Plan Overview

A Data Management Plan created using DMPTool

Title: Evaluation de la couverture de l'archivage du web suisse

Creator: Christelle Donius

Affiliation: Non Partner Institution

Principal Investigator: Christelle Donius

Data Manager: Christelle Donius

Funder: Digital Curation Centre (dcc.ac.uk)

Funding opportunity number: 41763

Template: Digital Curation Centre

Project abstract:

Ce travail de recherche vise à analyser et comparer les approches de deux acteurs de l'archivage de contenus web suisses : 1. La bibliothèque nationale (BN), qui archive un choix de sites, en fonction de son mandat de collecte des (e-)Helvetica, selon une sélection qualitative ; 2. L'Internet Archive (IA), une initiative américaine à but non lucratif, qui archive toutes sortes de contenus numériques, dont des pages web, en pratiquant le "harvest all" à l'aide de crawlers. À partir d'une observation des pratiques dans d'autres pays, nous examinerons plus particulièrement ces deux méthodes. Puis, nous étudierons le degré de couverture qui en résulte, d'une part en analysant la couverture commune ou différentielle entre des ressources archivées par la BN et par l'IA respectivement, d'autre part en étudiant l'exemple de quelques sites suisses. Nous nous intéresserons également aux questions d'accès et de diffusion des contenus archivés, notamment sous des aspects légaux.

Last modified: 06-04-2019

Evaluation de la couverture de l'archivage du web suisse

Dans le cadre de notre projet de recherche, nous nous emploierons à collecter les métadonnées relatives à l'opération d'archivage de sites web par deux institutions: la Bibliothèque nationale suisse et l'Internet Archive. Il s'agira aussi de collecter les métadonnées associées aux sites web concernés, permettant leur description et leur indexation dans les systèmes des deux institutions prémentionnées afin d'assurer leur découvrabilité et leur consultation lorsque celle-ci est possible.

Les informations collectées sont de nature numérique ou textuelle ; elles sont structurées (CSV ou JSON). De fait, pour faciliter leur traitement, mais également pour des raisons de poids de fichier, ou encore pour permettre une consultation pérenne du contenu des fichiers, ceux-ci seront enregistrés en format CSV ou TXT.

La collecte a pour but d'étudier la couverture de l'archivage du web. Aussi, les données représenteront un instantané du web suisse archivé à l'été 2019. Il n'existe pas de jeu de données antérieur que nous pourrions réutiliser. En revanche, les données pourront tout à fait être utilisées ultérieurement dans le cadre d'autres recherches.

Quant au volume des données, celles-ci étant purement textuelles, nous l'estimons inférieur à 50 GB. Nous ajusterons cette valeur au moment de la collecte effective.

Outre le traitement des données et la mise à disposition des résultats issus de leur analyse dans le cadre du projet de recherche, nous restons ouvertes à une publication ultérieure à laquelle les données seront également attachées.

Plusieurs méthodes de collecte des métadonnées sont envisagées par ordre de préférence et selon le succès de chaque opération:

1. interrogation d'API et de bases de données publiques,
2. requête directe auprès des institutions détentrices en expliquant la démarche de recherche,
3. éventuellement collecte automatisée directement sur les pages web des institutions, si le besoin s'en faisait ressentir.

Les méthodes d'interrogation et de collecte automatisée feront l'objet de rédaction de ligne de codes nécessaires à leur exécution qui seront au préalable testées sur de petits volumes afin de garantir leur efficacité, avant d'être appliquées à la récolte réelle. Les erreurs qui pourront découler de ces procédures seront également documentées pour garantir une transparence quant à l'acquisition et l'intégrité ou non des données.

La convention de nommage des jeux de données sera le suivant: *[nom de l'institution]_[date de récolte/traitement]_[objet]_[version].[txt/csv]*.

Chaque jeu de données sera accompagné d'un fichier info.txt qui décrira la méthode de récolte appliquée. Un fichier readme.txt sera également associé à l'ensemble des données afin de documenter la convention de nommage et la hiérarchie des répertoires.

Les données collectées suivent un schéma que nous décrirons, en précisant quelles sont les données qui nous intéressent réellement et qui seront traitées.

Les traitements appliqués aux données seront également documentés, par le biais des programmes utilisés qui le pratiquent, mais surtout par la complétion du carnet de bord partagé entre les membres de l'équipe. Toutes les opérations effectuées sur les données seront consignées de manière précise afin d'être décrites et retranscrites ultérieurement.

Seules des métadonnées d'archives de sites web, disponibles auprès des des institutions sondées, seront traitées. Aucune donnée confidentielle n'est concernée.

Bien qu'il ne soit pas question de traitement et de diffusion de contenu, nous veillerons à appliquer les mêmes règles que celles instaurées par la Bibliothèque nationale suisse sur ces aspects, par respect du droit d'auteur.

Toute source utilisée sera citée dans le rapport de recherche, conformément aux pratiques en vigueur.

Conformément à l'article 17 du règlement sur la formation de base (bachelor et master) de la HES-SO , "à l'exception des droits d'auteur, les droits sur les biens immatériels réalisés par les étudiant-e-s dans le cadre de leur formation ou d'un mandat de recherche confié par ou à l'école sont la propriété de l'école."

Cependant, au vu de la démarche de l'école de tendre vers l'Open Access, et relativement aux cours qui nous sont dispensés dans ce sens, les données récoltées dans le cadre du projet de recherche seront diffusées sous licence Creative Commons CC-BY.

Dans un souci d'accessibilité aux données par tous les membres de l'équipe du projet, mais également pour s'assurer de la préservation des données, nous utiliserons le cloud sécurisé mis à disposition par la HES-SO, à savoir Switch Drive. Le service propose un versionning des fichiers et conserve les fichiers supprimés pendant 3 mois, ce qui permet d'éviter la suppression intempestive de données. Le volume de données autorisé est de 50 GB, ce qui est tout à fait suffisant pour le stockage des données collectées et de nos documents de travail. Si le volume des données venait à avoir été sous-estimé, un serveur de le HES-GE dédié à la recherche pourra être utilisé pour le dépôt et le traitement des données.

Le service Switch Drive ne proposant pas de backup, les dossiers de travail seront régulièrement et automatiquement sauvegardées sur le NAS personnel de l'un des membres de l'équipe, avec synchronisation sur deux autres machines équivalentes de son réseau. Un backup manuel sera exécuté par l'autre membre de l'équipe à chaque étape charnière du projet pour les différents livrables ou "matières premières" (documents rédigés, données collectées).

Dans la mesure où aucune donnée sensible ne sera collectée, nous n'appliquerons pas de mesures de sécurité poussées. L'équipe de recherche n'étant composée que de deux membres, une confiance mutuelle est de rigueur, tout comme une transparence quant à l'utilisation des informations partagées sur les dossiers communs. Chaque membre possède ses propres identifiants pour accéder aux données. La gestion des identifiants est gérée sur et par Switch Drive et est liée aux identifiants fournis par l'établissement d'études. Les membres de l'équipe sont les seules habilitées à donner accès au contenu à un tiers en possession d'un compte Switch Drive.

Aucune obligation légale ou contractuelle ne gouverne la conservation des données collectées dans le cadre de ce projet.

Nous prévoyons de conserver deux jeux de données : les données brutes collectées et les données traitées. Dans la mesure du possible, nous utiliserons le format d'enregistrement ouvert CSV avec un encodage en UTF-8 afin de garantir leur accès sur le long terme.

Nous envisageons également de mettre à disposition les fichiers de traitement des données, en renseignant le programme utilisé, ainsi que son numéro de version. En effet, un tel fichier peut contenir un historique du traitement appliqué aux données facilitant une documentation et une reproductibilité de nos résultats.

Ces pratiques, nous l'espérons, permettront l'utilisation de nos données pour d'autres recherches similaires.

La préparation des jeux de données, leur documentation et leur dépôt dans un repository font partie intégrante du projet de recherche.

A l'heure actuelle, nous envisageons Zenodo comme dépôt pour nos données. En effet, ce dépôt se montre satisfaisant sur plusieurs points : types et formats de données autorisés, absence de coûts, possibilité d'attribuer une licence Creative Commons à nos données, et l'utilisation d'identifiants pérennes tant pour les données que la liaison avec ORCID.

En revanche, le volume de dépôt étant limité à 50 GB, en fonction du volume total final de nos données, nous nous autorisons le choix d'un autre dépôt ultérieurement.

Les données seront mises à disposition et documentées dans un dépôt en ligne, dès la fin du projet de recherche (janvier 2020). Elles seront partagées avec n'importe qui sous licence CC-BY. Le dépôt choisi permet l'attribution d'un identifiant pérenne via le dépôt. Les membres de l'équipe de recherche seront identifiées par leur ORCID, ce qui garantit qu'elles pourront être contactées en cas de besoin même lorsque leur adresse e-mail actuelle auprès de la HEG ne sera plus valide.

En cas de publication ultérieure, si la revue ne permet pas de joindre les données à l'article, le recours au DOI fourni au moment du dépôt permettra leur consultation indépendamment de l'éditeur.

Aucune restriction quant à la réutilisation des données ne sera appliquée.

Les données seront mise à disposition sous licence CC-BY.

Les deux membres du groupe de projet sont co-responsables pour l'ensemble des étapes du data management, que ce soit la collecte, la documentation, l'organisation des fichiers dans les dossiers, l'archivage, le partage des données, et toute autre opération nécessaire.

Nous nous baserons sur les enseignements reçus dans le cadre des cours de la HEG et sur les informations disponibles sur le site web du dépôt sélectionné. Il n'y aura pas besoin de ressources informatiques supplémentaires, ni de financement requis.
