RES
PA
DON

Colloque International
3, 4 et 5 avril 2023

(English version -> p. 8)

LE WEB : SOURCE ET ARCHIVE

Présentation du projet

Ce colloque international propose d'interroger la place des sources issues du web dans la recherche et de situer les pratiques d'archivage du web dans des démarches et des questionnements scientifiques pluriels.

Soutenu par le GIS CollEx-Persée et porté par le Service Commun de Documentation de l'Université de Lille et la Bibliothèque nationale de France, en partenariat avec le GERiiCO de l'Université de Lille, Sciences Po et le Campus Condorcet, le réseau ResPaDon se donne pour objectif d'ouvrir un espace de dialogue entre chercheurs et professionnels des bibliothèques et des archives. Dans cette perspective, le colloque international organisé par l'équipe « Usages » pilotée par Laurence Favier (Pr. Université de Lille, GERiiCO), Emmanuelle Bermès (Adjointe chargée des questions scientifiques et techniques auprès de la Direction des services et des réseaux, BnF) et Marie-Madeleine Géroudet (Responsable du Département Services à la recherche et aux chercheurs, SCD de l'Université de Lille) souhaite développer les échanges de savoirs et de pratiques entre les professionnels de l'IST et les équipes de recherche en engageant la communauté académique internationale à participer à ses travaux.

Organisé à LILLIAD (Learning Center Innovation de l'Université de Lille) sur le Campus de la Faculté des Sciences et Technologies à Villeneuve D'Ascq, l'événement se déroulera sur 2 jours et demi, les 3, 4 et 5 avril 2023. Les travaux de ces journées seront publiés.

Université de Lille · BnF Bibliothèque nationale de France · CollEx-Persée · SciencesPo · Campus Condorcet Paris-Aubervilliers · GériCo

Introduction générale – positionnement scientifique

C'est à condition de faire l'objet de pratiques d'archivage que le web devient corpus de recherche (Beaudouin, Pehlivan, 2016 : 15). Support d'inscription éphémère de signes, codes et données qui circulent à travers les réseaux numériques, le web comme matériau scientifique interroge les chercheurs en tant que source ou terrain de recherche et questionne les professionnels de l'information et de la documentation en tant qu'archive.

De 2013 à 2016, le projet « Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre » a le premier utilisé une collection des archives de l'internet pour lui appliquer des méthodes d'analyse globales. Un corpus de sites avait alors été spécifiquement délimité pour procéder à la collecte qui allait permettre d'analyser les pratiques amateurs autour des contenus patrimoniaux numérisés et mis à disposition par les bibliothèques. Au printemps 2022, le Datasprint organisé dans le cadre du projet Respadon a confirmé que la mobilisation des archives du web pouvait être complémentaire à l'étude du web vivant. L'exploration du web comme source ou comme terrain s'avère désormais indispensable à de nombreuses disciplines : histoire contemporaine, sciences politiques, sociologie, histoire des sciences, sciences de l'information et de la communication. Et parce qu'il permet de faire l'histoire d'un domaine académique et d'inscrire ses travaux dans une tradition, le web est devenu une source nécessaire à de nombreuses démarches scientifiques.

Or, les infrastructures scripturales (Denis, 2018) qui permettent la production de données et leur circulation à travers le web sont paradoxalement fragiles et les contenus volatiles. C'est la raison pour laquelle les institutions culturelles et patrimoniales comme la Bibliothèque Nationale de France (BnF) et l'Institut National de l'Audiovisuel (INA) se sont très tôt intéressées à l'archivage de l'Internet. Dès 1996, Internet Archive commence à constituer une archive internationale du web. En France c'est la loi dite « DADVSI » de 2006 (Droits d'Auteurs et Droits Voisins dans la Société de l'Information) qui étend le dépôt légal aux « signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique ». Elle contribue à définir l'espace du web français comme un patrimoine national représentatif de la production éditoriale de son temps (Bermès, 2019 ; Ilien et al., 2011).

De nombreux autres pays se sont dotés d'un cadre légal similaire et sont réunis dans le cadre du consortium IIPC (International Internet Preservation Consortium). En parallèle les chercheurs qui ont commencé à étudier les archives web comme objet se sont structurés en réseaux avec WARCnet et RESAW (Schafer, Musiani, Borelli, 2016 ; Brügger, 2018).

Inscrit dans la lignée des objectifs du projet Respadon (Réseau de partenaires pour l'exploration et l'analyse de données numériques) qui consiste à rendre les archives web plus accessibles pour les chercheurs de nombreuses disciplines, ce colloque propose de questionner les pratiques scientifiques où le web peut être saisi comme source et comme archive.
Cet événement propose de réunir à la fois des exemples concrets des usages scientifiques des archives web en lien avec leur rôle de source et de terrain, et des réflexions heuristiques sur la nature de ces archives et leurs modalités de préservation et de mise à disposition par les institutions patrimoniales et les établissements documentaires.

Les propositions de communication attendues porteront sur les thématiques suivantes reliées aux 3 axes principaux du colloque, décrits ci-dessous :

- Présentations de projets scientifiques mobilisant les corpus web et les humanités numériques
- Présentations de retours d'expérience de projets de recherche utilisant le web comme source : obstacles rencontrés, "success stories"
- Réflexions autour des pratiques professionnelles et académiques impliquant la collecte et la préservation de données issues du web
- Présentation d'expérimentations et de dispositifs favorisant l'accès aux archives numériques et aux corpus web
- Réflexions méthodologiques et épistémologiques sur les besoins d'accès à des données en ligne dans différentes disciplines et de préservation de ces données.

Axe 1 Le web à l'intersection de la mémoire et du savoir : enjeux épistémologiques

Le web comme terrain scientifique implique de considérer le matériau disponible en ligne comme gisement de traces de la mémoire collective. Comment élaborer un discours scientifique à partir d'une source par essence volatile ? De quelle mémoire s'agit-il lorsqu'elle est « distribuée » sur le web ? Quel est le statut des sources web dans les pratiques scientifiques qui élaborent un nouveau savoir et contribuent à l'émergence de sciences désormais assistées par le numérique ? Il s'agit ici de se questionner collectivement sur les méthodes, les modalités et le statut des résultats produits lorsque des données en ligne entrent en jeu dans la recherche, quelle que soit la discipline concernée (histoire contemporaine, sciences politiques, sociologie, histoire des sciences, sciences de l'information et de la

communication, informatique…). Seront appréciées les approches et propositions qui illustrent des travaux scientifiques utilisant le Web et ses traces ainsi que des réflexions sur la nature des archives et des données utilisées dans la recherche scientifique contemporaine.

Axe 2	Politiques, pratiques et techniques archivistiques et archives web : du document aux corpus

La communauté professionnelle a développé depuis plusieurs années des définitions techniques de l'archive numérique et des dispositifs associés. Cette pratique questionne la notion de « lieux des archives » et de préservation des corpus : l'offre de service autour des archives web doit permettre de faire émerger le terrain de travail avec les chercheurs, en prenant en compte des questionnement divers :

- En matière de souveraineté informationnelle, comment la diversité des initiatives privées et publiques existantes se concrétise-t-elle dans les politiques de constitution des archives ?
- En matière d'accès aux archives, quelle définition de la granularité, de l'unité de sens, de la cohérence du fonds archivistique, pour que les archives web soient exploitables par la recherche ?
- Comment développer des pratiques scientifiques efficientes dans le respect du cadre juridique et des contraintes liées à la notion de dépôt légal ? Comment faire évoluer ce cadre pour favoriser la recherche ?

Pour éclairer ces enjeux, on pourra imaginer de mobiliser l'histoire de l'archivistique ou une approche comparative des cadres internationaux.

Axe 3	Relations entre dispositif technique et données scientifiques : l'archive web en réseau

Une approche prospective de l'étude des sources web implique de les envisager non pas de manière isolée, mais en lien avec les différents types de données et de sources mobilisées dans le cadre de la démarche scientifique. Les approches méthodologiques qui mobilisent les matériaux numériques sont plurielles. Les analyses de traces, l'analyse de réseau, la

visualisation de données (approches quantitatives) peuvent être complétées par des approches qualitatives en ethnographie numérique qui cherchent à questionner la relation des traces aux utilisateurs. En lien avec les enjeux des données de la recherche, les pratiques de constitution des corpus par les chercheurs peuvent s'appuyer sur les savoir-faire archivistiques : cycle de vie des données, éditorialisation et documentation des sources, préservation des dispositifs techniques liés aux données, préservation des données de la recherche. On s'intéressera ici aux limites légales et techniques actuelles à la constitution de l'archive du web considérée comme une archive incomplète : l'archivage de certains objets-limites impliquant une interactivité technologique (ex. jeux vidéo en ligne, réseaux sociaux, logiciels) conduit à questionner la nature documentaire du web. A contrario, les technologies d'archivage numérique peuvent être mobilisées pour collecter et préserver des objets documentaires accessibles en ligne que l'on n'assimile pas habituellement aux archives web (presse quotidienne, articles scientifiques). Cet axe portera donc sur tous les enjeux, méthodes et réflexions permettant d'envisager le web comme source et archive non pas isolée, mais liée avec d'autres corpus et collections.

Calendrier prévisionnel

Diffusion de l'appel à communication	Mai 2022
Réception des propositions : – Résumés pour les conférences plénières (4500 caractères max., bibliographie comprise) – Pour les tables rondes (présentations courtes), les résumés représentent 1 page (1500 caractères, espaces compris)	~~Fin août 2022~~ **30 septembre 2022**
Retours du comité scientifique aux auteurs, acceptation des propositions	Fin novembre 2022
Réception des articles complets (25 000 à 35 000 signes espaces compris) et des courtes présentations	Mi-janvier 2023
Colloque	3,4 et 5 avril 2023

Procédure d'évaluation et participation

Les contributions anonymisées seront évaluées en double-aveugle.
Pour participer et adresser une contribution : https://respadon.sciencesconf.org/

Références

Emmanuelle BERMES (2019) « Quand le dépôt légal devient numérique : épistémologie d'un nouvel objet patrimonial », *Quaderni* [En ligne], 98 | Hiver 2018-2019, URL : http://journals.openedition.org/quaderni/1455 ; DOI : https://doi.org/10.4000/quaderni.1455

Niels BRÜGGER (2018) *The archived web : Doing history in the digital age*. MIT Press.

Valérie GAME, Gildas ILLIEN (2006) « Le Dépôt légal d'Internet à la Bibliothèque nationale de France », in Bulletin des bibliothèques de France (BBF), n°3, p. 82-85. En ligne, URL : http://bbf.enssib.fr/consulter/bbf-2006-03-0082-013. ISSN 1292-8399.
Illien, 2008

Valérie BEAUDOUIN, Philippe CHEVALLIER, Lionel MAUREL (2018) *Le web français de la Grande Guerre. Réseaux amateurs et institutionnels*. Presses universitaires de Paris Nanterre.

Valérie BEAUDOUIN, Zeynep PEHLIVAN (2016) « Cartographie de la Grande Guerre sur le Web : Rapport final ». Bibliothèque nationale de France ; Bibliothèque de documentation internationale contemporaine ; Télécom ParisTech.

Jérôme DENIS (2018) *Le travail invisible des données. Éléments pour une sociologie des infrastructures scripturales*, Presses des Mines, 208 p.

Gildas ILLIEN, Pascal SANZ, Sophie SEPETJAN, Peter STIRLING (2011) « La situation du dépôt légal de l'internet en France : retour sur cette nouvelle législation, sur sa mise en pratique depuis cinq ans, et perspectives pour le futur ». Actes du 77e congrès de la Fédération internationale des associations de bibliothécaires et d'institutions (IFLA), URL : http://conference.ifla.org/past-wlic/2011/193-stirling-fr.pdf

Emily MAEMURA (2021) « Data Here and There: Studying Web Archives Research Infrastructures in Danish and Canadian Settings ». University of Toronto, Faculty of Information, Doctoral Paper.

MAEMURA, Emily, BECKER, Christoph, et MILLIGAN, Ian (2016) « Understanding computational web archives research methods using research objects ». In : *IEEE International Conference on Big Data*, p. 3250-3259, DOI : 10.1109/BigData.2016.7840982

Frédéric MARTIN (2017), « Les archives de l'internet comme axe de coopération nationale ». In : *Webcorpora*, URL : https://webcorpora.hypotheses.org/394

Francesca MUSIANI (ed.) (2019) *Qu'est-ce qu'une archive du web ?* OpenEdition Press, URL : https://doi.org/10.4000/books.oep.8713

Jean-Charles PAJOU (2016) « L'Observatoire du dépôt légal : un certain regard sur l'édition », Bulletin des bibliothèques de France (BBF), n° 9, p. 134-144.
En ligne, URL : https://bbf.enssib.fr/consulter/bbf-2016-09-0134-002 ISSN 1292-8399.

Nick RUEST, Jimmy LIN, Ian MILLIGAN, Samantha FRITZ (2020) « The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives ». IEEE/ACM Joint Conference on Digital Libraries, Wuhan, Chine.
En ligne, URL : https://doi.org/10.1145/3383583.3398513

Valérie SCHAFER, Francesca MUSIANI, Marguerite BORELLI (2016) « Negotiating the web of the past ». *French Journal for Media Research*, La toile négociée/Negotiating the web, http://frenchjournalformediaresearch.com/lodel/index.php?id=963

Peter STIRLING (2017) « Le dépôt légal de l'internet dans le projet Corpus ». In : *Webcorpora*, URL : https://webcorpora.hypotheses.org/111

RES
PA
DON

International Conference

April 3rd, 4th and 5th 2023

THE WEB : SOURCE AND ARCHIVE

Project display

This international conference proposes to question the place of sources from the web in the scientific field and to situate web archiving practices in plural scientific approaches and questions.

Founded by the GIS CollEx-Persée, the project is undertaken by the University of Lille and the National Library of France, in partnership with the GERiiCO of the University of Lille, Sciences Po and the Condorcet Campus, it brings libraries and research teams together to think, experiment and share practices related to web archives. The main goal is to bring the producers and the users of the web archive collection closer together, with the help and the mediation of academic libraries. In this perspective, the international symposium organized by the scientific committee led by Laurence Favier (Pr. University of Lille, GERiiCO), Emmanuelle Bermès (Curator, PhD, Deputy director for services and networks – Bibliothèque nationale de France) and Madeleine Géroudet (Curator, Head of research support Departement - University of Lille Library) wishes to develop exchanges of knowledge between STI professionals and scientific teams by engaging the international academic community to participate in its work.

Organized at LILLIAD (Learning Center Innovation of the University of Lille) on the Campus of the Faculty of Science and Technology in Villeneuve D'Ascq, the event will take place over 2 and a half days, on April 3, 4 and 5, 2023.

The work of these days will be published.

General introduction – scientific positioning

It is only if it is the object of archiving practices that the web becomes a corpus of research (Beaudouin, Pehlivan, 2016 : 15). As a medium of ephemeral inscription of signs, codes and data that circulate through digital networks, the web as scientific material questions researchers as a source or research field and questions information and documentation professionals as an archive.

From 2013 to 2016, the project "Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre" (The online future of digitized heritage: the example of the Great War) was the first to use a collection from the Internet archives to apply global analysis methods. A corpus of sites was then specifically delimited to proceed with the collection that would allow for the analysis of amateur practices around digitized heritage content made available by libraries. In the spring of 2022, the Datasprint organized in the framework of the Respadon project confirmed that the mobilization of web archives could be complementary to the study of the living web. The exploration of the web as a source or as a field of study is now essential to many disciplines : contemporary history, political science, sociology, history of science, information and communication science. And because it allows one to make the history of an academic field and to inscribe one's work in a tradition, the web has become a necessary source for many scientific approaches.
However, the scriptural infrastructures (Denis, 2018) that allow the production of data and its circulation through the web are paradoxically fragile and the contents volatile. This is why cultural and heritage institutions such as the Bibliothèque nationale de France (BnF) and the National Audiovisual Institute (INA) took an early interest in Internet archiving.
As early as 1996, the Internet Archive began to build an international archive of the Web. In France, the law known as "DADVSI" of 2006 (Copyright and Neighboring Rights in the Information Society) extends legal deposit to « signs, signals, writings, images, sounds or messages of any kind communicated to the public by electronic means ». It contributes to defining the French web space as a national heritage representative of the editorial production of its time (Bermès, 2019 ; Ilien et al., 2011).
Many other countries have adopted a similar legal framework and are united in the International Internet Preservation Consortium (IIPC). In parallel, researchers who have started to study web archives as an object have structured themselves into networks with WARCnet and RESAW (Schafer, Musiani, Borelli, 2016 ; Brügger, 2018).

In line with the objectives of the Respadon project (Network of partners for the exploration and analysis of digital data), which consists in making web archives more accessible to researchers

from many disciplines, this conference proposes to question the scientific practices where the web can be seized as a source and as an archive.
This event proposes to bring together both concrete examples of the scientific uses of web archives in relation to their role as source and field, and heuristic reflections on the nature of these archives and the ways in which they are preserved and made available by heritage and documentary institutions.

The expected papers will focus on the following themes related to the 3 main axes of the conference, described below :

- Presentations of scientific projects using web corpora and digital humanities
- Presentations of feedback from research projects using the web as a source : obstacles encountered, success stories
- Reflections on professional and academic practices involving the collection and preservation of web-based data
- Presentation of experiments and devices promoting access to digital archives and web corpora
- Methodological and epistemological reflections on the needs of accessing and preserving online data in different disciplines.

Axis 1 The web at the intersection of memory and knowledge : epistemological issues

The web as a scientific field implies to consider the material available online as a deposit of traces of the collective memory. How to elaborate a scientific discourse from an essentially volatile source ? What kind of memory is involved when it is "distributed" on the web ? What is the status of web sources in the scientific practices that elaborate a new knowledge and contribute to the emergence of sciences henceforth assisted by the digital ? The aim here is to collectively question the methods, modalities and status of the results produced when online data come into play in research, whatever the discipline concerned (contemporary history, political science, sociology, history of science, media studies, computer science...). Approaches and proposals that illustrate scientific work using the Web and its traces as well as thoughts on the nature of archives and data used in contemporary scientific research will be appreciated.

Axis 2 Archival policies, practices, techniques and web archives : from document to corpus

For several years, the professional community has developed technical definitions of the digital archive and associated devices. This practice questions the notion of "archive places" and the preservation of corpora : the service offer around web archives must allow the emergence of the working field with researchers, taking into account various questioning :

- In terms of informational sovereignty, how does the diversity of existing private and public initiatives materialize in the policies for building archives ?

- In terms of access to archives, what definition of granularity, unity of meaning, and coherence of the archival collection is needed to make web archives usable for research ?

- How to develop efficient scientific practices while respecting the legal framework and the constraints related to the notion of legal deposit ? How can we make this framework evolve to promote research ?

To shed light on these issues, we could consider mobilizing the history of archiving or a comparative approach to international frameworks.

Axis 3 Relations between technical devices and scientific data : the networked web archive

A prospective approach to the study of web sources implies considering them not in isolation, but in relation to the different types of data and sources mobilized in the framework of the scientific process. The methodological approaches that mobilize digital materials are plural. Trace analysis, network analysis, and data visualization (quantitative approaches) can be complemented by qualitative approaches in digital ethnography that seek to question the relationship between traces and users. In connection with the issues of research data, the practices of constitution of corpora by researchers can be based on archival know-how : life cycle of data, editorialization and documentation of sources, preservation of technical devices related to data, preservation of research data. We will focus here on the current legal and technical limits to the constitution of the web archive considered as an incomplete archive : the archiving of some object-limits involving technological interactivity (e.g. online video games, social networks, software) leads to question the documentary nature of the web. On the other

hand, digital archiving technologies can be mobilized to collect and preserve documentary objects accessible online that are not usually considered as web archives (daily press, scientific articles). This axis will therefore focus on all the issues, methods and reflections that allow us to consider the web as a source and archive that is not isolated, but linked with other corpora and collections.

Provisional timetable

Distribution of the call for papers	May 2022
Receipt of proposals : – Summaries for plenary conferences (4500 characters max., bibliography included) – round tables (abstracts et short présentations) : 1 page, 1500 characters	~~End of August 2022~~ **September 30th, 2022**
Feedback from the scientific committee to the authors, acceptance of proposals	End of November 2022
Receipt of complete articles (25 000 à 35 000 characters, spaces included) and short presentations	Mid-january 2023
Conference	April 3,4 and 5 2023

Evaluation procedure and participation

Anonymized contributions will be double-blind evaluated.
To participate and submit a contribution : https://respadon.sciencesconf.org

References

Emmanuelle BERMES (2019) « Quand le dépôt légal devient numérique : épistémologie d'un nouvel objet patrimonial », *Quaderni* [En ligne], 98 | Hiver 2018-2019, URL : http://journals.openedition.org/quaderni/1455 ; DOI : https://doi.org/10.4000/quaderni.1455

Niels BRÜGGER (2018) *The archived web : Doing history in the digital age*. MIT Press.

Valérie GAME, Gildas ILLIEN (2006) « Le Dépôt légal d'Internet à la Bibliothèque nationale de France », in Bulletin des bibliothèques de France (BBF), n°3, p. 82-85. En ligne, URL : http://bbf.enssib.fr/consulter/bbf-2006-03-0082-013. ISSN 1292-8399.
Illien, 2008

Valérie BEAUDOUIN, Philippe CHEVALLIER, Lionel MAUREL (2018) *Le web français de la Grande Guerre. Réseaux amateurs et institutionnels*. Presses universitaires de Paris Nanterre.

Valérie BEAUDOUIN, Zeynep PEHLIVAN (2016) « Cartographie de la Grande Guerre sur le Web : Rapport final ». Bibliothèque nationale de France ; Bibliothèque de documentation internationale contemporaine ; Télécom ParisTech.

Jérôme DENIS (2018) *Le travail invisible des données. Éléments pour une sociologie des infrastructures scripturales*, Presses des Mines, 208 p.

Gildas ILLIEN, Pascal SANZ, Sophie SEPETJAN, Peter STIRLING (2011) « La situation du dépôt légal de l'internet en France : retour sur cette nouvelle législation, sur sa mise en pratique depuis cinq ans, et perspectives pour le futur ». Actes du 77e congrès de la Fédération internationale des associations de bibliothécaires et d'institutions (IFLA), URL : http://conference.ifla.org/past-wlic/2011/193-stirling-fr.pdf

Emily MAEMURA (2021) « Data Here and There: Studying Web Archives Research Infrastructures in Danish and Canadian Settings ». University of Toronto, Faculty of Information, Doctoral Paper.

MAEMURA, Emily, BECKER, Christoph, et MILLIGAN, Ian (2016) « Understanding computational web archives research methods using research objects ». In : *IEEE International Conference on Big Data*, p. 3250-3259, DOI : 10.1109/BigData.2016.7840982

Frédéric MARTIN (2017), « Les archives de l'internet comme axe de coopération nationale ». In : *Webcorpora*, URL : https://webcorpora.hypotheses.org/394

Francesca MUSIANI (ed.) (2019) *Qu'est-ce qu'une archive du web ?* OpenEdition Press, URL : https://doi.org/10.4000/books.oep.8713

Jean-Charles PAJOU (2016) « L'Observatoire du dépôt légal : un certain regard sur l'édition », Bulletin des bibliothèques de France (BBF), n° 9, p. 134-144.

En ligne, URL : https://bbf.enssib.fr/consulter/bbf-2016-09-0134-002 ISSN 1292-8399.

Nick RUEST, Jimmy LIN, Ian MILLIGAN, Samantha FRITZ (2020) « The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives ». IEEE/ACM Joint Conference on Digital Libraries, Wuhan, Chine.
En ligne, URL : https://doi.org/10.1145/3383583.3398513

Valérie SCHAFER, Francesca MUSIANI, Marguerite BORELLI (2016) « Negotiating the web of the past ». *French Journal for Media Research*, La toile négociée/Negotiating the web, http://frenchjournalformediaresearch.com/lodel/index.php?id=963

Peter STIRLING (2017) « Le dépôt légal de l'internet dans le projet Corpus ». In : *Webcorpora*, URL : https://webcorpora.hypotheses.org/111