3L.AM, EA 4335, Le Mans Université

Résumé : La base de données EuRED (*European Reading Experience Database*) vise à rassembler des expériences de lecture des Européens de l'invention de l'imprimerie à nos jours. Cette preuve de concept a nécessité la mise au point d'une ontologie, de thesaurus et l'adoption du standard xml-TEI et s'appuie sur la réutilisation de données existantes (35 000 fiches) qui proviennent de la base britannique UK-RED lancée en 2006. Cette dernière a été constituée par des bénévoles formés à l'utilisation et par le grand public lesquels, soit dépouillaient des corpus d'auteurs, soit déposaient des témoignages familiaux, ce qui a généré malheureusement un nombre si important d'erreurs diverses que la possibilité de réutiliser ces données s'en trouvait menacée. Ces obstacles nous ont conduit à encadrer la participation du public dans EuRED et plus encore dans un projet H2020 qui en est la suite et qui prévoit également l'exploration de données massives. Ici, notre approche de la science participative et du *crowdsourcing* repose sur deux axes : 1) la contribution sera toujours possible mais avec une saisie contrôlée pointant vers des référentiels externes et internes ; 2) l'amélioration continue de ces mêmes thesaurus et des algorithmes d'exploration sera rendue possible par les retours d'utilisateurs non-professionnels et de citoyens-experts. Ce sont les constats qui ont guidé cette réflexion et les orientations retenues que nous souhaitons présenter dans cet article.

Mots-clés : crowdsourcing, bases de données, thésaurus, EuRED, READ-IT

Abstract : The aim of the database EURED (European Reading Experience Database) is to gather european reading experiences from the invention of the printing press till now. This proof of concept has required the development of an ontology, of thesaurus and the choice of the xml-TEI standard. It is based on the reuse of existing data (35 000 files) from the british database UK-RED initiated in 2006. This one was completed by skilled and non-skilled volunteers who analysed and processed corpus of authors' works, or uploaded family records, but there were so many errors that it was difficult to reuse the produced data. To avoid these errors, the volunteers for the EURED project and for an H2020 project (with similar massive data mining) will have the support of professionals. Our approach of citizen science will be based on two axes : 1) it will still be possible to contribute but the data input will be controlled by external and internals experts ; 2) the continuous improvement of these thesaurus and of the data mining algorithms will be made possible by the feedbacks of non-professional and of citizen-experts. The aim of this paper is to present the observations that guided the reflection and the guidelines that were adopted.

Keywords : Crowdsourcing, database, thesaurus, EuRED, READ-IT

Cette contribution repose sur une étude de cas qui n'a pas nécessairement vocation à être généralisée. Elle permet cependant de mettre en évidence de manière très pratique des questions qui concernent de nombreux autres projets pour lesquels l'action de recherche met en œuvre des problématiques spécifiques de *crowdsourcing* et plus largement de science participative.

L'une des autres particularités du projet sur lequel s'appuie cet article est qu'il ne porte pas uniquement sur l'annotation ou la correction d'un corpus numérisé par un groupe de volontaires bénévoles comme on peut le voir dans les exemples les plus connus (Transcribe Bentham, par exemple), mais qu'il vise au contraire à construire un jeu de données dont les limites ne sont pas fixées, par le biais de la fouille de corpus existants ou bien par l'apport volontaire de données personnelles grâce à une interface spécifiquement dédiée qui permettra par ailleurs à l'utilisateur de créer son propre parcours de navigation à l'intérieur de la base de données et des collections du corpus. Celui-ci pourra à son tour faire éventuellement l'objet d'une analyse par certains chercheurs participant au projet.

Dans une période dominée par l'irruption de l'intelligence artificielle et des données massives dans les sciences humaines, il s'agit de montrer quels types de problèmes pose l'utilisation du *crowdsourcing* dans une base de données en matière de curation de contenu, tout en ayant bien à l'esprit qu'il est aujourd'hui strictement impossible de s'en passer. Ce sont donc les questions d'une part de la définition de la meilleure articulation entre le *crowdsourcing* et les objets spécifiques que sont les bases de données, et d'autre part des moyens à mettre en œuvre pour assurer la qualité, la cohérence et la durabilité des données produites à partir d'une démarche de science participative qui seront posées ici.

Origine et objectifs initiaux

Le projet READ-IT qui est au cœur de ces réflexions s'appuie sur un projet précédent : EuRED. La base de données EuRED (*European Reading Experience Database*) est une preuve de concept qui a été développée dans le cadre du projet P-RECIHC (*Reading in Europe : Contemporary Issues in Historical and Comparative Perspective*) financé par l'agence nationale de la recherche (ANR) entre 2014 et 2017. L'ambition de cette base de données était de recueillir des témoignages d'expériences de lecture d'Européens entre le XVI^{e} siècle et aujourd'hui sous forme de données textuelles exclusivement, l'expérience de lecture étant caractérisée au sens large comme le récit personnel ou non d'une interaction entre un être humain et un support écrit. Elle avait également pour objectif de tester la possibilité de réutilisation de données existantes, mais également d'offrir la possibilité de réutiliser les données produites par elle-même en se conformant aux principes FAIR (trouvables, accessibles, interopérables, réutilisables). Il fallait également prévoir l'intégration de contenus nativement numériques (tweets, SMS, posts de blogs…). Cette prise en compte a eu pour conséquence l'introduction d'éléments supplémentaires de caractérisation, le tout dans une logique d'acquisition de données massives et de fouille du Web. Au total donc, l'ambition affichée de cette preuve de concept était de tester les possibilités et d'identifier les verrous technologiques et conceptuels potentiels pouvant contrarier le développement

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850

d'un outil plus évolué permettant d'explorer des corpus multimodaux (textes, images, sons).

EuRED s'est appuyée sur l'expérience acquise par le projet britannique UK-RED (*United Kingdom Reading Experience Database*[1]). Cette base de données – toujours active et alimentée – a été lancée en 2006. Elle compte aujourd'hui près de 40 000 enregistrements d'expériences de lecture qui ont eu lieu entre les débuts de l'imprimerie et 1945. Elle s'appuie sur une technologie robuste et éprouvée. Son modèle de données permet de poser les bases de la caractérisation de l'expérience de lecture, mais il est d'une particulière complexité (une quarantaine de tables pour un peu moins de 200 champs). Surtout, il se limite à la sphère culturelle britannique quand il ne concerne tout simplement pas uniquement l'Angleterre[2]. Son mode d'alimentation s'est exclusivement appuyé sur le *crowdsourcing,* mais selon des modalités variées, directes et indirectes.

Une grande partie des données indexées dans UK-RED proviennent de l'exploration systématique et manuelle de corpus souvent non encore numérisés d'auteurs britanniques comme Robert-Louis Stevenson ou Joseph Conrad dont les œuvres et les papiers personnels ont été passés au crible par des équipes de bénévoles, formés, à la recherche de témoignages d'expériences de lecture. Ces derniers ont ensuite entré manuellement ces traces dans la base de données. Ensuite, UK-RED a bénéficié de campagnes de recueil de données primaires, comme les journaux de tranchées de soldats de la Grande Guerre, auxquelles elle a pu avoir un accès privilégié. Cependant, même dans ce cas, la méthode d'ingestion dans la base de données est restée entièrement manuelle et assurée par les mêmes bénévoles encadrés par un chercheur. Enfin, UK-RED est dotée d'une interface de contribution qui permet la saisie en ligne des éléments constitutifs d'une expérience de lecture antérieure à 1945 par tout utilisateur, que ces éléments proviennent de documents personnels ou bien du domaine public.

Afin de dépasser ces limites technologiques, conceptuelles et culturelles, EuRED a commencé par définir un modèle de données plus simple, mais permettant de prendre en compte de nouveaux éléments de caractérisation de l'expérience de lecture en matière de circonstances (environnement, position du lecteur…), tout en intégrant des éléments permettant de prendre en compte les spécificités culturelles de chaque Européen.

1. Cette base de données est accessible à l'adresse suivante : `https://www.open.ac.uk/Arts/reading/UK/`, consulté le 10 juillet 2020.

2. Certains items et les possibilités offertes par les listes de choix prédéfinies sont en effet particulièrement restrictifs pour les religions (il n'y a pas de réponse prévue pour l'athéisme) et en matière linguistique puisqu'il n'est tenu compte ni du gaélique (irlandais ou écossais) ni du gallois par exemple ou de toute autre langue parlée ou écrite dans l'Empire colonial britannique.

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850

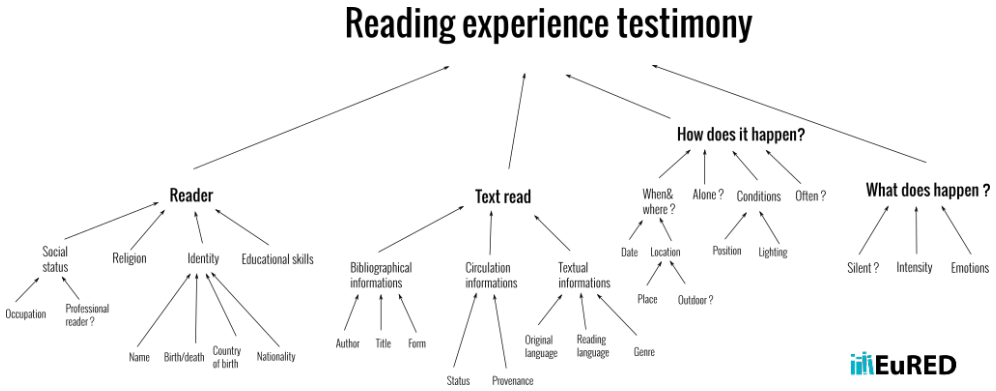


Figure 1 – Formalisation du modèle de données de EuRED

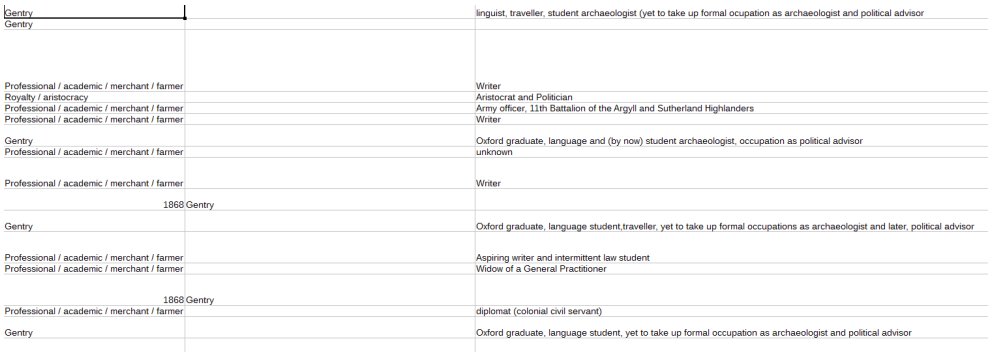
Ce modèle de données a ensuite été exprimé en xml-TEI avec le développement d'un jeu de balises spécifiques dans le but de s'appuyer sur des standards éprouvés et de garantir le meilleur niveau d'interopérabilité. En outre, il a dès le départ été prévu de construire un jeu de thésaurus et de vocabulaires contrôlés dédiés en vue de procéder à un maximum d'alignements automatiques lors de l'ingestion de données externes. Enfin, pour un certain nombre de champs (personnes, titres, lieux. . .), un système de liage a été déployé vers des référentiels externes (VIAF, data.bnf.fr, geonames . . .) en utilisant les technologies du Web sémantique.

Afin de tester tous ces développements, nous avons pris comme base de travail une extraction de UK-RED représentant environ 34 000 enregistrements créés entre 2006 et 2011. Après une série de tests effectués sur un échantillon de 200 enregistrements, la confrontation du jeu de données provenant de UK-RED et du modèle de données de EuRED a été conduite tout au long du deuxième trimestre 2016. Des tables de correspondance entre les champs ont été construites par itérations successives afin de faciliter les alignements. Un premier bilan a pu faire apparaître que, si les correspondances ont été globalement très satisfaisantes, il n'en est pas de même avec les données sources dont la qualité pose souvent un problème majeur susceptible de nuire à leur réutilisation ultérieure et qui n'avait pu être identifié lors de l'exploitation de l'échantillon-test.

Les difficultés rencontrées ont été multiples et de nature différente. Comme on pouvait s'y attendre, les ambiguïtés sont nombreuses en ce qui concerne les noms de personnes, les titres des œuvres, les localisations, ce qui ne surprend pas parce qu'aucun système de lien vers des référentiels externes n'a été mis en place dans UK-RED. Plus fâcheux en revanche, l'analyse plus fine des enregistrements convertis a fait apparaître des erreurs extrêmement fréquentes, non seulement dans la structure de la base de données elle-même, mais surtout dans la nature même des données renseignées. On trouve ainsi des champs renseignés par erreur, des fautes de frappe, et surtout, de manière globale, un certain manque de compréhension dans la nature des champs et des informations à indexer comme on le verra dans les exemples qui suivent. Au total, le taux d'erreur calculé sur la base de la totalité des 34 000 enregistrements est très

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850

légèrement inférieur à 16 % (soit un peu plus de 5 400 enregistrements concernés). Ce chiffre est considérable et amène à remettre en cause la fiabilité globale des données produites.



Gentry		linguist, traveller, student archaeologist (yet to take up formal ocupation as archaeologist and political advisor
Gentry		
Professional / academic / merchant / farmer		Writer
Royalty / aristocracy		Aristocrat and Politician
Professional / academic / merchant / farmer		Army officer, 11th Battalion of the Argyll and Sutherland Highlanders
Professional / academic / merchant / farmer		Writer
Gentry		Oxford graduate, language and (by now) student archaeologist, occupation as political advisor
Professional / academic / merchant / farmer		unknown
Professional / academic / merchant / farmer		Writer
1868	Gentry	
Gentry		Oxford graduate, language student,traveller, yet to take up formal occupations as archaeologist and later, political advisor
Professional / academic / merchant / farmer		Aspiring writer and intermittent law student
Professional / academic / merchant / farmer		Widow of a General Practitioner
1868	Gentry	
Professional / academic / merchant / farmer		diplomat (colonial civil servant)
Gentry		Oxford graduate, language student, yet to take up formal occupation as archaeologist and political advisor

Figure 2 – Confusions et erreurs de saisie dans les champs

L'exemple ci-dessus montre que les champs, dans bien des cas, n'ont pas été renseignés correctement. Ici, il y a une confusion dans la nature des champs. La colonne de gauche doit faire apparaître à quel groupe socio-économique appartient la personne qui relate son expérience de lecture. On y trouve toujours la même date (1868), alors que la mention d'appartenance à la *gentry* se trouve rejetée dans un autre champ. Le caractère répété de cette erreur peut avoir deux explications : l'une, liée à la technique et pourrait être en fait un dysfonctionnement temporaire de l'interface de saisie ; l'autre, en raison de son caractère répété qui porte sur les mêmes valeurs, tiendrait à des facteurs humains qui peuvent faire penser soit à un problème d'attention lors de la saisie, soit à un problème dans la formation du bénévole et/ou dans la documentation de l'interface.

L'exemple suivant porte sur le champ « Occupation [métier] ». Il pose des questions différentes mais tout aussi cruciales pour assurer la meilleure qualité possible des données produites. Ce champ a pour vocation de recueillir des informations sur le métier occupé par la personne qui relate son expérience de lecture. Plusieurs constats s'imposent. Le premier est qu'il existe une confusion manifeste entre le fait de détenir une position familiale et donc d'avoir des liens de parenté (*daughter*, « fille ») et le fait d'avoir une activité professionnelle. Ceci peut éventuellement parfois se concevoir, mais il s'agit toujours de situations très particulières et très exceptionnelles (familles régnantes par exemple). Le deuxième est qu'on peut trouver à la suite de la mention de ces liens familiaux un autre degré de confusion puisque s'y trouve ajoutée l'indication d'une profession (« fille d'un historien de la musique, puis écrivaine »). Cette absence de normalisation et de standardisation, malgré tout l'intérêt de ces informations, conduit à abandonner toute possibilité de traitement statistique. Le troisième enfin tient au fait que la base contient en outre des fautes de frappe (« *landowning* »/« *land owning* »), ce qui est bien entendu un risque inhérent à la saisie manuelle des données. L'inconvénient ici est que ces trois types d'erreurs ont des effets

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850

cumulatifs qui rendent très problématique la confiance dans les données produites qui deviennent naturellement plus difficiles à exploiter par la suite.

daughter of Lieutenant
daughter of manufacturer; later schoolmistress and head of English
daughter of merchant and banker
Daughter of monarch
daughter of MP
daughter of music historian Charles Burney; later writer
daughter of novelist and journalist
daughter of Prime Minister
daughter of Prison Governor
daughter of railway clerk
daughter of Revd Peter Leigh
daughter of Richardson's late friend
Daughter of Scottish land owning family
Daughter of Scottish landowning family
Daughter of stonemason
Daughter of the fifth Earl of Kingston upon Hull

Figure 3 – Extraction du champ « Occupation »

Ces exemples que l'on aurait pu multiplier montrent les difficultés et les limites auxquelles sont confrontées les expériences de *crowdsourcing* quand elles touchent à la saisie répétitive et manuelle de données et moins à la transcription collective d'un corpus. Il existe cependant des moyens de limiter considérablement ce type de risques sans renoncer pour autant à une stratégie de participation du grand public.

Pour la mise au point du prototype EuRED, nous avons très vite fait le choix de mettre en place une autre procédure de saisie de données pour contourner ce genre de problème afin de garantir la meilleure qualité en entrée et en sortie. Il repose sur trois piliers : mise en place de façon quasi systématique de listes de choix fermées dans l'interface de saisie ; mise en place de connecteurs vers des référentiels externes ; et liens vers des thésaurus structurés et des vocabulaires contrôlés internes (au nombre de 23 [3]). Ces orientations ne visent toutefois pas à automatiser la saisie, mais n'ont d'autre objectif que de limiter le plus possible le nombre d'erreurs humaines. Dans notre prototype, les titres des œuvres citées dans les témoignages d'expériences de lecture sont extraits de référentiels après une auto-complétion. Le système de gestion des thésaurus permet quant à lui une structuration fine des notions et concepts en même temps qu'il génère automatiquement une liste de choix dans laquelle l'utilisateur doit en principe trouver le terme qui convient à l'élément constitutif de l'expérience qu'il doit décrire. En parallèle, les indices liés à chaque terme et la publication en ligne permettent de leur côté de garantir la stabilité de l'ensemble dans le temps.

3. Sur la distinction entre « thésaurus » et « vocabulaire contrôlé », voir Gilchrist, Alan (2003) « Thesauri, taxonomies and ontologies – an etymological note », *Journal of Documentation*, vol. 59, Issue 1, p. 7-18. Accessible sur `https://www.emerald.com/insight/content/doi/10.1108/00220410310457984/full/html`, consulté le 20 juillet 2020. Ces thésaurus et vocabulaires contrôlés sont accessibles à l'adresse suivante : `http://eured.univ-lemans.fr/thesaurus/`).

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850



FIGURE 4 – Vue partielle du thésaurus « Occupations » de EuRED. `http://eured.univ-lemans.fr/thesaurus/occupation`

Au total, le retour d'expérience lors de la réutilisation des données provenant de UK-RED fait clairement apparaître que la résolution de certains problèmes liés à la participation de bénévoles (qu'ils en soient à l'origine ou non) dans l'alimentation d'une base de données de ce type doit être incluse dans la réflexion de départ. Au cœur de celle-ci doit se trouver la nécessité de mettre en œuvre une interface dont l'ergonomie doit être pensée pour des utilisateurs non professionnels. Ceci implique notamment d'accéder en temps réel à la documentation du modèle de données ; la réduction la plus drastique possible des champs de saisie de texte, ce qui signifie que le recours aux listes de choix liées à des thésaurus doit être le plus systématique possible ; la récupération automatique des informations normalisées contenues dans des référentiels bibliographiques externes pour les identités, les titres et les localisations en s'appuyant sur des processus d'auto-complétion. Toutefois, il conviendra de prendre garde à ce que, partant d'un *crowdsourcing* de contenu valorisant, on assiste à un appauvrissement des tâches et qu'on en arrive à une forme de *crowdsourcing* d'activités routinières [4].

Surtout, il est absolument essentiel qu'intervienne une procédure de modération – sous une forme à définir – qui permettra la validation et par conséquent la publication ou non des données entrées par l'ensemble des contributeurs, cette modération pouvant alors elle-même être confiée à des groupes de bénévoles spécialement formés et encadrés. C'est à cette condition uniquement que la durabilité de ce type de plate-forme – qui repose sur la potentialité d'une réutilisation future des données produites – peut exister et donc le projet perdurer au-delà de l'extinction des financements qui ont permis sa réalisation.

4. En suivant les catégories définies par Burger-Helmchen, Thierry, et Pénin, Julien (2011) : « *Crowdsourcing* : définition, enjeux, typologie », *Management & Avenir*, 2011/1, n° 41, p. 254-269. Accessible sur `https://www.cairn.info/revue-management-et-avenir-2011-1-page-254.htm`, consulté le 20 juillet 2020.

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850

READ-IT : *crowdsourcing*, curation et durabilité

La preuve de concept EuRED et l'expérience acquise sont aujourd'hui fortement mises à contribution dans la définition de la future base de données et des interfaces qui seront produites par le projet READ-IT (*Reading Europe Advanced Data Investigation Tool*) retenu dans le cadre du programme européen *Joint Programming Initiative for Cultural Heritage*. Ce projet de recherche regroupe cinq partenaires (Le Mans Université, France ; CNRS-IRISA, France : Universiteit Utrecht, Pays-Bas ; Open University London, Royaume-Uni et l'Institut de Littérature tchèque de l'Académie des Sciences de la République Tchèque). Il est financé pour la période 2018-2022. Ce projet a pour but de permettre l'identification et le partage de témoignages d'expériences de lecture d'Européens entre le XVIII^{e} siècle et aujourd'hui. À la différence de UK-RED et de EuRED, les données proviendront de sources multimodales (textes, images fixes, enregistrements sonores, contenus nativement numériques) et seront multilingues (français, anglais, allemand, néerlandais et tchèque). Une grande partie des données proviendront de réservoirs de données existants (bibliothèques numériques par exemple) et de l'exploration du Web. Cependant, le *crowdsourcing* n'en est pas pour autant exclu, l'un des objectifs majeurs de READ-IT étant bien de créer les conditions d'une participation active et efficace du grand public à un processus de mise en valeur du patrimoine et d'enrichissement de la connaissance. Le grand public pourra participer aux activités du projet et alimenter la base selon plusieurs modalités. La première est liée à la validation du modèle de données et de l'ontologie associée par la mise en place de campagnes d'annotation ; la deuxième, plus indirecte, repose sur la validation et l'annotation des documents candidats avant leur intégration dans la base de données ; enfin, la troisième consiste en la mise en place d'une interface spécifique de dépôt et de navigation permettant de recueillir des souvenirs d'expériences de lecture par des contributions volontaires.

En raison de sa très large automatisation, l'exploration de corpus numérisés en vue de l'identification des expériences de lecture, puis leur indexation nécessitent la mise en œuvre d'un modèle de données souple et robuste d'autant plus que le projet s'appuie sur un ensemble de technologies liées au traitement du langage naturel (NLP) et à l'apprentissage profond, lesquels exigent des phases d'itération très nombreuses qui doivent permettre aux algorithmes d'améliorer à chaque fois leurs performances. Ici, la participation du grand public interviendra à la suite d'une première phase d'annotation et de validation de la part d'experts du domaine de l'histoire de la lecture, puis d'une deuxième qui s'est déroulée au printemps 2020. Cette campagne poursuivait plusieurs objectifs : achever la validation du modèle et de l'ontologie ; produire plusieurs milliers d'annotations pour améliorer la performance des algorithmes de détection, améliorer l'ergonomie de l'interface dédiée [5] et concevoir un guide d'annotation suffisamment simple et précis pour être mis à disposition du grand public. Elle a nécessité la participation pendant 10 semaines en mars, avril et mai 2020 de 9 étudiants de Master 2 Études culturelles internationales qui, après avoir été formés, ont dû annoter chacun environ 1 000 commentaires de lecture provenant des plateformes Babelio (français) et Goodreads (anglais) en se conformant aux instructions données,

5. Cette interface est accessible à l'adresse `https://read-it.hum.uu.nl/explore`

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850

à savoir l'identification des composantes de l'expérience de lecture [6]. Des réunions hebdomadaires tenues en visioconférence ont permis de faire remonter les difficultés rencontrées tant sur le plan conceptuel, méthodologique que technique et d'améliorer significativement les outils disponibles et les performances des algorithmes de détection. Ces réunions plénières ont été complétées par des réunions par groupes de trois qui ont permis d'affiner les résultats. Une fois les enseignements de cette campagne tirés, nous pourrons ouvrir ce type d'opérations au grand public. Il est ainsi prévu que des séances de *crowdsourcing* se déroulent dans le cadre d'événements de dissémination un peu partout en Europe où le public pourra se prononcer sur la qualité de l'annotation d'images candidates et éventuellement les corriger. Ces séances – qui pourraient s'apparenter à un système de *focus group* ou à un *hackathon* en raison du caractère non professionnel des participants – seront dirigées et encadrées à chaque fois par un membre du projet. En dehors de ce travail de validation, l'intérêt de ce type d'action est également de recueillir de nouveaux éléments du vocabulaire associé à la description de ces situations de lecture et ainsi d'enrichir les instruments terminologiques proposés par le projet en y ajoutant des termes qui auraient pu échapper aux experts. En d'autres termes, il s'agit ici de mettre en œuvre un *crowdsourcing* contrôlé qui ne nécessitera pas ultérieurement – en théorie – une opération de vérification des données et de la cohérence de la base.

Les techniques de *Webcrawling* qui seront mises en œuvre dans READ-IT feront indirectement intervenir le grand public puisqu'il s'agit de chercher des traces de lecture contemporaines qui seraient contenues dans des médias numériques et autres réseaux sociaux. Bien entendu, ces opérations seront menées en totale conformité avec les obligations juridiques en matière de protection des données personnelles. Ici, il s'agit de recueillir un matériau original en perpétuelle mutation produit par des lecteurs le plus souvent non professionnels à la recherche de nouvelles modalités de l'expérience de lecture, afin à la fois d'enrichir une fois encore les outils terminologiques et le modèle de données, et d'augmenter les connaissances dans le champ de l'étude de la lecture à l'ère numérique, à l'échelle européenne.

La plate-forme créée dans le cadre de READ-IT est également dotée d'une interface spécifique [7] qui permettra à court terme le recueil de souvenirs de lecture auprès du grand public, et la navigation dans des collections numériques sur un double modèle « *my stories* »/« *my collections* ». L'approche est double et elle repose sur un principe de co-curation qui inclut les déposants et le projet lui-même. De manière schématique, il s'agit donc, très classiquement, de faire appel au grand public afin de recueillir des données et des documents personnels relatifs à des expériences de lecture qui seraient alors déposés sur une application spécifique. Dans le même temps, nous souhaitons également offrir à tout un chacun intéressé par le sujet la possibilité de créer ainsi des parcours de navigation et des collections personnelles (*scrapbooks*) à travers ses propres documents, mais également à travers d'autres sélectionnés au sein de la base de données. Ces contenus pourraient alors être mis en ligne et partagés avec l'accord de leurs créateurs. L'intérêt de ce type de dispositif est multiple. Pour les membres

6. C'est-à-dire, l'identification du sujet lisant, des verbes d'action, du contenu, de la forme et des circonstances dans laquelle s'est déroulée cette expérience.

7. Cette interface est accessible à l'adresse `https://read-it.in-two.com/`

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850

du projet, il y a bien entendu en premier lieu la possibilité de recueillir des documents de première main et le plus souvent uniques. Ensuite, par l'analyse de la nature et du contenu des annotations et des commentaires portés par les utilisateurs, c'est une autre façon efficace d'enrichir le vocabulaire disponible. En outre, en suivant et en évaluant les utilisateurs de ces interfaces [8] dans leur comportement et la façon dont ils construisent leurs collections et leurs commentaires, les chercheurs qui participent au projet auront pour certains l'opportunité de travailler en temps réel l'ergonomie de l'interface, quand d'autres pourront mesurer les apports des formes de sérendipité qui ne manqueront pas de s'exprimer dans l'exploration des données et la façon dont elles seront utilisées. D'autres encore auront en outre la possibilité d'utiliser ces récits pour apporter des éléments de connaissance irremplaçables à l'étude des pratiques de lecture passées et présentes en Europe.

Cependant, ce type d'approche qui encourage la participation du public n'est pas sans poser des problèmes spécifiques en matière de protection des données personnelles, mais aussi en matière de protection de la propriété intellectuelle pour les œuvres encore sous droits qui seraient mentionnées dans ces expériences de lecture et dont des extraits seraient cités, ou pour des documents graphiques non encore tombés dans le domaine public qui ne bénéficieraient pas d'une licence de type *Creative Commons*. En ce qui concerne la protection des données à caractère personnel, chaque création de compte dans l'interface – que ce soit pour y déposer et/ou pour construire une collection – fera l'objet d'un consentement éclairé explicite de la part du participant en même temps que sera mise en place une procédure d'anonymisation. De plus, comme ce projet prévoit la mise en place de deux interfaces de consultation [9], ce consentement inclura une procédure d'*opt-in* qui autorisera ou non la visualisation de ce témoignage dans l'interface publique. Pour les documents susceptibles de contenir des éléments sous droits, une première phase d'examen des métadonnées associées sera effectuée de manière automatique, puis, en tout état de cause, une validation humaine sera conduite dans le cadre d'un processus de modération.

Cette phase de modération est au cœur de la réflexion sur les conditions de durabilité (*sustainability*) de ce projet afin qu'il continue à vivre au-delà de l'extinction de son financement. Cette préoccupation permanente pour tout programme de recherche devient encore plus prégnante dans le cadre d'une approche reposant sur une approche de science participative en général et de *crowdsourcing* en particulier. Ici, la démarche retenue ne réside pas dans une stratégie de transfert de compétences à des groupes de bénévoles volontaires formés, mais elle s'inscrit plutôt dans un cadre de développement de la formation en matière de gestion des données à destination des étudiants de sciences humaines et sociales – mais pas uniquement – qui repose sur deux types d'actions. D'une part, pendant la durée du projet, l'organisation de plusieurs *summer schools* mêlant des étudiants de niveau M en informatique et en SHS est prévue afin qu'ils puissent réfléchir conjointement aux enjeux spécifiques de la gestion de ce type de données. D'autre part et sur un plus long terme, l'objectif est de s'appuyer sur le renforcement, voire la mise en place de formations à destination des étudiants de

8. Ces interfaces appartiennent à la catégorie des PEA (*Public Engagement Applications*). Dans READ-IT, il n'est en revanche pas question de mettre en place des jeux sérieux.

9. L'une pour le grand public, et l'autre, avec accès authentifié, pour des chercheurs.

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850

l'université du Mans dans le domaine des humanités numériques de niveau L et M pour lesquels READ-IT constituerait un terrain idéal d'acculturation puis d'entraînement aux problèmes spécifiques posés par la curation de données hétérogènes et massives en sciences humaines et sociales. Ce faisant, certaines des conditions d'une modération efficace et durable paraissent en mesure d'être réunies.

Published in : Cécile Meynard et Thomas Lebarbé (dir.), *Le Crowdsourcing : Partager, enrichir et publier des sources patrimoniales*, Editions des archives contemporaines, ISBN : 9782813003850, doi : 10.17184/eac.9782813003850