Théia

1 | 2024
Retours d'expérience en édition numérique de sources en histoire et histoire de l'art

Éditer la recherche et les données de la recherche : conjuguer plateforme de livres et entrepôt de données

Publishing research and research data: combining a book platform and a data warehouse

Christine Chadier

http://publications-prairial.fr/theia/index.php?id=161

DOI : 10.35562/theia.161

Electronic reference
Christine Chadier, « Éditer la recherche et les données de la recherche : conjuguer plateforme de livres et entrepôt de données », *Théia* [Online], 1 | 2024, Online since 17 avril 2025, connection on 20 septembre 2025. URL : http://publications-prairial.fr/theia/index.php?id=161

Éditer la recherche et les données de la recherche : conjuguer plateforme de livres et entrepôt de données

Publishing research and research data: combining a book platform and a data warehouse

Christine Chadier

OUTLINE

Pourquoi compléter des sites de publications en ligne par un entrepôt de données ?
Pourquoi choisir Nakala ?
  Une plateforme portée par une infrastructure de recherche
  Nakala ? Qu'est que c'est ?
Construire ses données dans Nakala
  Les métadonnées dans Nakala
    Les métadonnées générées par Nakala
    Les métadonnées à ajouter par les déposants
    Où peuvent se retrouver les métadonnées en dehors de Nakala ?
  Les identifiants dans Nakala
  Les images dans Nakala
    IIIF : International Image Interoperability Framework
  Générer un site web personnalisé avec « Nakala_Press »
La « collection » Chrétiens et Sociétés sur Nakala
  Le rôle de l'éditrice ou de l'éditeur

AUTHOR'S NOTES

Cette communication se veut « un retour d'expérience » dans le cadre des choix éditoriaux faits au sein du LARHRA pour les éditions « Chrétiens et Sociétés ». Il n'a pas pour ambition de valoriser tel ou tel procès, mais d'exposer les raisons qui ont conduit un laboratoire-éditeur à faire de tels choix en matière de stockage de données. Le dispositif est évolutif et continue d'être développé actuellement.

TEXT

Pourquoi compléter des sites de publications en ligne par un entrepôt de données ?

1 En France, l'ouverture des données de recherche a été prévue dans le premier volet de la loi pour une République numérique (loi dite « Axelle Lemaire » n° 2016-1321 du 7 octobre 2016) puis inscrite depuis dans le code de la recherche :

> Article 533-4-II du code de la recherche
> I.- Lorsqu'un écrit scientifique issu d'une activité de recherche financée au moins pour moitié par des dotations de l'État, des collectivités territoriales ou des établissements publics, par des subventions d'agences de financement nationales ou par des fonds de l'Union européenne est publié dans un périodique paraissant au moins une fois par an, son auteur dispose, même après avoir accordé des droits exclusifs à un éditeur, du droit de mettre à disposition gratuitement dans un format ouvert, par voie numérique, sous réserve de l'accord des éventuels coauteurs [...]
> II.- Dès lors que les données issues d'une activité de recherche [...] ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, leur réutilisation est libre.
> III.- L'éditeur d'un écrit scientifique mentionné au I ne peut limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication.

2 Les jeux de données rassemblés au cours d'une recherche deviennent ainsi aussi importants que la publication des résultats scientifiques. Leur mise à disposition à l'issue d'un programme de recherche peut même être exigée par les partenaires financiers (demandes de financement de projets européens H2020, programmes ANR [1]). L'éditeur, particulièrement l'éditeur scientifique public, se doit donc

d'accompagner ce mouvement afin que la publication des données n'entre pas en concurrence avec la publication des écrits scientifiques (articles ou ouvrages) et afin de ne pas déconnecter les données de la recherche des résultats de la recherche.

3 Si on veut que les données soient publiées selon les critères FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable) qui favorisent, outre l'ouverture des données, leur mise à disposition avec un triple objectif de qualité des données et des métadonnées, d'inscription dans un cycle de vie maitrisé par les scientifiques et enfin de pérennité des données sur le long terme (accès, intégrité, contextualisation de la production des données), il convient de proposer, en appui des revues et des collections d'ouvrages une solution pour l'hébergement des données de la recherche [2] et de déterminer la solution la mieux adaptée à la discipline et aux choix éditoriaux.

Pourquoi choisir Nakala ?

4 Avec l'obligation d'ouvrir les données de la recherche, différents entrepôts de données se mettent en place dont plusieurs spécialisés en SHS, mais certains sont institutionnels (Collection Sciences Po - data.sciencespo réservé à la communauté académique de Sciences Po et partenaires de Sciences Po), d'autres réservés à un type de données comme Quetelet-Progedo (données quantitatives – données d'enquêtes, données quantitatives tabulées) ou le Conservatoire national des données 3D qui utilise un Schéma de métadonnées « maison » vu la spécificité des données hébergées [3].

Une plateforme portée par une infra-structure de recherche

5 Nakala, portée par l'infrastructure de recherche Huma-Num mise en œuvre par le CNRS avec le Campus Condorcet et Aix-Marseille Université a pour principale mission de construire, avec les communautés et à partir d'un pilotage scientifique, une infrastructure numérique de niveau international inscrits dans les principes FAIR. Engagée dans l'European Open Science Cloud, elle porte la participation de la France dans l'European Research

Infrastructure Consortium [ERIC DARIAH (Digital Research Infrastructure for the Arts and Humanities)].

6 Depuis janvier 2021, Huma-Num porte avec OpenEdition et Métopes le Consortium de moyens mutualisés pour des services et données ouvertes en SHS (COMMONS). COMMONS, projet lauréat de l'appel à manifestations d'intérêt « Équipements structurants pour la recherche : EquipEx+ » du Programme d'investissements d'avenir (PIA 3), associe trois infrastructures françaises qui jouent un rôle essentiel dans les Sciences humaines et sociales et en particulier dans l'édition en SHS. COMMONS permet d'aborder l'ensemble de la chaîne de production des connaissances : de la gestion des données à la diffusion et l'édition. Publiant en accès ouvert nos ouvrages et revue sur OpenBooks et OpenJournals, les plateformes d'OpenEdition et utilisant la chaine Métopes [4] pour le travail éditorial, il était donc logique que nous nous tournions vers Huma-Num et Nakala pour choisir le lieu d'hébergement pour publier les données ayant été utilisées pour la rédaction des publications éditées par nos soins ou les données complémentaires qui pourront ainsi plus facilement être réutilisées que des banque d'images, des tableaux statistiques, des reproduction de documents, etc.

7 Nous avons donc choisi pour accompagner la revue *Chrétiens et Sociétés* XVI^e^-XXI^e^ *siècles* (https://journals.openedition.org/chretienssocietes/) et la collection « Chrétiens et Sociétés. Documents et mémoires » (https://books.openedition.org/larhra/634) de créer une collection dans Nakala.

Nakala ? Qu'est que c'est ?

8 Nakala est l'un des centres de référence de l'écosystème national Recherche Data Gouv.fr, Un écosystème au service du partage et de l'ouverture des données de recherche. Se rapprocher de réseaux structurés comme ceux sur lesquels s'appuie Nakala permet de bénéficier de services offrant stabilité et visibilité… Contrairement à d'autres entrepôts de données comme Zenodo porté par le CERN (Conseil européen pour la recherche nucléaire), il s'agit d'un entrepôt de données spécialisé en SHS. Lancé dès 2014, il répond aux besoins d'accès persistant et interopérable aux données numériques. Il se compose de deux ensembles :

- un visible de tous permettant l'accès aux données
- un qui sera utilisé par les robots pour pouvoir moissonner les données qui contient la présentation des métadonnées

9 La qualité et le soin apporté au traitement des métadonnées sont la clef de voûte pour une meilleure exposition des données de la recherche. Les métadonnées, c'est-à-dire des informations de description ou « données à propos des données », permettent de décrire les caractéristiques d'une donnée ou d'un jeu de données, par exemple :

- auteurs/autrices
- description du contenu
- date de création
- lieu de capture/de production
- raison pour laquelle les données ont été générées
- comment les données ont été créées
- etc.

10 Les métadonnées doivent être aussi riches, précises et exactes que possible.

11 Nakala propose des services interopérables de présentation des métadonnées

- exposition dans un triplestore RDF [5]
- accès via un entrepôt OAI-PMH [6].

12 Outre cette structuration des données, Nakala assure l'accès aux données :

- accès et citation par un identifiant pérenne (Digital Object Identifier, DOI)
- stockage sécurisé
- accès permanent.

Construire ses données dans Nakala

13 Un dépôt, au sens où Nakala l'entend, est un regroupement d'informations contenant

des données : des fichiers numériques (au minimum 1)
- des métadonnées
- un identifiant pérenne de type DOI attribué automatiquement à chaque donnée publique et déclaré auprès de DataCite.

14 Nakala contient aussi d'autres informations

- des collections (regroupements de références à des données)
- des référentiels (vocabulaires utilisés dans les descriptions).

15 C'est cet ensemble qui doit être appréhendé pour comprendre et pouvoir réutiliser les données.

Les métadonnées dans Nakala

16 Les données déposées, mais aussi les collections et les référentiels, doivent être accompagnés de métadonnées :

- certaines sont générées par Nakala
- les autres sont ajoutées par le déposant.

Les métadonnées générées par Nakala

17 Les métadonnées générées automatiquement par Nakala se composent d'informations de gestion :

- dates de dépôt et de dernière modification, version, statut, identification du déposant, identifiant du dépôt

18 et d'informations techniques sur les fichiers :

- taille (en octets), type de média (mime), empreinte digitale (sha-1).

Les métadonnées à ajouter par les déposants

19 Ce sont les informations documentaires qui vont permettre d'identifier la donnée à deux niveaux.

- Au niveau du dépôt :

20 avec cinq champs obligatoires : Titre(s), Auteur(s), date de création, licence, type
et des champs facultatifs : le déposant peut ajouter l'ensemble des champs du vocabulaire Dublin-Core qualifié [7].

- Au niveau des fichiers :

21 la date de visibilité (embargo) : par défaut c'est la date du dépôt
une description facultative mais fortement conseillée afin de valoriser les données.

Où peuvent se retrouver les métadonnées en dehors de Nakala ?

22 Excepté la page d'administration qui n'est visible que par le seul déposant, tous les autres accès sont susceptibles d'être consommés par des services extérieurs à Nakala. Toutes les personnes et services qui moissonnent ces métadonnées sont susceptibles de les réexposer et leurs réexpositions d'être moissonnées et réutilisées à nouveau par d'autres personnes ou services.

23 Cela peut se faire :

- automatiquement par la découverte par des robots
- par les moteurs de recherche type Google, Google data search, Google scholar
- dans leurs articles, ouvrages, blogs, etc.

Les identifiants dans Nakala

24 Chaque donnée déposée dans Nakala se voit attribuer un identifiant pérenne : un Digital Object Identifier (DOI) [8]. Les DOI sont mis à disposition par Datacite [9]. L'identifiant (DOI) sert à construire des URLs pour obtenir l'accès aux différents éléments d'un dépôt

25 Pour accéder à la page d'accueil (*landing page*) d'un dépôt les URLs sont les suivants :

- https://nakala.fr/[DOI] ou https://doi.org/[DOI]
- https://nakala.fr/10.34847/nkl.4b33r2h4
- https://doi.org/10.34847/nkl.4b33r2h4.

26 Par extension cet URL est souvent considéré comme l'identifiant du dépôt.

Les images dans Nakala

27 Nakala intègre le protocole IIIF (International Image Interoperability Framework). Intégrer la communauté permet de mieux valoriser les documents numérisés (reproduction de documents, photographies d'œuvres ou de lieux, etc.) en donnant aux utilisateurs la possibilité d'accéder facilement aux images et au contexte de production de celles-ci. Mais surtout il s'agit d'un format interactif qui apporte aux images des fonctionnalités avancées de visualisation et de manipulation des documents (zoom profond, visualisation de documents en haute résolution, gains d'accessibilité, facilités d'accès et de partage des ressources grâce à des liens en principe pérennes, etc.).

IIIF : International Image Interoperability Framework

28 IIIF désigne à la fois une communauté et un cadre d'interopérabilité pour diffuser, présenter et annoter des images et documents audio/vidéo sur le Web, qui s'est imposé en quelques années comme un standard et une brique technologique essentielle pour décloisonner les collections numérisées des institutions patrimoniales à l'échelle mondiale. Cela recouvre un modèle pour présenter et annoter des contenus numériques (images, documents audio et vidéo) et un cadre de réflexion sur les mécanismes d'échange des images entre entrepôts numériques : Nakala, BnF, BL, Cornell, Los Alamos National Laboratory, NL of Norway, Oxford, Stanford…

29 IIIF propose un cadre technique commun grâce auquel les fournisseurs et créateurs d'images numériques peuvent :

- délivrer leurs images de manière standardisée sur le Web
- les rendre consultables, manipulables et annotables
- rendre cela possible pour n'importe quelle application ou logiciel compatible.

30 Utiliser les standards du IIIF c'est offrir aux images utilisées pour la rédaction d'ouvrages et d'articles scientifiques une visibilité et une interopérabilité qu'elles ne pourraient avoir en restant seulement

insérées dans un texte même publié en accès ouvert sur une plateforme [10].

Générer un site web personnalisé avec « Nakala_Press »

31 À terme, il est prévu de générer un site web avec Nakala-Press, lorsque nous aurons accumulé dans notre collection Nakala suffisamment de données liées à *Chrétiens et Sociétés* pour l'alimenter de façon conséquente. Nakala-Press est un module de publication qui permet de créer un site web autour de données publiques déposées dans Nakala. Ce site web permettra de mettre en valeur les données rassemblées et d'en faciliter l'accès à un public plus large.

La « collection » Chrétiens et Sociétés sur Nakala

32 Le premier volume publié en association OpenBooks/Nakala sera le tome 4 des *Matériaux pour l'histoire religieuse du peuple français Bourgogne, Franche-Comté et Lyonnais Savoie et Dauphiné Grand Midi et Algérie* (figure 1) [11].

33 Les notices seront mises en ligne sous la forme d'un volume de la collection *Chrétiens et Sociétés. Documents et Mémoires* sur OpenBooks (https://books.openedition.org/larhra/634) [12].

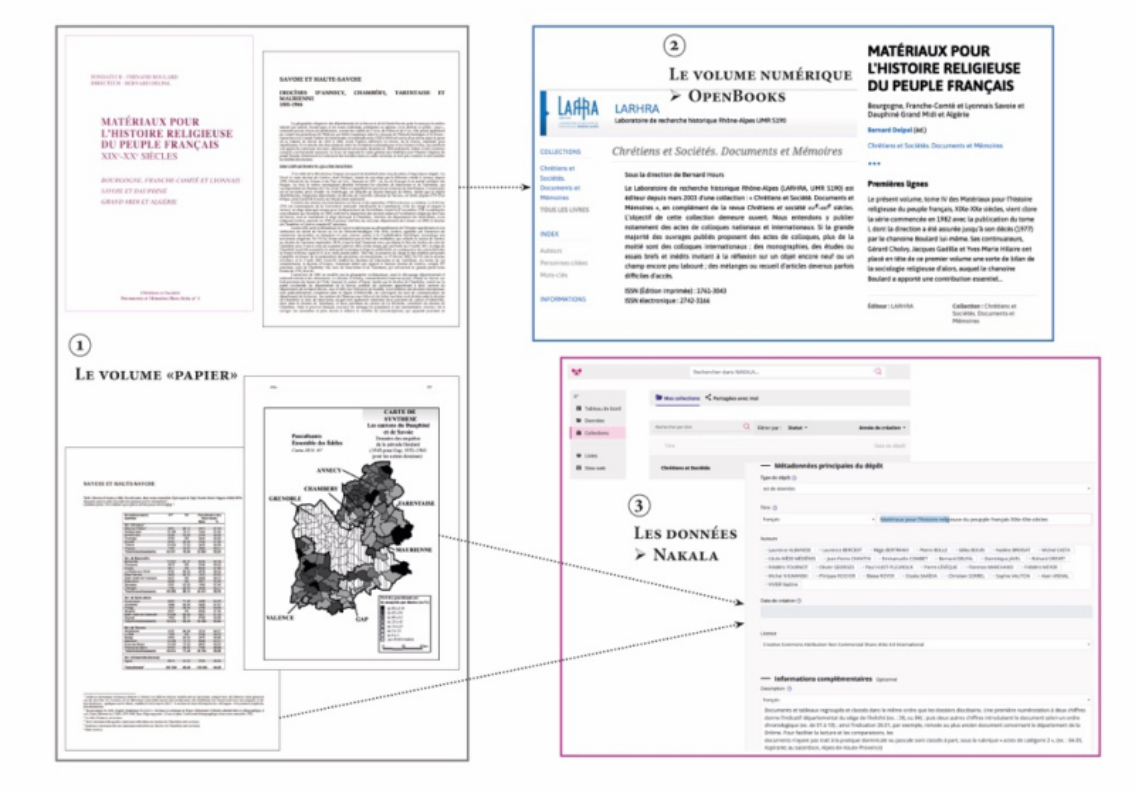
34 Il s'agira ensuite de déposer le reste du volume sur Nakala. Nous aurons deux types de données :

- une « donnée » [au sens Nakala] qui rassemblera l'ensemble des tableaux avec un fichier par département, ce qui reprend la structuration du volume papier, avec une feuille pour chaque tableau, le tout en .csv [13].
- une « donnée » qui rassemblera l'ensemble des cartes au format pdf. Le format pdf a été retenu de façon à pouvoir conserver la même configuration que pour les tableaux, soit un fichier par département contenant un ensemble de carte.

35 La convention de nommage des fichiers reprendra celle des tableaux de la version papier. Le choix de la même structure pour les fichiers des tableaux et des cartes permettra de mettre facilement en regard

les tableaux mais ne permettra pas d'appliquer le protocole IIIF. C'est un choix éditorial pour s'intégrer au mieux à la structure des textes du volume mis en ligne sur OpenBooks. Cela n'empêchera pas de créer un jeu de métadonnées développé pour donner visibilité et accessibilité tant aux cartes qu'aux tableaux.

Figure 1. Mise en ligne du tome 4 des *Matériaux pour l'histoire religieuse du peuple français Bourgogne, Franche-Comté et Lyonnais Savoie et Dauphiné Grand Midi et Algérie*



Réalisation Christine Chadier

Le rôle de l'éditrice ou de l'éditeur[14]

36 Ce nécessaire travail de relecture et d'harmonisation qui apparaît dans le cadre de la mise en ligne des données des « Matériaux Boulard » soulève la question de l'éditorialisation des données[15]. En effet, si on peut envisager, dans le cadre d'un projet de recherche ou d'un simple « *work in progress* », de déposer des données directement issues de la recherche qui constituent en soi un ensemble unique, cela me paraît plus difficile dans le cadre d'une

collection en appui de plateformes d'édition. L'édition impliquant en effet tout un travail d'harmonisation et de suivi afin de valoriser le texte publié, il convient de prévoir le même travail d'harmonisation sur les données si l'on souhaite les valoriser au-delà de la seule mise à disposition de la communauté. Le travail d'éditorialisation est essentiel pour favoriser la lisibilité des données. Le dépôt des données « brutes » suffit pour de la simple récupération de données en vue de leur réutilisation ou le moissonnage par les machines, mais si l'on veut qu'elles puissent être consultées aisément, il faut un travail de préparation.

37 Dans ce cadre, il me paraît indispensable d'avoir des données structurées de façon identique, par exemple par ouvrage ou par numéro de revue, avec des conventions de nommage qui permettent de rattacher facilement la donnée à son ouvrage ou son numéro de revue et obtenir des URL un minimum standardisées. C'est pourquoi il me semble nécessaire que la personne responsable de la collection dans Nakala soit également la·le responsable des sites de publication sur les plateformes. La publication des données viendrait donc s'ajouter aux compétences de l'éditrice ou de l'éditeur. L'édition électronique est venue s'ajouter à l'édition imprimée à la fin des années 1990 et au début des années 2000 comme la publication des données et la gestion d'un entrepôt de données vient s'ajouter aujourd'hui à l'édition électronique.

38 L'évolution des compétences doit donc suivre l'évolution des types de contenus à publier et des outils. Cela nécessite une collaboration accrue avec les auteurs pour valider l'éditorialisation des données qu'il a précieusement accumulées au cours de ses recherches et qu'il n'avait pas forcément prévu de publier. Le rôle des chercheurs est de veiller à la qualité des données publiées et le rôle de l'éditrice ou de l'éditeur, comme elle·il le fait déjà pour les revues ou les ouvrages, est de veiller au respect du travail du chercheur mais également à la bonne visibilité des données, aux conditions de réutilisation et à ce qu'elles puissent bénéficier d'un stockage « pérenne », autant que nous puissions l'envisager.

39 Il est à noter que cette montée en compétences de l'éditrice ou de l'éditeur, rendue possible en s'inscrivant, par exemple, dans des réseaux « métiers » ou des listes de diffusion [16], permet une montée

en qualité des revues ou des collections et de quitter le statut de « laboratoire-éditeur » pour rejoindre le cercle des 106 éditeurs scientifiques publics recensés par Caroline Dandurand lors de son enquête de 2022 [17] aux côtés des Presses universitaires de Lyon, des Presses universitaires de Rennes, de Sorbonne Université, des Presses UGA Éditions, des Publications de l'École française de Rome, d'ENS Éditions, de MOM Éditions - Maison de l'Orient et de la Méditerranée, etc. En choisissant de rejoindre en 2023 Nakala après avoir rejoint en 2008 OpenEdition, Chrétiens et Sociétés suit le même chemin que *GalliaArchéologies des Gaules*, revue CNRS. Cette évolution ne serait pas possible sans s'appuyer sur un contenu scientifique reconnu par la communauté et validé par ses pairs. La raison d'être du travail réalisé est d'apporter des outils de diffusion de qualité à un contenu de qualité.

Figure 2. Le rôle de l'éditrice ou de l'éditeur

Musique d'ensemble, Testu & Massin (Paris) [18], EST 5 C/43, Estampe https://selene.bordeaux.fr/ark:/27705/330636101_EST_5_C_43/v0001.simple.highlight=musique%

NOTES

1 Cécile Arènes, Lionel Maurel, Stéphanie Rennes, *Guide d'application de la Loi pour une République numérique pour les données de la recherche*, Comité pour la science ouverte, 2022 ⟨hal-03968218⟩

2 Romain Feret, Françoise Catherine Gouzi, Sandra Guigonis, Hélène Jouguet, Nicolas Larrousse, *et al.*, Recommandations aux revues souhaitant définir une « politique de données » liées aux publications [Rapport Technique], Comité pour la science ouverte, 2020, 8 p. ⟨hal-03594383⟩

3 « Publication d'une liste d'entrepôts pour les données en SHS », *Le blog d'Huma-Num et de ses consortiums*, URL : https://humanum.hypotheses.org/9880

4 Métopes fournit un ensemble d'outils et de méthodes permettant de créer des fonds éditoriaux structurés et d'assurer la diffusion des produits éditoriaux, numériques ou imprimés, à partir de fichiers texte et facilite ainsi la diffusion des résultats et des données de la recherche.

5 Un triplestore RDF (Resource Description Framework) agit comme une base de données relationnelle : il stocke des données et les récupère via un langage de requête. Mais contrairement à une base de données relationnelle, un triplestore ne stocke qu'un seul type de données : le triplet (sujet - prédicat - objet).

6 L'Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) est un protocole informatique développé par l'Open Archives Initiative afin d'échanger des métadonnées.

7 Le Dublin Core est un format descriptif simple et générique créé en 1995 à Dublin (Ohio) par OCLC (Online Computer Library Center) et le NCSA (National Center for Supercomputing Applications). L'objectif du Dublin Core est de fournir un socle commun d'éléments descriptifs pour améliorer le signalement et la recherche de ressources au-delà des diverses communautés et des nombreux formats descriptifs propres à chaque spécialité, tout en restant suffisamment structuré. Pour en savoir plus : https://www.bnf.fr/fr/dublin-core.

8 Les DOI sont attribués pour les données scientifiques mais aussi pour d'autres types de ressources telles que des publications, des jeux et bases de données, des logiciels, des images, des cartes, etc.

9 DataCite est un consortium international de bibliothèques et services spécialisés dans les sciences de l'information, qui vise à faciliter l'archivage numérique ainsi que l'accès aux ressources numériques sur Internet, notamment par l'attribution d'un DOI à chacune d'entre elles. DataCite a été fondé le 1er décembre 2009 à Londres

10 Pour en savoir plus sur le IIIF : https://francearchives.gouv.fr/fr/article/705250527.

11 Les données ont été préparées par Anna Sébille, stagiaire du Master Sciences des religions et sociétés.

12 Suite à la fermeture de la plateforme OpenBooks du 19 février 2024 jusqu'à fin mars 2024 au plus tôt pour permettre la migration du système d'exploitation, la mise en ligne du volume est retardée.

13 Le format .csv (Comma Separated Values) est un format ouvert contrairement au format Document Microsoft Excel (.xls) ou Document Apple Numbers (.numbers) : c'est donc un format qui va maximiser la possibilité de réutilisation des données.

14 Nous entendons par « éditrice ou éditeur », l'ingénieur chargé de l'édition d'un ouvrage ou d'une revue et/ou la responsable éditoriale d'une collection.

15 Stéphane Renault, Blandine Nouvel, Micaël Allainguillaume, Astrid Aschehoug, Nicolas Coquet et Marie-Adèle Turkovics, « Harmoniser les pratiques éditoriales numériques des revues françaises d'archéologie », *Humanités numériques* [En ligne], 2 | 2020, mis en ligne le 01 juin 2020, consulté le 14 juin 2023. URL : http://journals.openedition.org/revuehn/483 ; DOI : https://doi.org/10.4000/revuehn.483

16 Par exemple le réseau Médici, réseau des métiers de l'édition scientifique ou la liste accesouvert@groupes.renater.fr, liste de discussion de la communauté du libre accès francophone.

17 Caroline Dandurand, *Préfiguration d'une structuration collective des éditeurs scientifiques publics engagés dans la science ouverte* [Rapport de recherche], Comité pour la science ouverte, 2022, 86 p. ⟨hal-03713434⟩

ABSTRACTS

Français
Il s'agit de présenter ici pourquoi et comment optimiser la mise en ligne d'un d'un article ou d'un ouvrage à l'ère de la science ouverte en présentant la démarche retenue pour les éditions « Chrétiens et Sociétés » publiée par le Laboratoire de recherche historique Rhône-Alpes.

English
The aim here is to explain why and how to optimise the online publication of an article or a book in the era of open science by presenting the approach adopted for the 'Chrétiens et Sociétés' editions published by the Rhône-Alpes Historical Research Laboratory.

INDEX

Mots-clés
science ouverte, entrepôt de données, Nakala, Huma-Num, OpenEdition, Métopes, COMMONS

Keywords
open science, data warehouse, Nakala, Huma-Num, OpenEdition, Métopes, COMMONS

AUTHOR

Christine Chadier
Université de Lyon Jean Moulin Lyon 3, LARHRA UMR 5190
IDREF : https://www.idref.fr/132141450
ISNI : http://www.isni.org/0000000357866609