”Dix ans avec CAHIER”. Bilan du consortium CAHIER (2011-2021) de la TGIR Huma-Num

Fatiha Idmhand

TGIR Huma-Num

Bilan du Consortium CAHIER (2011-2021)

Bilan préparé par Fatiha IDMHAND (Coordinatrice-adjointe du consortium de 2015 à 2021)

consortium cahier

Sommaire

TGIR Huma-Num - Bilan final du Consortium CAHIER

1. Identité du consortium

Nom du consortium : CAHIER Corpus d'Auteurs pour les Humanités : Informatisation, Édition, Recherche
Adresse Web : https://cahier.hypotheses.org/

2. Historique de la gouvernance

Années	NOM, Prénom	Institutions	Fonction
2011 – 2015 Le consortium est officiellement né le 01/01/2012 mais Marie-Luce Demonet travaillait à sa mise en place dès 2011.	DEMONET, Marie-Luce	Université de Tours	Professeur des Universités
2014 – 2015	DEMONET, Marie-Luce LEBARBE, Thomas	Université de Tours Université Grenoble Alpes	Professeur des Universités Professeur des Universités
2015 – 2019	LEBARBE, Thomas IDMHAND, Fatiha	Université Grenoble Alpes Université de Poitiers	Professeur des Universités Professeur des Universités
2019 – 09/09/2021	LEBARBE, Thomas IDMHAND, Fatiha	Université Grenoble Alpes Université de Poitiers	Professeur des Universités Professeur des Universités
09/09/2021 – 31/12/2021	IDMHAND, Fatiha	Université de Poitiers	Professeur des Universités
01/01/2022 – mise en place du nouveau projet	*LAVRENTEV, Alexey* *DORD-CROUSLE, Stéphanie*	*CNRS, CNRS, UMR-IHRIM, Lyon* *CNRS, CNRS, UMR-IHRIM, Lyon*	*Ingénieur de recherches CNRS* *Chargée de recherche CNRS*

3. Financement : liste des institutions gestionnaires

Années	Institutions	Personnes gestionnaires
2011 – 2015	Centre d'Etudes Supérieures de la Renaissance, Tours	RAGEOT, Laurence (MSH Val de Loire, gestionnaire administrative) Sandrine Vicente (CESR, gestionnaire financière)
2016 – 2019	Maison des sciences de l'Homme, Val de Loire, Tours	RAGEOT, Laurence (MSH Val de Loire, gestionnaire administrative) Alexandra Magné puis Sandrine Vicente (MSH Val de Loire, gestionnaire financière)
2019 – 09/09/2021		RAGEOT, Laurence (MSH Val de Loire, gestionnaire administrative) Claudie Vinet (MSH Val de Loire, gestionnaire financière)
09/09/2021 – 31/12/2021		
01/01/2022 – mise en place du nouveau projet		

4. Résumé des activités de 2021

a) Bilan année 2021. Budget année 2022.

L'année 2020, perturbée par la crise sanitaire, a profondément affecté les activités du consortium CAHIER. Le confinement de l'automne-hiver 2020-2021 et la réouverture tardive, en juin 2021, des espaces et lieux de recherches scientifiques ont conduit à l'annulation de toutes les activités prévues à partir d'octobre 2020 jusqu'à juin 2021. En conséquence, lorsque le budget a été présenté à Huma-Num le 30 octobre 2020, celui-ci faisait apparaître, comme chaque année, les dépenses réelles et les prévisions de dépenses de la fin d'année 2020. Or toutes les activités ont été suspendues et reportées au second semestre 2021. Ainsi, le consortium avait prévu :

- 10.000€ pour l'assemblée générale de novembre 2020 : celle-ci a dû être annulée de même que les réunions du comité de pilotage et des groupes de travail qui la précèdent habituellement, ainsi que la troisième et dernière journée de formation de l'atelier annuel (journée dédiée à la formation à EVT[1])
- 3.000€ pour la formation TEI2 initialement prévue en octobre 2020, annulée elle aussi et finalement organisée avec une année de retard en septembre 2021 (ces 3.000€ ont donc été dépensés en 2021)
- 3.500€ pour des subventions à 2 projets mais il s'agissait de subventions approuvées en 2019[2] et qui accusaient un retard de versement qui a été lui-même accentué par la crise sanitaire (l'une, de 2800€, n'a été versée qu'en 2021 et l'autre ne sera finalement pas réclamée)
- Enfin, 2.113,99€ étaient destinés à la cotisation du consortium CAHIER au consortium TEI. Elle n'avait pu être versée en 2020 en raison de difficultés, de la part du consortium TEI, à téléverser la facture dans ChorusPro (méthode devenue obligatoire) et a donc été payée en 2021, avec retard.

Au total, sur l'ensemble de l'année 2020, seule la première partie de l'atelier annuel a pu être organisée en présentiel (en septembre 2020 à Lorient) et deux groupes de travail ont pu se réunir. Aucune activité en présentiel n'a pu se tenir jusqu'au colloque « Dix ans avec CAHIER » du 7 au 10 juin 2021 à Bordeaux. Les stages des mastérants se sont déroulés à distance.

En conséquence, nos reliquats de l'année 2020 se sont élevés à **35.097,81€**, auxquels sont venus s'ajouter **22.816,60€**. Il s'agit d'un transfert en attente depuis plusieurs années, depuis le compte en ressources propres de CAHIER au Centre d'Etudes Supérieures de la Renaissance (Tours) lequel gérait le consortium durant la première labellisation de CAHIER. Ce retard s'explique d'abord par le fait que le CESR a souhaité solder toutes les factures en attente à la fin de la première labellisation (2015-2016) avant de réaliser ce transfert. Ensuite, le consortium a rencontré des difficultés dues à l'inertie des acteurs, aussi bien du côté du CESR que de la DR8. L'arrivée d'une nouvelle gestionnaire financière à la MSH Val de Loire a permis de débloquer la situation et de réaliser le versement dû et survenu, en 2020.

Ainsi, entre les difficultés énoncées, les retards accumulés et la crise sanitaire, le consortium prévoit pour la fin de l'année 2021, un nouveau reliquat de **41.945,87€** (+ 648,48€ en SE qui seront dépensées avant la fin 2021). **En conséquence, le consortium ne formule aucune demande financière pour l'année 2022.**

CAHIER souhaite utiliser son reliquat pour finaliser les actions entreprises en 2020-2021 et qui concerneront trois domaines prioritaires : la FAIRisation des données, la finalisation et la diffusion des livrables des groupes de travail et le projet post-labellisation. Les méthodes mises en œuvre pour atteindre ces objectifs sont décrites, avec précision, dans les différentes sections de ce bilan :

- la FAIRisation des données repose sur l'accompagnement personnalisé dans la préparation des données, plusieurs guides méthodologiques et l'utilisation d'une application web dédiée (voir ci-après section n°10.b)

[1] Voir programme en ligne sur : https://cahier.hypotheses.org/5329

[2] Conformément aux recommandations du conseil scientifique d'Huma-Num, il n'y a plus eu d'intégration de nouveaux membres depuis 2020.

- la finalisation et la diffusion des livrables des groupes de travail, et notamment l'article scientifique dédié au thésaurus « Typologie des textes », sa traduction en anglais, la traduction des 365 concepts du thésaurus (voir ci-après section n°4.b.2). Avec la reprise annoncée des conférences internationales, CAHIER pourra disséminer et partager ces résultats avec la communauté scientifique internationale.
- et le projet post-labellisation dont les jalons ont déjà été posés et approuvés par les membres mais qui demande dorénavant l'implication exclusive du groupe de travail « CAHIER après CAHIER » (voir ci-après section n°10.a et annexe n°7)

TGIR HUMA-NUM - Consortium "CAHIER"				
tableau type de synthèse de présentation budgétaire par nature et actions structurantes				
Bilan 2021				**Prévisions 2022**
Dans le bilan (version word), les dépenses réalisées sur l'année devront être mises en regard du budget prévisionnel que le consortium avait présenté à N-1 ainsi que du budget exécuté l'année précédente. Les différences budgétaires devront être expliquées. Les dépenses au-delà de 5 000€ doivent être détaillées dans le tableau ci-dessous.				
DEPENSES	**Exécuté 2021**			**Prévisions pour 2022**
	SE	**RP**		**RP**
Dépenses de fonctionnement	**37 851,52 €**	**25 922,50 €**		**35 000,00 €**
Subvention au projet membre (subvention accordée en 2019)		2 800,00 €		
Gouvernance du consortium : réunions, journées de travail internes	695,36 €			
Assemblée générale du consortium		10 000,00 €		
Colloque "10 ans de Cahier" (organisation et publication des actes)	15 477,20 €			
Fairisation des données (+ PGD)	5 650,00 €			6 000,00 €
Groupe "CAHIER après CAHIER"	802,34 €	122,50 €		7 000,00 €
Groupes de travail - réunions	2 916,71 €			3 000,00 €
Organisation de formation et soutien à la formation des membres du consortium (soutien aux formations ANF, TEI et EnExedy et financement de participation à des formations pour les membres du consortium)	12 309,91 €	3 000,00 €		11 000,00 €
Soutien à la participation à des conférences internationales				2 000,00 €
Soutien à la publication de corpus et à la traduction				1 000,00 €
Gestion du consortium		10 000,00 €		5 000,00 €
Dépenses de personnel non permanent	**0,00 €**	**6 546,04 €**		**7 000,00 €**
Stagiaires rémunérés sur la dotation - Fairisation des données		6 546,04 €		7 000,00 €
TOTAL DES DEPENSES	**37 851,52 €**	**32 468,54 €**		**42 000,00 €**
RESSOURCES VERSEES PAR HUMA-NUM	**2021**			**2022**
	SE	**RP**		**RP**
Dotation versée par la TGIR Huma-Num (part en SE et en RP)	**38 500,00 €**	**16 500,00 €**		**0,00 €**
Reliquats éventuels de l'année N-1		**35 097,81 €**		**41 945,87 €**
		22 816,60 €		
TOTAL DES RESSOURCES	**38 500,00 €**	**74 414,41 €**		**41 945,87 €**
RESTANT DISPONIBLE (ressources-dépenses)	**648,48 €**	**41 945,87 €**		**-54,13 €**
AUTRES RESSOURCES DU CONSORTIUM	**2021**			
	SE	**RP**		
Indiquer les autres sources de financement en les détaillant par ligne				
TOTAL DES RESSOURCES	**0,00 €**	**0,00 €**		

Bilan financier année 2021 et prévisionnel année 2022

b) Bilan des groupes de travail : année 2021

1. Groupe de travail [Data_Cahier] : FAIRisation des données du consortium et réalisation du Plan de gestion des données

Porté par F.Idmhand et I.Galleron
Trois stagiaires ont été recrutés pour accompagner les travaux de ce GT :
-Un stagiaire en informatique (2 x 6 mois) : Ala Eddine Laouir (Université Haute Alsace) chargé du développement de l'application web Mynkl connectée à l'API de Nakala
-Un stagiaire en Humanités numériques (1 x 6 mois) : Andrés Echavarria (Université de Bretagne-Sud) chargé de mettre à jour, avec les porteurs de projets, leurs métadonnées et leurs données des projets en vue du dépôt sur Nakala
Une stagiaire en documentation et bibliothèques (1 x 1 mois puis financement d'une prestation) : Laurène L'Hermite (Université de Poitiers) chargée du suivi de la rédaction des plans de gestion des données des projets
Résultats : un outil développé « Mynkl », 1 guide et 1 mémorandum, 1PGD de structure et 10 PGD pour les projets membres

1. Développement de l'application *Mynkl*[3]: http://mynakala.huma-num.fr/ (documentation : http://mynakala.huma-num.fr/docs)

Pour accompagner la FAIRisation des données du Consortium, et après l'organisation de plusieurs [FAIR_Lines] durant le printemps 2020, le travail opérationnel a été mené durant l'année 2021. Les stagiaires qui ont été recrutés disposaient de profils complémentaires : l'un, issu d'un Master 2 en informatique, a été chargé de construire l'application connectée à l'API de Nakala, et l'autre, issu d'un Master en Sciences humaines avec mineure en « Humanités Numériques » a été chargé d'organiser la révision des fichiers XML-TEI et des tableurs de données en vue du dépôt sur Nakala.

L'application *Mynkl* a été créée pour accompagner la FAIRisation des données de CAHIER[4]. Une meilleure insertion des ressources de CAHIER dans l'écosystème du Web passe par la publication des données et des métadonnées à l'intérieur du Web de données « Linked data » : le dépôt des données dans un entrepôt ouvert et sécurisé qui fournit des identifiants pérennes contribue à cette insertion. *Mynkl* participe à cette démarche en facilitant le dépôt en masse des données. Cette fonctionnalité n'a pas encore été développée par l'équipe de Nakala, elle devrait être disponible à la fin de l'hiver 2021-2022 : *Mynkl* répondait donc à une urgence : aider les projets membres à FAIRiser les données qu'ils ont constituées depuis plusieurs années.

L'application fonctionne selon un principe simple : elle assure le rôle de « passeur » en accompagnant les données depuis les dossiers de l'utilisateur jusqu'à l'entrepôt Nakala via le service que propose cet entrepôt : une API (*Application Programming Interface*). *Mynkl* ne stocke aucune information personnelle : la connexion de l'utilisateur passe par la clé API que fournit Huma-Num à toutes celles et ceux qui créent un Huma-NumId. Une fois connecté à Nakala, via l'outil *Mynkl*, l'utilisateur entre dans son environnement de travail personnel puisqu'il accède à ses propres données stockées sur Nakala. L'application assure ensuite le transfert des données, pour cela, elle réalise un *mapping* des métadonnées qui ont été préparées selon les recommandations du guide[5] (qu'il s'agisse d'un tableur ou de métadonnées XML-TEI) vers les métadonnées attendues/exigées par Nakala. Ensuite, en fonction de la vitesse de la

[3] Voir plus loin la section n°10.b qui explique les raisons pour lesquelles une application web a été développée.
[4] Voir : http://mynakala.huma-num.fr/ , application développée dans le cadre du stage de Ala Eddine LAOUIR (Master 2 en « Informatique et mobilités », Université Haute Alsace). Il est possible de se connecter en utilisant la clé api fournie par Huma-Num, voir documentation : http://mynakala.huma-num.fr/docs Le « mémo » d'utilisation de Mynkl a été déposé sur HAL le 28/10/2021 : https://halshs.archives-ouvertes.fr/halshs-03408209.
[5] Voir « Guides pour la FAIRisation des données » : V1 https://halshs.archives-ouvertes.fr/halshs-02889777 et V2 https://halshs.archives-ouvertes.fr/halshs-03037748 en ligne sur HAL

connexion internet, le dépôt des données est plus ou moins rapide.

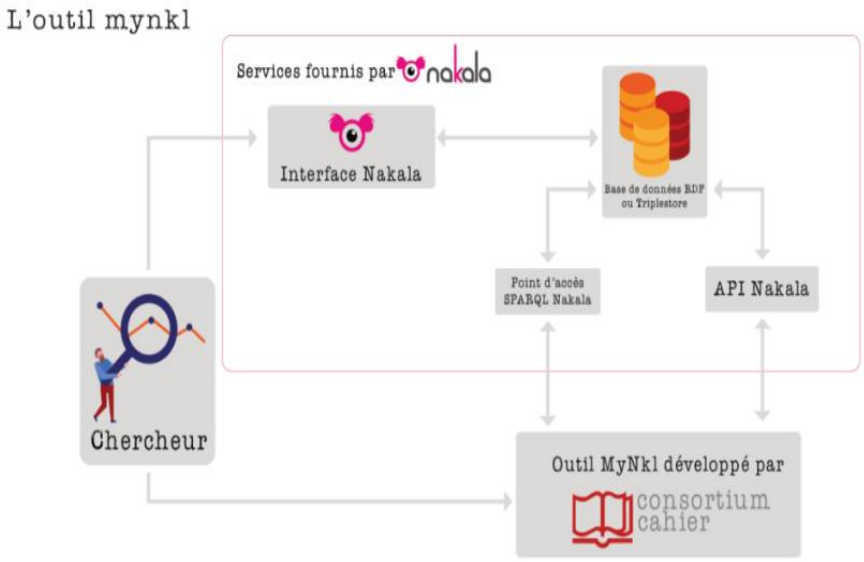


Schéma décrivant le fonctionnement de l'application *Mynkl*

Mynkl a été enrichi d'une interface minimale d'exploration des données ; elle suggère des contenus à l'utilisateur en fonction de ses habitudes de dépôt. Cette fonctionnalité s'appuie sur le filtrage de contenu (*Content based filtering*), elle examine les contenus présents dans Nakala et qui partagent des propriétés avec les objets que l'utilisateur a déposés dans l'entrepôt. L'opération est réalisée sans collecte de données personnelles et sur la base exclusive des données déposées par l'internaute : l'algorithme fouille les métadonnées, repère des éléments de similarité entre plusieurs paires d'éléments et génère une liste de recommandations à l'utilisateur. En proposant cette fonctionnalité, le groupe de travail a souhaité ouvrir des perspectives pour la recherche scientifique : le signalement de ressources (ici, dans Nakala) fait partie des perspectives prometteuses de ces entrepôts de données. Cette fonctionnalité permet de donner des exemples précis à la communauté scientifique, à la fois de l'utilité de nouveaux services comme les API et de l'intérêt à disposer de données FAIR pour leur réutilisation dans de nouvelles recherches scientifiques.

Le rapport remis par CAHIER en 2020 visait la FAIRisation d'au moins 40 projets ; à ce stade seuls 10 projets ont FAIRisé leurs données, 2 travaillent actuellement sur des collections qui restent privées à ce stade mais qui existent déjà dans Nakala et 6 d'entre eux ont utilisé l'application *Mynkl* pour déposer leurs données. A la date du 31 octobre 2021, le nombre de ressources FAIR du consortium CAHIER s'élève à 61.107 ressources dont 57.073 images et 4.034 fichiers XML-TEI[6]. Plus précisément, il s'agit de :

- 1604 images et 1793 fichiers XML-TEI disponibles sur Ortolang (1 projet a utilisé Ortolang)
- 167 images et 167 fichiers XML-TEI disponibles sur Zenodo (1 projet a utilisé Zenodo)
- et de près de 55.302 images correspondant à peu près à 16.030 fichiers de données, dont 2077 fichiers XML TEI (6 projets ont utilisé Nakala)

2. Rédaction des plans de gestion des données de CAHIER et des projets membres

La FAIRisation des données du Consortium s'est accompagnée de la rédaction du document technique décrivant le cycle de vie des données produites : le Plan de Gestion des données. Laurène L'Hermite a été recrutée par le consortium (stage puis prestation) pour accompagner la collecte des informations et assurer, avec les projets volontaires, la préparation de leurs PGD.

Le Plan de Gestion des Données du consortium CAHIER se veut à la fois un bilan des actions du consortium au bout de dix ans et une structure globale pouvant servir d'exemple et de guide de recommandation pour tous les projets dédiés aux Corpus d'auteurs. Ainsi, CAHIER a rédigé deux types de

[6] Voir plus loin la section n°13 a). Ces calculs sont basés sur les entrepôts de 10 projets (cellules grisées)

plans de gestion des données : un Plan dit « de structure » (voir Annexe n°3a) et un plan de gestion des données par projet volontaire[7].

Les plans proposés par CAHIER s'inspirent des recommandations publiées en 2016 par le programme Horizon Europe[8], du modèle fourni par l'Agence Nationale de la Recherche et des modèles mis en ligne par la plateforme DMP OPIDoR (https://opidor.fr/), l'outil d'aide à la création en ligne de plans de gestion de données (Data Management Plan ou DMP). Après avoir étudié ces exemples, le groupe de travail a rédigé un modèle adapté aux besoins de la communauté scientifique du Consortium CAHIER et des projets scientifiques dédiés aux corpus d'auteurs. Celui-ci sera mis en ligne sur Opidor durant l'automne 2021. Le plan de gestion des données du consortium est joint à ce bilan en annexe n°3a et peut être consulté en ligne sur HAL : https://halshs.archives-ouvertes.fr/halshs-03409421.

2. Groupe de travail [Typologie des textes]

Porté par M.-L. Demonet
Résultats : un thésaurus de 365 concepts, 1 guide et 1 article scientifique en cours de finalisation

Créé en novembre 2019, le groupe de travail s'est donné pour objectif de développer une ontologie des typologies de textes à l'aide de l'outil OpenTheso (développé par Miled Rousset, https://github.com/miledrousset/opentheso). Un thésaurus de cette nature (typologie textuelle) contribue à faire avancer les principes FAIR de CAHIER en offrant une ressource accessible pour l'indexation des données textuelles : chaque concept décrit est associé à un identifiant stable. Plusieurs réunions, plénières ou partielles, la plupart en visioconférence, ainsi que deux ateliers en présentiel ont permis d'effectuer d'importants remaniements et de proposer des avancées significatives :

1) la hiérarchie des hypergenres ou micro-thésaurus a été entièrement redistribuée en neuf groupes, chacun sur trois niveaux au maximum, afin de bien identifier les nœuds, les classes et les propriétés ;
2) de nouveaux concepts ont été ajoutés avec leurs définitions ; leur nombre est actuellement de 365 ; les identifiants handle fournis par Huma-Num ont été intégrés au thésaurus (mai 2021), ils sont accessibles sur le serveur (https://opentheso.huma-num.fr/opentheso/api/theso/Typologie) et déjà moissonnés par Nakala ;
3) un article « Décrire un corpus d'auteurs : un thésaurus pour les types de textes », exposant les étapes de réalisation et les principes théoriques tels qu'ils ont été élaborés pendant quatre ans est en cours de rédaction ;
4) une version 1 du guide de bonnes pratiques décrivant les étapes d'intégration des concepts à différents types de corpus (XML/TEI, ou Dublin Core sous Omeka par exemple) a été élaborée et mise à la disposition de la communauté sur HAL : https://halshs.archives-ouvertes.fr/halshs-03402679;
5) plusieurs tests ont été réalisés depuis mai pour intégrer les concepts avec leur identifiant handle dans les en-têtes TEI ou DC, avec démonstration et exemples lors du colloque de Bordeaux (juin 2021).

L'année 2022 prévoit 4 à 6 réunions pour l'article, une version 2 du guide et d'améliorer l'intégration des concepts dans les métadonnées en fonction des projets. Le thésaurus sera présenté lors de colloques et pourra être enrichi de nouveaux concepts.

3. Groupe de travail [(Ré)utilisabilité]

Porté par Anne Garcia Fernandez, Elisabeth Greslou et R. Walter
Résultats : rédaction d'un guide en cours

Le groupe de travail (Ré)utilisabilité a été créé lors de l'AG CAHIER du 27 novembre 2020.
Il est animé par Anne Garcia-Fernandez (CNRS, UMR Litt&Arts, Grenoble), Elisabeth Greslou

[7] Voir exemples en Annexe n°3b, 10 projets ont rédigé un Plan de gestion des données et 3 autres plans sont en cours de rédaction
[8] H2020 Programme « Guidelines on FAIR Data Management in Horizon 2020 », https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf#page=10

(UGA, UMR Litt&Arts, Grenoble) et Richard Walter (CNRS, UMR THALIM, Paris) et a réuni une vingtaine de personne autour de la problématique du R de FAIR.
Objectif : « Mener une réflexion sur les actions et outils permettant que les **données**, au-delà de leur seule exposition, soient concrètement **(ré)utilisables** et **(ré)utilisées** ».
Livrable : trois réunions et un atelier de trois jours en présentiel ont permis la rédaction d'un vade-mecum qui sera présenté lors de l'AG du consortium CAHIER en novembre 2021.
Wiki du groupe de travail : https://groupes.renater.fr/wiki/gt-re-utilisabilite-cahier/index

5. Résumé des objectifs et du programme scientifique du consortium

Lors de sa création en juillet-août 2011, le consortium CAHIER s'était d'abord donné pour objectif la construction d'une « fédération de projets existants ou projetés en France » portant sur les « corpus d'auteurs » et ayant pour activité et/ou but, une édition numérique (accompagnée ou non d'une édition papier). Le consortium CAHIER visait à réunir au sein d'une organisation transversale interdisciplinaire, des chercheurs porteurs de projets scientifiques numériques en vue de partager des expériences, des pratiques, des problèmes (et des solutions), et des données sur les corpus textuels. Le programme scientifique de CAHIER n'a pas changé entre ses deux labellisations mais il s'est étoffé au fil des ans comme en attestent les différents bilans remis entre 2012 et 2020[9]. Ainsi, **depuis 2011, le consortium a interrogé la chaîne de « conversion numérique » des textes** (pour reprendre l'expression de Milad Doueihi[10]) et ses travaux ont suivi les trois axes de développement suivants :

- **AXE 1 : La construction de données « interopérables »** : ce concept a récemment été supplanté par celui de données « FAIR » mais depuis 2011, l'objectif de CAHIER est d'acculturer la communauté scientifique à l'utilisation de standards et de normes ouvertes facilitant l'interopérabilité et donc, la réutilisation des données (B.A12 et B.A13). Pour atteindre cet objectif, CAHIER a fait porter un effort significatif sur l'acquisition de données de qualité (image et texte), proposé le partage de normes de transcription définies selon des objectifs éditoriaux clairement énoncés et il a réfléchi à l'indexation des données à travers l'élaboration de métadonnées compatibles avec les standards de catalogage, d'archivage, d'identification et de protection des données. Enfin, le consortium a également testé des dispositifs d'affichage « du mode texte et du mode image » (B.A12 et B.A13) de façon à opérer des choix pertinents en fonction des publics mais également de la qualité des données produites.
- **AXE 2 : L'approche critique des pratiques de publication numérique de corpus d'auteurs** : le consortium s'était également donné comme objectif de travailler sur l'évaluation des pratiques d'édition numérique (B.A13 et B.A14), en fonction des projets. Il s'agissait de discuter les critères (modes d'accès, volumétrie, outils utilisés, adéquation aux objectifs du consortium) et choix en vue d'harmoniser des pratiques. Enfin, le consortium a également interrogé les conditions de réalisation des éditions numériques, notamment les conditions juridiques, de façon à travailler sur les éventuels verrous et tenter de les lever.
- **AXE 3 : Les outils numériques des sciences des textes** : le consortium s'était donné pour but de tester les prototypes qui « circulent dans le domaine des humanités numériques » (B.A14) afin d'en mesurer l'efficacité pour la construction et/ou l'exploitation de corpus numériques, de former aux outils repérés ou utilisés par la communauté scientifique afin d'étendre son panel méthodologique et de soutenir, le cas échéant, des développements informatiques.

6. État comparatif entre la programmation initiale et les actions réalisées en fin de cycle 4+4+2

En lien avec les objectifs énoncés ci-dessus, le consortium CAHIER a connu trois grandes phases qui correspondent aux trois temps du cycle 4+4+2 :

[9] Les bilans seront cités sous la forme suivante : « B.A12 » = bilan de l'année 2012. « B.A13 » = Bilan de l'année 2013, etc.
[10] Doueihi, Milad, La grande conversion numérique, Paris, Seuils La Librairie du XXIe siècle, 2008.

- Cycle 1 : 2011-2015. Construction, ouverture et consolidation du réseau
- Cycle 2 : 2015-2019. Production de livrables par le réseau
- Cycle 3 : 2019-2021. Harmonisation des travaux et FAIRisation

Le consortium avait pour ambition de faire évoluer et de consolider les formats de l'édition numérique (dite « brute, critique, génétique, encyclopédique » cf. B.A12 et B.A15) et de favoriser les échanges de données et de méthodes en soutenant la rétroconversion ou la migration vers de nouveaux formats de corpus anciens, dispersés, déjà édités ou non. Ces objectifs ont été nuancés durant le « second cycle de vie » du consortium et révisés pour la seconde labellisation. En effet, avec l'apparition de nouveaux acteurs de la numérisation entre 2011 et 2015 (les MSH, qui ont ouvert des appels à projets de numérisation, le programme spécifique de CollEx-Persée ou l'ANR), CAHIER a reconsidéré son rôle de soutien à la numérisation. L'aide de 3000€ a été repensée afin de soutenir l'amorçage ou la réalisation d'une tâche liée à un projet numérique sur un corpus d'auteurs plutôt que des tâches de numérisation. Par contre, l'extension du réseau est devenue un objectif important du second cycle, car elle est apparue comme le moyen de favoriser l'échange d'expériences. De même, le second cycle témoigne d'une plus grande implication dans l'Informatisation, l'Edition et surtout la Recherche scientifique sur les corpus numériques créés (cf. D.15)[11].

Pour accroître son audience, le consortium a davantage misé sur la **formation** des chercheurs, jeunes chercheurs et ingénieurs à de nouvelles compétences en soutenant la participation de ses membres à des formations aux outils et méthodes numériques notamment. CAHIER a créé des formations « à la demande » pour ses membres (atelier thématique annuel, formations spécifiques) et encouragé sa communauté à échanger avec la communauté scientifique internationale en soutenant la présentation de travaux lors de colloques internationaux : ADHO et TEI notamment.

Les groupes de travail ont connu un important dynamisme durant la seconde labellisation : ils se sont chargés de traiter des questions de recherche précises concernant les (ou des) corpus d'auteurs et/ou de lever des verrous en un temps limité. Ils ont régulièrement été composés de six à dix personnes en moyenne, membres ou non de CAHIER et ont tous produit au moins un livrable sous forme de guide, d'articles scientifiques ou d'outils. Ils ont également accompagné la démonstration et le transfert des connaissances et compétences acquises, notamment lors du cycle 3 du consortium, en valorisant les productions et les résultats obtenus par les membres.

Enfin, la création d'un conseil scientifique international en 2018 propre à CAHIER a contribué à promouvoir les activités du consortium au-delà des frontières nationales tout en lui permettant de bénéficier des conseils d'experts internationaux reconnus dans le domaine des Humanités numériques. Depuis 2018, le conseil scientifique est composé de Bertrand Jouve (CNRS, qui a quitté le CS en 2019 en raison d'autres engagements professionnels), de Susan Schreibman (Maastricht University), de Lou Burnard (Oxford University) et de Elisabeth Burr (Lepizig University).

Le tableau synoptique proposé ci-dessous permet d'apprécier l'évolution entre la programmation initiale proposée pour le cycle 1 et la réalisation des objectifs durant les cycles 2 et 3. La liste exhaustive des livrables est présentée de façon détaillée plus loin dans le dossier en section n°12, elle fait état, pour l'ensemble des trois cycles, 2011-2021, de :

- **AXE 1 : La construction de données « interopérables »**
  - 6 publications de type guides méthodologiques
  - 61.107 ressources FAIR dont 57.073 images et 4.034 fichiers XML-TEI[12] :
    - 1604 images et 1793 fichiers XML-TEI disponibles sur Ortolang (1 projet a utilisé Ortolang)
    - 167 images et 167 fichiers XML-TEI disponibles sur Zenodo (1 projet a utilisé Zenodo)
    - ~55.302 images et ~16.030 fichiers de données à ce jour dont 2077 fichiers XML TEI (6 projets ont utilisé Nakala)

439 fichiers XML-TEI sont, par ailleurs, disponibles sur github (2 projets ont utilisé github)

  - 3 083 résultats (au 26/10/21) signalées dans Isidore

[11] Les dossiers de demandes de labellisation seront nommés « D.15 » = dossier labellisation de l'année 2015, « D.12 » = dossier de labellisation de l'année 2012, etc.

[12] Voir plus loin la section 13 a). Ces calculs sont basés sur les entrepôts des 10 projets (cellules grisées)

- 3 outils de communication à destination de la communauté (1 blog, 2 comptes sur les réseaux sociaux, 2 listes de diffusion mail)

- **AXE 2 : L'approche critique des pratiques de publication numérique de corpus d'auteurs**
  - 1 colloque scientifique
  - 23 participations à des colloques scientifiques
  - 1 ouvrage collectif
  - 36 publications recensées sur HAL

- **AXE 3 : Les outils numériques des sciences des textes**
  - 14 formations ou écoles thématiques
  - 3 outils numériques développés pour la communauté scientifique de CAHIER et au-delà
  - soutien au développement d'au moins 3 outils développés par d'autres communautés scientifiques

Etape du cycle	Programmation initiale	Actions réalisées en fin de cycle	Livrables (liste exhaustive et détails ci-après en point n°13)
Cycle 1 : 2011-2015 Construction, ouverture et consolidation du réseau	AXE 1 Construction de données « interopérables »	La fédération de projets constituée en 2011 associe à ce stade des projets scientifiques numériques mais pas encore de données. L'échange de méthodologies reste la priorité, la création de groupes de travail a pour but de répondre aux problèmes (création du groupe de travail (GT) « Questions juridiques » en 2012)	❖ 1 GT « Questions juridiques »
	AXE 2 Approche critique des pratiques de publication numérique de corpus d'auteurs	Le réseau s'étend par l'intégration progressive de nouveaux projets. **Le nombre de projets membres passe successivement de 12 (B.A11) à 23 (B.A12), puis à 29 (B.A14).** Le réseau compte en 2014-2015 : 103 chercheurs impliqués (PU, MCF, CR et DR CNRS), 54 ITA et 2 conservateurs de bibliothèque. Lors du dépôt du dossier de labellisation en 2015, le réseau compte **42 membres** et couvre une large partie du territoire national. La liste exhaustive des membres du consortium est présentée plus loin en section n°13a) et 13 c). Elle présente, pour chaque projet, les personnes impliquées.	❖ 42 membres (voir carte du réseau depuis 2014 en Annexe n°1)
	AXE 3 Outils numériques des sciences des textes	Création des ateliers annuels du Consortium CAHIER : ces formations organisées sur plusieurs jours et délivrées par des experts du domaine permettent de travailler sur un verrou scientifique et/ou technique à l'aide d'un ou plusieurs outils ou langages informatiques. En 2014 : 3 ateliers annuels avaient été organisés	❖ 3 ateliers thématiques ❖ 1 soutien aux outils (Philologic)
Cycle 2 : 2015-2019 Production de livrables par le réseau	AXE 1 Construction de données « interopérables »	En lien avec la priorité du D.15 « **Informatisation** » La création de groupes de travail portant sur la correspondance, la granularité du moissonnage des données, la typologie des métadonnées	❖ 2 GT « Questions juridiques » et « Correspondances » et 2 guides publiés ❖ 3 ateliers annuels soutenus ❖ 3 soutiens à des formations (ANF, etc.)
	AXE 2 Approche critique des pratiques de publication numérique de corpus d'auteurs	En lien avec les priorités du D.15 « **Edition** » et « **Recherche** » L'édition ouverte des sources de la recherche (corpus, textes, travaux) est la priorité du Consortium. Le terme édition doit être entendu à la fois comme édition de sources en libre accès mais également dans sa dimension érudite. La recherche scientifique au sein du consortium concerne les liens entre informatisation et édition en vue de construire de nouvelles connaissances et de nouveaux savoirs. Différents groupes de travail ont réfléchi sur de nouvelles thématiques (comme le crowdsourcing), sur l'édition numérique et le cas des correspondances. La valeur ajoutée évidente a été la création de	❖ 65 projets membres (voir carte du réseau depuis 2014 en Annexe n°1) ❖ 1 GT « Event » et 1 guide publié ❖ 1 GT « R2CAHIER » et 2 articles scientifiques publiés ❖ 3 articles scientifiques (cf. liens HAL) ❖ Création du conseil scientifique international de CAHIER

		nouveaux concepts : les différents types de publications numériques, le crowdreading, la description des genres textuels. La possibilité de partager ces résultats scientifiques a été maintenue par la mise en place d'ateliers et de groupes de travail ouverts, au-delà des seuls membres de CAHIER. CAHIER s'est également doté, durant cette période, d'un conseil scientifique composé de quatre membres chargés de formuler des recommandations à CAHIER.	
	AXE 3 Outils numériques des sciences des textes	En lien avec la priorité du D.15 « **Informatisation** », et pour poursuivre la formation et l'accompagnement de projets en gestation sur la partie informatisation des données, CAHIER a continué à organiser et soutenir des formations techniques. Le but du consortium était d'avoir un impact structurant majeur.	❖ 1 GT « Data_Cahier » et 1 outil livré et finalisé : WebOai ❖ 14 ateliers thématiques ❖ 3 formations spécifiques ❖ 2 soutiens aux outils (TXM et Philologic)
Cycle 3 : 2019-2021 Harmonisation des travaux et FAIRisation	AXE 1 Construction de données « interopérables »	Le travail de facilitateur du consortium s'est poursuivi en 2020 et 2021 conformément aux intentions déclarées lors de la demande de renouvellement de 2015 et aux recommandations formulées par le CS d'Huma-Num mais, cette fois, la priorité a été donnée à l'évaluation de l'état des données du consortium et à l'accompagnement de la FAIRisation de celles-ci. Le groupe de travail « Data_Cahier » a réalisé l'étude de terrain, l'analyse des données de CAHIER et des choix opérés par les projets et identifiés ainsi que les barrières (cf. https://halshs.archives-ouvertes.fr/halshs-03224294). Au rythme des mises à jour de l'outil Nakala dont la refonte a été réalisée par les équipes d'Huma-Num entre 2019 et 2020, le GT a préparé la FAIRisation des données du consortium grâce à la rédaction d'un « Guide pour la FAIRisation des données » et en levant le verrou du dépôt massif des données par le développement d'une application web *Mynkl* (voir extraits du mémorandum en Annexe n°2) Le consortium a également accompagné la rédaction d'un plan de gestion des données (PGD) pour tous les projets membres qui ont accepté d'être accompagnés dans cette tâche et préparé son propre plan de gestion des données (cf. Annexes n°3a et n°3b)	❖ 1 GT « Data_Cahier » ❖ 1 outil développé « *Mynkl* » ❖ 1 mémorandum publié ❖ 1 PGD modèle et 10 PGD publiés ❖ 1 video youtube : https://www.youtube.com/watch?v=hygpiLsCJMY ❖ Sur Nakala : ~55.302 images et ~16.030 fichiers de données à ce jour dont 2077 fichiers XML TEI ❖ 2 soutiens à des formations (ANF, etc.)
	AXE 2 Approche critique des pratiques de publication numérique de	Après dix années de travaux, le consortium CAHIER a souhaité réunir ses membres pour venir présenter les corpus d'auteurs qu'ils ont construits et détailler les différents types d'exploitation menées sur ces corpus: analyses littéraires, linguistiques, poétiques ou génétiques assistées par ordinateur ainsi que les difficultés et les défis de l'édition numérique, et plus	❖ Composition finale du réseau : 58 membres actifs en 2021 (voir carte du réseau depuis 2014 en Annexe n°1) ❖ 1 colloque scientifique ❖ 1 ouvrage collectif

	corpus d'auteurs	particulièrement de l'annotation en XML/TEI. Ce colloque a réuni 45 personnes en présentiel et une moyenne de 25 personnes en visio. A l'issue de cette rencontre, un ouvrage collectif a été préparé, il est actuellement sous presse chez l'éditeur EAC : Editions des archives contemporaines (https://eac.ac/) L'intensification de l'action internationale s'est poursuivie malgré le confinement par une collaboration virtuelle aux rencontres internationales : DH et TEI, ainsi que par la collaboration avec DARIAH, Humanistica et Huma-Num dans le cadre d'EOSC Pillar (cf. vidéo « FAIRification of data in mass »	❖ 1 article scientifique (cf. liens HAL) ❖ 1 conseil scientifique international composé de 3 membres au sein de CAHIER
	AXE 3 Outils numériques des sciences des textes	Le consortium est resté structuré autour du langage de balisage XML-TEI et de son exploitation via XSLT et xQuery. Le consortium utilise également la norme Dublin Core pour décrire ses ressources numériques. Ces standards sont bien partagés au niveau national et international. Un nouveau verrou a été identifié par CAHIER, celui de la description des genres dans les headers TEI et dans les mots clés (Dublin Core par exemple). Pour le lever, CAHIER a construit un thésaurus complet décrivant 365 concepts et genres textuels. Ce thésaurus est en ligne sur OpenTheso et chaque concept est pourvu d'un *handle* : il s'agit d'une plus-value importante en vue du web sémantique.	❖ 1 GT « Typologie des genres textuels » et 1 guide déposé sur HAL « Décrire les textes dans le cadre d'une édition numérique. Le thésaurus "Typologie textuelle" du Consortium CAHIER » ❖ 1 atelier thématique ❖ 1 formation spécifique

7. Difficultés rencontrées et éventuels changements stratégiques par rapport au programme scientifique initial

Si ce n'est la reconfiguration l'aide de 3000€ que le consortium a pu apporter à ses projets membres, aucun changement stratégique par rapport au programme scientifique initial n'a été opéré. Toutefois, durant sa seconde labellisation, CAHIER a accentué ses efforts en vue de produire davantage de littérature grise (travaux de son « AXE 2 : L'approche critique des pratiques de publication numérique de corpus d'auteurs ») et de lever plusieurs verrous concernant la FAIRisation de ses données[13] sans que cela ne conduise à des changements stratégiques majeurs puisque ces actions ont été portées par des groupes de travail.

On peut toutefois signaler une difficulté d'ordre épistémologique qui n'avait pas été identifiée par les membres du consortium en amont mais qui ressort à ce stade du bilan. Elle concerne les différents domaines scientifiques associés dans le consortium : la philologie d'une part, la critique littéraire de l'autre et la génétique des œuvres en dernier lieu. Tous trois partagent un même domaine scientifique, un même objet, et de nombreux questionnements sur le texte et l'édition numérique (écrits, corpus, auctorialités, etc.). Toutefois, leurs visées et finalités scientifiques sont différentes et, en conséquence, leurs projets et résultats numériques sont eux aussi différents : la philologie vise l'édition et, le plus souvent, l'établissement d'une version stable du texte, la critique littéraire vise l'interprétation du sens du texte tandis que la génétique des œuvres combine les deux objectifs avec un intérêt particulier pour le processus créatif et donc l'histoire du texte. Ces trois visées ont amené les projets à opter pour différents types de publications numériques et à accorder plus ou moins d'intérêt à certaines métadonnées descriptives, à certaines technologies plutôt qu'à d'autres et à certains langages de balisages, plutôt qu'à d'autres. Cette disparité apparaît aujourd'hui à la fois comme une richesse et parfois un frein à la réutilisation des corpus. Elle demande la mise en place de réflexions spécifiques, que seule une communauté mature, comme celle de CAHIER, peut mener, en s'appuyant sur la structuration et l'expérience déjà acquise.

Pour observer, comprendre et décrire les résultats numériques qu'il a produits, le consortium s'est doté d'outils, et notamment d'un guide d'analyse en novembre 2018[14] : « Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations ». Par rapport aux catégories décrites dans ce guide, on constate que les membres du CAHIER ont principalement produit deux types de publications numériques scientifiques :

- **32 projets d'archives éditorialisées (AE[15])** : il s'agit de projets offrant une scénarisation minimale des ressources en lien avec la question de recherche. L'éditorialisation est donc orientée par la question de recherche et se reflète dans le choix des documents, dans les médiations choisies, dans les métadonnées et dans les éventuels discours d'accompagnement : c'est le cheminement scientifique qui organise le « vrac » des documents. On constate que la plupart des projets qui ont opté pour ce type de publications s'inscrivaient dans les domaines de la génétique des œuvres et de la critique textuelle.
- **26 projets d'éditions enrichies (EE)** : il s'agit de publications qui proposent un texte profondément enrichi d'informations documentaires et contextuelles. Celui-ci peut avoir été préparé de façon à permettre un affichage selon différents critères éditoriaux, l'interrogation par facettes et l'exploitation étendue des données. Ce type d'édition suit nécessairement des pratiques harmonisées et se réfère à des standards soutenus par de larges communautés internationales, comme celle du consortium TEI. On constate qu'au sein de CAHIER, les projets ayant opté pour ce type d'éditions s'inscrivaient dans le domaine de la philologie numérique.
- Enfin, et toujours en lien avec les critères établis par le guide « Les publications numériques de corpus

[13] Voir ci-après section n°10. « Les trois principales difficultés dans l'activité du consortium sur la période 4+4+2 »

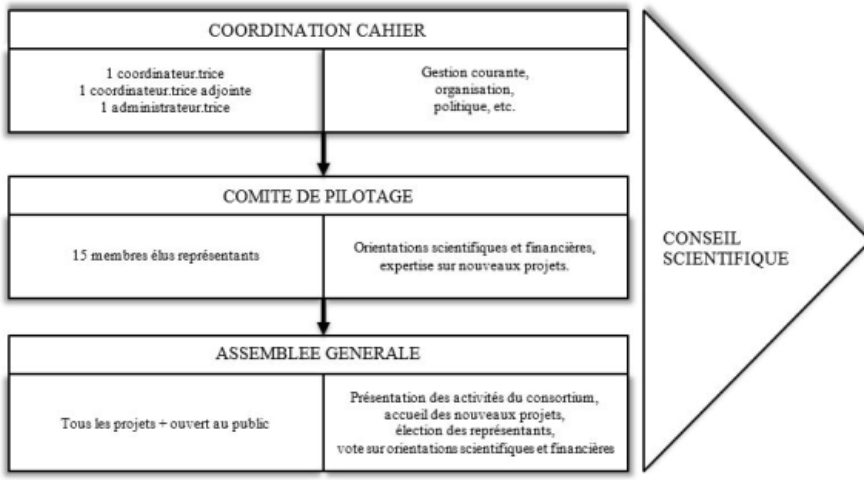
[14] Ioana Galleron, Marie-Luce Demonet, Cécile Meynard, Idmhand Fatiha, Elena Pierazzo, et al.. « Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations » (V1-Novembre 2018). [Rapport de recherche] Huma-Num. 2018, 19 p. (halshs-01932519) https://halshs.archives-ouvertes.fr/halshs-01932519

[15] Voir plus loin en section n°13 a) la liste des projets.

d'auteurs - Guide de travail, grille d'analyse et recommandations », on constate qu'**aucune édition de lecture (« *reading edition* » EL)** n'a été proposée. Il s'agit de publications, relues, corrigées et nettoyées des erreurs humaines (ou de celles qui sont issues de la reconnaissance automatique de caractères) qui vont au-delà de la mise à disposition du texte car elles l'accompagnent de toute une série d'éléments qui le mettent en perspective. Ainsi par exemple, il peut s'agir d'une édition en XML/TEI qui se limite à l'encodage des grandes structurations du texte et qui fournit des informations de base, sans toutefois avoir recours à toute une série de balises optionnelles, permettant une manipulation poussée du texte.
Alors que l'édition de lecture n'a été portée par aucun projet membre de CAHIER, elle apparaît pourtant, à ce stade des travaux de CAHIER, comme une option pertinente qui favorise l'exploitation poussée des corpus à l'aide des outils de l'informatique et facilite la réutilisabilité des éditions numériques[16]. Cette édition pourra être recommandée, voire proposée « par défaut » dans le prochain projet de CAHIER en temps qu'« état » liminaire du projet éditorial. Elle permettrait d'augmenter significativement la quantité de données disponibles.

8. Modifications éventuelles dans l'organisation du pilotage du consortium

La gouvernance et le pilotage du consortium CAHIER n'ont pas été modifiés durant les second et troisième cycle de CAHIER. Suite à la démission du coordinateur du consortium, Thomas Lebarbé (Université Grenoble Alpes), survenue le 09/09/2021, c'est la coordinatrice-adjointe Fatiha Idmhand (Université de Poitiers) a poursuivi le travail de coordination. Le comité de pilotage de CAHIER associe toujours, chercheurs, enseignants-chercheurs et ingénieurs.



https://cahier.hypotheses.org/gouvernance

- **Coordination du consortium**

Fatiha Idmhand (Professeur, Université de Poitiers)
Laurence Rageot (Ingénieur de recherche CNRS, MSH Tours Val de Loire)

- **Comité de pilotage**

Céline Bonhert (Maître de conférences, Université de Reims)
Pierre-Yves Buard (Ingénieur de recherche, MSH Caen)
Marie-Luce Demonet (membre honoraire) (Professeur émérite, Université de Tours)

[16] Galleron I., Idmhand F., « De l'interopérabilité à la réutilisabilité des éditions électroniques », *Humanités numériques* [En ligne], 1 | 2020, mis en ligne le 01 janvier 2020. URL : http://journals.openedition.org/revuehn/350 ; DOI : https://doi.org/10.4000/revuehn.350 ; https://isidore.science/document/10670/1.j76gme

Stéphanie Dord-Crouslé (Chargée de recherche CNRS, IRHIM Lyon)
Ioana Galleron (Professeur, Université Sorbonne-Nouvelle)
Elisabeth Greslou (Ingénieur de recherche, UMR-LITT&ARTS, Grenoble)
Fatiha Idmhand (Professeur, Université de Poitiers)
Chiara Lastraioli (Professeur, Université de Tours)
Alexey Lavrentev (Ingénieur de recherche CNRS, IRHIM Lyon)
Giancarlo Luxardo (Ingénieur d'études CNRS, UMR-PRAXILING)
Jean-Sébastien Macke (Ingénieur d'études CNRS, UMR-ITEM)
Cécile Meynard (Professeur, Université d'Angers)
Wioletta Miskiewicz (Chargée de recherches CNRS, UMR-AHP)
Emmanuelle Morlock (Ingénieur de recherche CNRS, UMR-HiSoMA)
Anne Réach-Ngô (Maître de conférences HDR, Université de Mulhouse)
Richard Walter (Ingénieur de recherche CNRS, UMR-THALIM)
Geoffrey Williams (Professeur, Université de Bretagne-Sud)

- **Conseil Scientifique**

Lou Burnard (Oxford University)
Elisabeth Burr (Leipzig university)
Susan Schreibman (Maastricht University)

Depuis l'Assemblée générale de novembre 2019, toutes les réunions de gouvernance et les activités du consortium ont été virtuelles en raison de la crise sanitaire. Une seule réunion du comité de pilotage a pu être organisée en présentiel : celle du lundi 7 juin 2021, veille de l'ouverture du colloque « Dix ans avec CAHIER » qui s'est tenu en mode « hybride ». La prochaine réunion du Comité de pilotage et l'Assemblée générale auront lieu les 24, 25 et 26 novembre 2021 à Paris. Durant ces deux dernières années, les réunions du comité de pilotage ont principalement concerné la gestion des activités courantes du consortium régulièrement mises en difficulté, en retard, annulées ou reportées en raison de la crise sanitaire.

Le comité de pilotage a engagé la réflexion attendue par le Conseil scientifique sur son avenir post-labellisation. Elle a été source de nombreuses réunions et de difficultés décrites ci-après section n°10. Toutefois, elles ont débouché sur un cadrage et sur un projet qui sera au cœur des travaux de l'année 2022.

9. Les trois réussites majeures dans l'activité du consortium sur la période 4+4+2

a) La plus-value de l'organisation en consortium

Indéniablement, l'une des principales réussites du consortium est **la création d'un réseau scientifique reconnu sur les corpus d'auteurs numérisés**. Ancré dans le paysage scientifique, il a développé une expertise dans l'accompagnement des projets sur les corpus d'auteurs, dans la conception de données ouvertes et dans la formation et la promotion des Humanités Numériques et des bonnes pratiques du domaine. La croissance du consortium témoigne de son dynamisme : fondé par 12 projets de recherche et regroupant une vingtaine de membres, il a multiplié par six sa taille et son réseau et est devenu un acteur et un interlocuteur incontournable des sciences humaines numériques en France. Préserver ce réseau est une priorité du nouveau projet de CAHIER.

b) Des formations attractives

Une seconde réussite du consortium est **l'attractivité de ses formations. L'atelier annuel**, école thématique récurrente, est l'action de formation qui rencontre le plus de succès : en dix ans, elle a formé près de 200 participants aux méthodes et technologies les plus récentes sans tomber dans la « routine » puisque l'atelier adapte tous les ans son programme à l'actualité des techniques, des méthodes et des concepts. CAHIER a également organisé des **formations spécifiques ciblées** pour répondre à des besoins précis : 46

personnes en ont bénéficié et ont pu se former aux langages de manipulation de XML comme XSLT ou XQuery et ont été initiés à la programmation en Python. Enfin, CAHIER a également soutenu des formations organisées par des partenaires auxquelles ont participé près de 340 personnes ; l'une d'elles est également la plus ancienne : la formation à XML-TEI de Tours. Elle comporte dorénavant un niveau 1 et un niveau 2 et est l'une des plus reconnues du pays, l'encodage en XML-TEI étant la méthode de référence pour la réalisation d'éditions philologiques numériques.

Exceptionnellement, pour sa dernière année d'existence, le consortium CAHIER n'a pas organisé d'atelier annuel au printemps 2021 mais a proposé à l'ensemble des projets adhérents (mais aussi des représentants d'autres projets et initiatives numériques) de se retrouver dans le cadre d'un **colloque organisé à Bordeaux « Dix ans avec CAHIER »**, en présentiel, du 7 au 10 juin 2021 (https://cahier10.sciencesconf.org/program) afin de faire un état des lieux des nouveaux savoirs produits dans le domaine des sciences du texte grâce aux corpus numériques et aux bases de données. Plutôt qu'un bilan du consortium, cet espace de dialogue entre les différents spécialistes des corpus d'auteurs avait pour but de faire le point sur les résultats de l'exploration des ressources numériques après leur constitution, gestion, publication, pérennisation, ou après la création de nouvelles applications ou plateformes.

Deux axes de réflexion ont été proposés : 1) Nouveaux regards sur l'histoire littéraire ou l'histoire des idées et 2) Linguistique, poétique et génétique numériques ; le colloque a accueilli presqu'autant de contributions dans l'un et l'autre. L'ouvrage collectif préparé à l'issue de cette rencontre[17] rend compte de la façon dont de nouvelles idées et perspectives ont pu émerger des activités d'informatisation et d'édition réalisées depuis dix ans, au sein de CAHIER bien sûr, mais surtout au sein des sciences des textes. Il souligne de quelles façons les pratiques ont évolué en même temps que les progrès de la technologie : on y trouve aussi bien des réflexions sur l'apport du numérique à l'entreprise d'édition, que des questionnements sur les façons de faire et les nouvelles difficultés liées au changement de médium. Ces nouvelles idées et problématiques sont autant de pistes de travail pour la rédaction du projet « CAHIER après CAHIER », que de perspectives en vue de la réutilisation des données et du passage dit « à l'échelle ». Elles ont révélé que la création de corpus n'est plus le seul défi des sciences des textes, la notion même de « corpus » en milieu numérique est en pleine mutation avec la possibilité de créer des liens entre des objets, entre différents témoins d'un même texte, entre textes différents, entre textes d'un même auteur ou d'auteurs distincts, entre états génétiques, entre texte de base et variantes, entre texte cité et texte citant, auteur(s) citant(s), entre auteur(s) source, compilateurs, scripteurs, éditeurs, etc. Au-delà des problèmes techniques que peut poser la mise en lien de ces objets, le « corpus » évolue et est plus ouvert que jamais. C'est pourquoi cette question a été identifiée comme l'un des axes du prochain projet de CAHIER[18]. De même, les discussions du colloque ont encore une fois débattu des obstacles juridiques que rencontrent les chercheurs qui travaillent sur les écrivains contemporains et, de façon plus générale, sur la circulation numérique de la mémoire contemporaine et les traces écrites. Ceci constituera également un axe de travail commun aux sciences humaines, historiques et juridiques, sur lequel veut se pencher le nouveau consortium.

[17] Livre actuellement sous presse chez l'éditeur Editions Archives Contemporaines. Voir introduction et sommaire du livre en Annexe n°4.
[18] Voir plus loin Annexe n°7.

Affiche du colloque « Dix ans avec CAHIER »

c) Des groupes de travail dynamiques

Enfin, une troisième réussite du consortium est **le dynamisme de ses groupes de travail.** Durant les différents cycles du consortium, six groupes de travail rassemblant entre quatre et dix personnes ont été créés et ont produit des livrables remarqués. Ils ont eu pour but de répondre à une question précise en un temps limité, et chacun des groupes constitués s'est donné pour objectif de lever le verrou identifié. Parfois, la rédaction d'un guide a suffi, comme dans le cas des groupes « Questions juridiques », « Correspondances », « Event » ou « (Ré)utilisabilité », pour d'autres, il a été nécessaire de produire, en sus, des articles scientifiques en vue de consolider un concept (comme celui de « crowdreading »), de présenter une approche (« Typologies des textes ») ou de questionner certains défis (comme « la réutilisabilité » ou les « textes en données »). A d'autres reprises, il a été nécessaire de développer tout un vocabulaire (référentiel) ou de construire un prototype (comme le « Thésaurus Typologie » ou l'application web « *Mynkl* »). Dans tous les cas, les livrables des groupes de travail témoignent à la fois de leur productivité et de leur efficacité au sein de CAHIER ; au total, six guides méthodologiques ont été rédigés et un septième, celui du groupe (Ré)utilisabilité, est en cours de finalisation ; deux outils numériques ont été créés, quatre articles scientifiques publiés et un thésaurus complet décrivant 365 concepts a été constitué[19]. Le fonctionnement du prochain consortium reposera davantage encore sur les groupes de travail afin de préserver ce dynamisme et d'échapper à tout schéma routinier.

10. Les trois principales difficultés dans l'activité du consortium sur la période 4+4+2

Les principales difficultés connues par CAHIER ont été rencontrées pendant les dernières années de sa labellisation (2019-2021).

a) Le projet post-labellisation

La construction du projet post-labellisation a été tout particulièrement affectée par la crise sanitaire en ne permettant pas aux membres du consortium de se rencontrer, de discuter et de travailler sur le projet. La réflexion sur l'avenir de CAHIER a été lente et difficile car le modèle « consortium » et la « fédération » de projets sont très souples mais très difficiles à ancrer dans un autre cadre.

Dès 2016-2017, les coordinateurs de CAHIER et le comité de pilotage avaient étudié la possibilité de créer un GIS, mais le faible nombre de GIS soutenus, la lourdeur d'un tel dossier qui aurait nécessité les signatures des institutions de tous les projets membres dans un contexte où les chercheurs et enseignants-chercheurs sont débordés de dossiers administratifs (ANR, LabEx, EquipEx, maquettes de formations, etc.) a découragé les coordinateurs de CAHIER.

En 2017-2018, une seconde option, plus légère, a été étudiée par certains membres du CoPil : constituer une société savante. Celle-ci aurait permis à CAHIER de porter des appels à projets pour trouver des financements et de préserver le fonctionnement en réseau de projets. Mais le modèle « associatif » n'a pas recueilli l'adhésion de tous les membres du CoPil, notamment de ceux qui voulaient continuer à travailler avec Huma-Num et qui trouvaient l'option peu opportune dans un périmètre où il existait déjà une société savante des Humanités numériques francophones : l'Association Humanistica[20].

En 2019, après son entretien avec le Conseil scientifique d'Huma-Num, **CAHIER a successivement exploré les trois pistes qui lui avaient été suggérées par le CS : l'implantation du réseau dans les MSH, la création d'un réseau « méthodes et pratiques » et la création d'un nouveau consortium.**

- **1.** Si l'implantation du réseau dans les MSH avait déjà fait l'objet de discussions internes du comité

[19] Voir plus loin les tableaux de la section n°12 recensant l'ensemble de ces livrables.
[20] www.humanisti.ca/

de pilotage en 2018, ce n'est que durant l'année 2020, que des réunions de travail ont été organisées avec le RnMsh et ses directeurs. En raison de la crise sanitaire, elles se sont essentiellement tenues en visioconférence. Malgré un intérêt certain du RnMsh, ces rencontres ont été infructueuses en raison, d'une part, de l'absence de modèles ou d'espaces facilitant l'intégration du « réseau CAHIER » dans le réseau des MSH, et ce malgré les collaborations de membres du consortium avec des MSH, et, d'autre part, en raison du financement puisqu'aucune proposition chiffrée n'a pu été discutée.

- **2.** En juin 2021, le projet de création de réseau « méthodes et pratiques » a été exploré. Une proposition a été élaborée en vue d'être soumise aux membres du consortium (voir Annexe n°5 : Projet de consortium de réseau « REST CAHIER »). Parallèlement, une seconde proposition a été également formulée, celle de la création d'un consortium dédié aux outils, « OLIO » (Voir Annexe n°6, Projet de consortium « OLIO »), et un fil de discussion « CAHIER NEXT » a été lancé.

Avec deux projets en juillet 2021, le consortium a ouvert une phase de dialogue entre les membres et les porteurs des deux projets et organisé une réunion ouverte du comité de pilotage le 15 juillet 2021. Les porteurs ont été invités à étoffer leurs propositions pour le début du mois de septembre 2021 en vue d'une Assemblée générale exceptionnelle qui se tiendrait le 17 septembre 2021. Suite à cette AG, les porteurs des projets « REST » et « OLIO » devaient réviser leurs propositions avant le 30 septembre 2021 voire, éventuellement, les fusionner en vue d'un vote au début du mois d'octobre 2021.

Le début du mois de septembre a été marqué par la démission du coordinateur Thomas Lebarbé le 09/09/2021 et, la veille de l'assemblée générale du 17/09/2021, par le retrait du projet « OLIO » qui a décidé de présenter son projet de consortium à Huma-Num. L'assemblée générale de CAHIER n'a donc discuté que d'un seul projet, le réseau « méthodes et pratiques REST CAHIER », dont elle a pointé l'insuffisance du cadrage scientifique.

- **3.** Suite à cette AG, les porteurs du projet « REST CAHIER » ont considéré qu'élaborer un cadrage scientifique avec des objectifs et des livrables ciblés revenait à rédiger une proposition de nouveau consortium. Ayant identifié des défis comme celui de la poursuite de la « création de corpus numériques » de qualité, c'est-à-dire de corpus faciles à exploiter et à réutiliser ou du perfectionnement des outils et des méthodes développées pendant les deux labellisations, les porteurs ont pris contact avec le consortium COSME pour partager avec eux ces pistes de travail et ont soumis aux membres, le 30/09/2021, un projet préfigurant un nouveau consortium : « CAHIER après CAHIER » (Voir Annexe n°7, Projet « CAHIER après CAHIER »).

Sur la base des propositions énoncée dans le projet « CAHIER après CAHIER », un vote a été organisé en ligne (Belenios a été utilisé) du 4 au 5 octobre 2021 et les membres ont été invités à voter pour le nouveau projet ou pour la fin des travaux. Les résultats du vote ont été les suivants :

- Nombre d'électeurs : **55**
- Nombre de votants : **50**
- Nombre de voix : **51** (1 personne avait une procuration)
- Résultats :
  - Poursuite des travaux de CAHIER dans le cadre d'un nouveau projet de consortium (document "Projet CAHIER 2") : **44 voix**
  - Fin des travaux de CAHIER : **7 voix**

CAHIER va donc consacrer le premier semestre de l'année 2022 à étoffer le projet de consortium, à circonscrire sa nouvelle communauté et ses questions de recherches pour 2023-2027. Le but sera de préserver le meilleur de CAHIER au sein d'une communauté élargie à de nouveaux domaines scientifiques. Ce travail sera coordonné par Alexey Lavrentev (IR CNRS, UMR-IRHIM Lyon) et Stéphanie Dord-Crouslé (CR CNRS, UMR-IRHIM Lyon) durant l'année 2022. Il commencera dès l'assemblée générale des 25 et 26 novembre 2021 prochains durant laquelle des ateliers thématiques seront organisés ; il se poursuivra durant l'année 2022 avec une réunion de travail par mois.

b) La FAIRisation des données du consortium

Dans son retour sur le bilan de CAHIER de l'année 2019 (B.19), le conseil scientifique avait souligné avec raison que celui-ci ne précisait pas suffisamment quel avait été le travail d'accompagnement

des projets qui avait été réalisé en vue d'un stockage sur Nakala[21]. Également retardée par la crise sanitaire, cette tâche n'a pu être réalisée que durant l'année 2020, et essentiellement à distance. Les principales difficultés de la période ont été la faible disponibilité des membres de CAHIER pendant la crise d'une part, et de l'autre, les travaux entrepris par Huma-Num pour mettre à jour le code informatique de Nakala durant la même période. Ces travaux ont eux-mêmes été retardés par la crise sanitaire puisque la livraison de l'outil, initialement prévue pour le printemps 2020, n'a pu être réalisée qu'en décembre 2020. Enfin, il faut également souligner que cette tâche a rencontré quelques oppositions au sein de CAHIER, certains membres considérant cette exigence comme une « injonction » de la coordination du consortium.

C'est le groupe de travail [Data_Cahier] qui a assumé cette tâche et qui l'a organisée en deux étapes : informé durant l'automne 2019 des travaux entrepris en vue du développement de la V2 de Nakala, il a d'abord consacré la fin de l'année 2019 et le début de l'année 2020 à étudier l'état de maturité des projets en vue de leur FAIRisation. Ensuite, il s'est appliqué à lever les verrous à travers trois actions : la sensibilisation des membres du consortium aux critères FAIR, leur accompagnement personnalisé dans le dépôt de leur données et le développement de l'outil qui manquait alors pour réaliser les dépôts de grande quantité de données.

-1. L'analyse de l'état des corpus et la préparation des données.

L'examen mené par le GT entre décembre 2019 et mars 2020 avait estimé le potentiel de données à près de 327.450 fichiers (total des fichiers annoncés sur les sites web des projets membres) et à, potentiellement, 500.000 images ; toutefois, seuls 13.201 fichiers semblaient réellement accessibles sur les sites web pour à peu près 150.000 images.

A la date du 31 octobre 2021, le nombre de ressources FAIR du consortium CAHIER est de 61.107, soit : 57.073 images et 4.034 fichiers XML-TEI[22]. Plus précisément, il s'agit de :

- 1604 images et 1793 fichiers XML-TEI disponibles sur Ortolang (1 projet a utilisé Ortolang)
- 167 images et 167 fichiers XML-TEI disponibles sur Zenodo (1 projet a utilisé Zenodo)
- ~55.302 images et ~16.030 fichiers de données à ce jour dont 2077 fichiers XML TEI (6 projets ont utilisé Nakala)

Sur la base du « FAIR Data Maturity Model: specification and guidelines » publié par la Research Data Alliance en mars 2020[23], un examen précis des projets membres (sites web) permet classer en trois groupes leurs réalisations :

- **FAIR** : seuls six projets remplissent la totalité des conditions présentées par le « FAIR Data Maturity Model » et disposent de données totalement FAIR et stockées sur Nakala. Trois d'entre eux utilisent deux autres entrepôts : Zenodo, l'outil développé par le CERN, et Ortolang.
- **FAI (cas a)**: dans ce premier cas, on trouve les projets qui fournissent leurs fichiers sources en téléchargement (sur leurs sites ou via github) et ceux qui ont utilisé un CMS pour exposer leurs données. Dans les deux cas ils offrent des données trouvables, accessibles et accompagnées de métadonnées riches et dont les fichiers peuvent être téléchargés. Les projets qui ont utilisé *Omeka* (majorité des cas) disposent, a minima, de quinze métadonnées normalisées grâce au Dublin Core (DC) et donc interopérables. Toutefois, certains des enrichissements apportés pour personnaliser les publications numériques ne sont pas pris en compte par les champs DC (ceci supposera un traitement informatique pour gérer ces métadonnées). Les données de ces projets, les images notamment, sont stockées sur des machines locales (MSH, institution de rattachement du porteur) ou virtuelles (délivrées par Huma-Num le plus souvent) et ne présentent pas de garanties de pérennité. L'absence d'identifiant pérenne tel que le DOI ne rend pas ces données totalement FAIR.

  FAI (cas b): dans ce second cas, on trouve des projets qui ont développé en interne ou non, un site

[21] Le travail a commencé tardivement au sein de CAHIER, mais ces réflexions étaient déjà présentes au sein des consortiums d'Huma-Num depuis 2018. Cf. Adeline Joffres, et al.. "The Impact of FAIR Principles on Scientific Communities in (Digital) Humanities. An Example of French Research Consortia in Archaeology, Ethnology, Literature and Linguistics". *DH 2018 Digital Humanities Conference,*, Jun 2018, Mexico, Mexico. https://halshs.archives-ouvertes.fr/hal-02153030

[22] Voir plus loin la section n°13 a). Ces calculs sont basés sur les entrepôts des 10 projets (cellules grisées)

[23] Voir : https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines-0

d'exposition pour leurs éditions enrichies. Dans ce cas, les données sont stockées en interne sur une machine (MSH, Institution ou sur une machine virtuelle délivrée par Huma-Num) et l'édition enrichie est exposée via un site web développé *ad-hoc*. Mais comme les données (fichier TEI et images) ne sont pas stockées de façon pérenne et qu'elles ne disposent pas d'identifiant pérenne tel que le DOI, elles ne sont « **FAIR** ». Parfois, il peut s'agir d'une édition enrichie terminée qui est publiée par une institution qui est aussi une maison d'édition ; dans ce cas la publication peut disposer d'un identifiant pérenne de type DOI, mais pas les données et fichiers sources : c'est la raison pour laquelle les membres du groupe [Data_Cahier] ont considéré qu'il s'agit de projets « **FAI** ».

- **XXXX** : quelques membres ne présentent qu'un site de présentation du projet et aucune donnée n'est consultable, soit parce que le projet n'a pas (encore) conçu de données, soit parce que le projet a restreint l'accès à ses données, soit, enfin, parce que des mises à jour n'ont pas été effectuées et que les liens sont cassés.

Sur la base de cet état des lieux, la première tâche du groupe de travail a consisté à présenter ces résultats aux membres lors de réunions virtuelles intitulées [FAIR_Line] et organisées à partir du printemps 2020. A cette occasion, Laurent Capelli a été invité à présenter les travaux en cours sur Nakala. Ensuite, un guide a été rédigé, à destination des membres et non-membres de CAHIER, pour leur permettre, grâce aux critères de la Research Data Alliance que le GT a traduit en français, de mesurer le degré d'ouverture de leurs projets et d'identifier les informations (métadonnées) nécessaires au dépôt dans Nakala (champs DC ou balises TEI). Le guide décrit également de façon théorique le *mapping* des métadonnées vers Nakala[24] qui allait être implémenté dans l'application web *Mynkl*.

-2. Le dépôt en masse de données accumulées pendant plusieurs années nécessitait le développement d'un outil de dépôt massif. Comme Nakala ne prévoyait pas d'implémenter cette fonctionnalité dès 2020, CAHIER l'a développée pour répondre à ses objectifs de fin de labellisation. Les outils existants, développés avant 2019 par d'autres collègues ou projets étaient inutilisables car Nakala révisait également son modèle de données (RDF). Pour suivre les évolutions du développement de Nakala et accompagner, techniquement, la FAIRIsation des données, un stagiaire en Master Informatique « Sécurité des données » et « Génie logiciel et mobilités » a été recruté pour 6 mois en 2020 et 6 autres mois en 2021. Il a été encadré par Fatiha Idmhand, par Allel Hadjali[25] (spécialiste de la modélisation des données) et par Amin Mesmoudi[26] (spécialiste de Sparql et RDF) du Laboratoire LIAS (Laboratoire d'Informatique et d'Automatique pour les Systèmes) équipe « Ingénierie des Données et des Modèles »[27], puis par Germain Forestier[28] (spécialiste de l'analyse de données) du laboratoire « IRIMAS – Institut de Recherche en Informatique, Mathématiques, Automatique et Signal »[29]. Grâce aux compétences de l'étudiant et de ses encadrants, il a été possible de travailler sur le développement de l'outil de dépôt malgré l'instabilité de l'API de test de Nakala durant l'année 2020. Le code de l'application web a été préparé durant le stage de 2020, et lorsque Nakala a été stabilisé en décembre 2020, l'outil a pu être mis à jour rapidement lors du stage de 2021. Immédiatement présenté aux membres de CAHIER et aux utilisateurs de Nakala (voir ci-après la présentation des livrables du groupe de travail section n°12) lors des [FAIR_Line], une formation à l'outil a été organisée en présentiel lors du colloque de Bordeaux et trois jours de permanence ont été proposés pour accompagner les utilisateurs[30]. En permettant de déposer, en une fois, plusieurs milliers de données, l'outil *Mynkl* lève un verrou majeur, car les membres du consortium ont accumulé des centaines, voire des milliers de données durant dix ans. Toutefois, il ne résout pas les deux autres problèmes rencontrés par le consortium : le temps supplémentaire demandé par certains membres pour réévaluer les questions juridiques liées au dépôt (c'est pour cela que cette tâche sera prolongée durant l'année 2022) et le refus de

[24] Voir « Guides pour la FAIRisation des données » : V1 https://halshs.archives-ouvertes.fr/halshs-02889777 et V2 https://halshs.archives-ouvertes.fr/halshs-03037748 en ligne sur HAL
[25] https://www.lias-lab.fr/members/allelhadjali
[26] https://www.lias-lab.fr/members/aminmesmoudi
[27] https://www.lias-lab.fr/teams/data-engineering
[28] https://germain-forestier.info/
[29] https://www.irimas.uha.fr/
[30] Voir la vidéo qui retrace ce suivi : https://www.youtube.com/watch?v=hygpiLsCJMY

certains porteurs de projets. Ce bilan, qui témoigne de la difficulté de la tâche, montre toutefois que les avancées de l'année 2021 ont engagé une dynamique favorable qui pourra se poursuivre durant l'année 2022.

c) La crise sanitaire

Enfin, sans revenir en détail sur les conséquences des deux confinements du printemps et de l'hiver 2020, la fermeture des lieux de recherches durant presqu'une année entre mars 2020 et juin 2021 a eu un impact non négligeable sur l'activité de CAHIER et de ses membres. De nombreuses activités initiées ou programmées ont été interrompues, reportées, retardées voire annulées en raison de la crise. Certains groupes de travail n'ont pu poursuivre leurs activités tandis que d'autres ont pallié l'absence de rencontres par l'organisation de réunions par visioconférence et l'édition collaborative de documents (ces rencontres en visio se superposant elles-mêmes aux multiples activités réalisées à distance). Ainsi, et pour la première fois de son histoire, CAHIER n'a pu ni organiser son Assemblée générale annuelle 2020 en présentiel, ni organiser le troisième temps de son atelier annuel 2020, ni les formations prévues (TEI) et envisagées (Heurist).

11. Activités consolidées et réalisations du consortium pour la période 4+4+2

Une présentation détaillée des livrables associés à ces activités est proposée ci-après dans les différents tableaux de la section n°12. La liste ci-dessous en offre une vue synthétique organisée selon les trois cycles de CAHIER :

- Cycle 1 : 2011-2015. Construction, ouverture et consolidation du réseau
- Cycle 2 : 2015-2019. Production de livrables par le réseau
- Cycle 3 : 2019-2021. Harmonisation des travaux et FAIRisation

Activités de coordination dans la/les communautés cibles du consortium	Cycle 1 : 2011-2015	Création du réseau, élaboration d'un processus d'adhésion, présentation du réseau à la communauté-cible : celle des sciences des textes
	Cycle 2 : 2015-2019	Extension du réseau et formalisation du processus d'adhésion. Diffusion des travaux dans la communauté cible grâce à un réseau inter-professionnel : enseignants-chercheurs, chercheurs, doctorants, ingénieurs et personnels d'appui à la recherche
	Cycle 3 : 2019-2021	Fin de l'ouverture du réseau Diffusion des résultats dans la communauté cible grâce au réseau inter-professionnel constitué d'enseignants-chercheurs, de chercheurs, de doctorants, d'ingénieurs et de personnels d'appui à la recherche
Activités de mise à disposition et de mutualisation des données, des outils, des méthodes	Cycle 1 : 2011-2015	Formation aux outils et méthodes existants, création d'un atelier annuel. Création de groupe de travail et rédaction d'un premier guide
	Cycle 2 : 2015-2019	Création de plusieurs groupes de travail et rédaction de plusieurs guides Développement de petits outils ou connecteurs permettant de travailler avec plusieurs outils. Poursuite des ateliers annuels pour monter en compétences
	Cycle 3 : 2019-2021	Activités de clôture : FAIRisation des données et développement d'un outil spécifique *Mynkl* pour FAIRiser les données du consortium sur Nakala, rédaction de nouveaux guides, organisation de formations
Bilan des actions de formation (nombre de formations organisées, de personnes formées, analyse du retour des participants…)	Cycle 1 : 2011-2015	Voir ci-après section n°12 compilant la liste exhaustive des formations organisées ou soutenues et le nombre de personnes formées. Il n'y a pas eu d'enquêtes post-formations organisées de façon systématique. Les retours des participants peuvent toutefois être appréciés de deux façons : le nombre de participants est constant bien que le public change et des demandes d'adhésion au consortium ont pu venir après la participation à une action de formation
	Cycle 2 : 2015-2019	
	Cycle 3 : 2019-2021	

Encadrement de stagiaires de Master1 ou Master2	Cycle 1 : 2011-2015	/
	Cycle 2 : 2015-2019	Au cours des trois dernières années, CAHIER a recruté quatre stagiaires de niveau Master 1 et Master 2 pour accompagner les travaux de FAIRisation des données et de clôture du consortium : *Une stagiaire pour l'expertise des fichiers XML-TEI du consortium (1 x 6 mois) : Julie Laurent (Université de Poitiers) *Un stagiaire en informatique (2 x 6 mois) chargé du développement de l'application web *Mynkl* connectée à l'API de Nakala : Ala Eddine Laouir (Université Haute Alsace) *Un stagiaire en Humanités numériques (1 x 6 mois) chargé de préparer les fichiers et de mettre à jour, avec les membres de CAHIER, métadonnées et données en vue du dépôt sur Nakala : Andrés Echavarria (Université de Lorient) *Une stagiaire en documentation et bibliothèques (1 x 1 mois puis financement d'une prestation) chargée du suivi de la rédaction des plans de gestion des données des projets : Laurène L'Hermite (Université de Poitiers) Si leurs compétences ont été très utiles au consortium, cette formation pratique leur a permis d'en acquérir de nouvelles et de construire leurs projets post-masters. A la suite de ce stage, trois d'entre eux ont décroché un contrat doctoral grâce aux connaissances acquises et développées au sein du Consortium et la quatrième a été retenue dans un Master 2 à Madrid pour poursuivre sa formation en *deep learning*. - Ala Eddine Laouir : doctorant en informatique et sécurité des données à partir de septembre 2021 au laboratoire INRIA/LORIA (Université Strasbourg) - Andrés Echavarria : doctorant en études hispaniques et humanités numériques à partir de décembre 2021 au laboratoire IRIEC (Université de Montpellier) - Laurène L'Hermite : doctorante en études théâtrales et archivistique à partir de janvier 2022 au laboratoire CRHIA (Université de La Rochelle) - Julie Laurent : titulaire du Master 2 Humanités numériques de l'Université Complutense de Madrid, actuellement en formation aux méthodes du *deep learning* à l'Université Madrid Complutense.
	Cycle 3 : 2019-2021	

Développement, mise en place d'outils pour la/les communauté(s) cible(s) du consortium	Cycle 1 : 2011-2015	Soutien au développement de Philologic Développement de WebOai
	Cycle 2 : 2015-2019	Finalisation du développement de WebOai Développement d'un connecteur TXM-Métopes
	Cycle 3 : 2019-2021	Développement d'un thésaurus sur OpenTheso Développement de *Mynkl*
Métriques d'accès aux données si celles-ci sont pertinentes (consultation des sites, des bases, des données…)	Cycle 1 : 2011-2015	Voir ci-après section n°12 compilant la liste exhaustive des métriques collectées
	Cycle 2 : 2015-2019	
	Cycle 3 : 2019-2021	
Respect des principes FAIR pendant le cycle de vie des données traitées par le consortium (choix technologiques et méthodologiques, guides de bonnes pratiques, utilisation de services d'Huma-Num, etc.)	Cycle 1 : 2011-2015	Utilisation et recommandation de formats et langages ouverts et interopérables pour les métadonnées: XML-TEI, XML-EAD et Dublin Core Critères de qualité pour les données (utilisation des guides d'Huma-Num pour la numérisation d'images, son, etc.)
	Cycle 2 : 2015-2019	
	Cycle 3 : 2019-2021	Choix prioritaire de Nakala comme entrepôt de stockage des données mais sans contrainte
Initiatives dans le cadre de la Science Ouverte	Cycle 1 : 2011-2015	Pas d'initiative particulière mais promotion de l'application des principes de la science ouverte : les données FAIR en font partie, de même que le dépôt des travaux sur HAL, etc.
	Cycle 2 : 2015-2019	
	Cycle 3 : 2019-2021	
Elaboration et diffusion de bonnes pratiques	Cycle 1 : 2011-2015	Voir ci-après section n°12 compilant la liste exhaustive des guides produits. Ils sont également présentés sur https://cahier.hypotheses.org/guides
	Cycle 2 : 2015-2019	
	Cycle 3 : 2019-2021	

Collaborations avec d'autres Consortiums-HN	Cycle 1 : 2011-2015	Travail avec le consortium COSME dans le cadre du groupe de travail « Questions juridiques ». Voir : https://cahier.hypotheses.org/publication-editions-textes et la journée d'études https://cahier.hypotheses.org/2031
	Cycle 2 : 2015-2019	Travaux avec les consortiums CORLI et MASA en vue des interventions collectives lors des colloques d'ADHO de Mexico (2018) et Utrecht (2019)
	Cycle 3 : 2019-2021	Echanges de pratiques sur Nakala Travail avec COSME en vue d'un nouveau projet de consortium
Collaborations interdisciplinaires	Cycle 1 : 2011-2015	Collaboration avec les sciences des textes historiques et le consortium COSME
	Cycle 2 : 2015-2019	/
	Cycle 3 : 2019-2021	Collaboration avec les sciences de l'informatique pour le traitement des données de CAHIER
Collaborations européennes et/ou internationales (RDA, DARIAH, CLARIN, actions COST, participation à des programmes H2020, colloques internationaux, etc.)	Cycle 1 : 2011-2015	/
	Cycle 2 : 2015-2019	Participation aux ateliers de Dariah (certains membres de CAHIER sont associés à des actions COST, à CLARIN et à des projets H2020)
	Cycle 3 : 2019-2021	Utilisation des travaux de RDA sur les données FAIR
Activités de recherche impulsées, rendues possibles par le consortium et/ou réalisées dans le cadre du consortium (publications, journées d'études...)	Cycle 1 : 2011-2015	Les membres de CAHIER étaient invités à échanger sur leurs questionnements scientifiques
	Cycle 2 : 2015-2019	Articles scientifiques co-écrits par des membres du consortium et signés au titre des travaux menés au sein du Consortium
	Cycle 3 : 2019-2021	Organisation d'un colloque scientifique en juin 2021 : https://cahier10.sciencesconf.org/

12. Description de l'apport à la/aux communauté(s) scientifique(s) concernée(s) par le consortium : liste exhaustive des livrables.

En sus de la description des activités pour la période 4+4+2, les tableaux de synthèse suivants présentent les principaux indicateurs quantifiés des activités du Consortium CAHIER depuis sa création en 2011.

a) Formations, écoles thématiques et rencontres scientifiques organisées (14)

Type d'Activités et thèmes	Date	Nombre	Nombre de participants	Commentaires
Colloque (1)				
« Dix ans avec CAHIER » – Colloque conclusif des travaux du consortium CAHIER – **Bordeaux**	**Du 07 au 10 juin 2021**	1	45 en présentiel 25 en visio	Programme complet : https://cahier10.sciencesconf.org/ Livre des résumés publié sur HAL le 30/07/2021 : https://halshs.archives-ouvertes.fr/halshs-03207669 *Statistiques en date du 30/10/2021* *Consultations de la notice : 286* *Téléchargements de fichiers : 188*
Ecoles thématiques annuelles (9)				
« La TEI pour la recherche: questionner et visualiser un corpus en XML-TEI » – Atelier annuel du Consortium CAHIER **– Lorient**	**Du 3 au 4 septembre 2020** **Et le 26 novembre 2020 (en visio)**	1	30	Atelier annuel organisé par un des membres du consortium Programme complet : https://cahier.sciencesconf.org/ Billet sur : https://cahier.hypotheses.org/category/suivi-du-consortium/atelier-annuel
« Exploiter les corpus d'auteur » – Atelier de formation annuel du consortium Cahier – **Poitiers**	**Du 18 au 20 juin 2019**	1	45	Atelier annuel organisé par un des membres du consortium Programme complet : https://cahier.sciencesconf.org/resource/page/id/1

« Rétro-numérisation de documents historiques et partage dans le Web sémantique : l'exemple de la lexicographie » – Atelier de formation annuel du consortium Cahier – **Montpellier**	**Du 26 au 29 juin 2018**	1	33	L'atelier a obtenu le soutien des ERICS CLARIN et DARIAH. Atelier annuel organisé par un des membres du consortium Programme complet : https://www.mshsud.org/agenda/85-atelier-du-consortium-huma-num-cahier?iccaldate=2019-11-1
« Les dispositifs numériques : des corpus aux usages » – Atelier de formation annuel du consortium CAHIER - **Bordeaux**	**Du 27 au 30 juin 2017**	1	26	Atelier annuel organisé par un des membres du consortium Programme complet : https://cahier.hypotheses.org/3170 Compte-rendu : https://cahier.hypotheses.org/3235
« Structuration et exploitation de données textuelles » – Atelier de formation annuel du consortium CAHIER - **Caen**	**Du 5 au 8 juillet 2016**	1	23	Atelier annuel organisé par un des membres du consortium Programme complet : https://cahier.hypotheses.org/2278
« Préserver, éditer et exploiter les sources de la recherche » – Atelier de formation annuel du consortium CAHIER – **Lille –**	**Du 29 juin au 3 juillet 2015**	1	50	Atelier annuel organisé par un des membres du consortium Programme complet : https://cahier.hypotheses.org/1892
« Édition analytique » – Atelier de formation annuel du consortium CAHIER – **Lyon –**	**Du 30 juin au 4 juillet 2014**	1	30	Atelier annuel organisé par un des membres du consortium Programme complet : https://cahier.hypotheses.org/1059
« Sources éditées : du texte à la structure, de la structure aux formes » – Atelier de formation annuel du consortium CAHIER – **Grenoble**	**Du 8 au 12 juillet 2013**	1	30	Atelier annuel organisé par un des membres du consortium Programme complet : https://cahier.hypotheses.org/689
"Sources éditées : du texte à la structure, de la structure aux formes" – Atelier de formation annuel du consortium CAHIER – **Caen**	**Du 9 au 13 juillet 2012**	1	20	Premier atelier annuel organisé par l'un des membres du consortium Programme complet : https://cahier.hypotheses.org/209 Compte-rendu : https://cahier.hypotheses.org/date/2012/07
Formations spécifiques (3)				
" Formation à l'outil OpenTheso et au format SKOS"	**Du 30 au 31 janvier 2018**	1	13	Formation organisée par le consortium. Formateur : Miled Rousset

"XSLT pour la création de pages HTML à partir de dossiers XML-TEI – niveau avancé"	**Du 26 au 27 janvier 2017**	1	13	Formation organisée par le consortium. Formatrice : Elena Pierazzo
"Introduction au langage XSLT pour la création de pages HTML à partir de dossiers XML-TEI"	**Du 21 au 22 mars 2016**	1	20	Formation organisée par le consortium. Formatrice : Elena Pierazzo
Actions Nationales de Formations, Ecoles d'été (2)				
ANF « Concevoir et exploiter les sources numériques de la recherche en SHS »	**2021** **2019** **2018** **2017** **2015** **2014** **2013** **2012**	1 1 1 1 1 1 1 1	20 29 26 34 30 30 39 42	Soutien financier à l'organisation de l'ANF portée par la MSH Val de Loire Soutien financier de chercheurs non CNRS
Ecole d'été éditions numériques EDEEN - 28 mai - 2 juin 2018 – Grenoble	**2018**	1	45	Soutien à l'organisation de l'école d'été portée par la MSH Alpes
Soutien à la formation des membres du consortium				
« TEI 2 : Encoder en XML-TEI, niveau avancé » - 10-11 septembre 2021, Tours	**2021**	1	23	Soutien financier à l'organisation de la formation portée par le CESR + Prise en charge des frais de mission de membres du consortium
"Initiation à l'encodage XML-TEI des textes patrimoniaux" - CESR-BVH - Tours	**2021** **2019** **2018** **2017** **2015** **2014** **2013** **2012**	1 1 1 1 1 1 1 1	20 29 26 34 30 30 39 42	Prise en charge de 4 inscriptions par le consortium (frais d'inscription et de mission de membres du consortium) pour participer à la formation. Soutien financier à l'organisation de la formation (nouvelles modalités de financement de la formation mise en place par le CESR)

Journées « Ontologie en Sciences Humaines et Sociales » - Consortium MASA – MSH Val de Loire – Tours	**9 – 10 novembre 2015**	1	1 mission prise en charge par Cahier	Prise en charge par le consortium des frais d'inscription et de mission de membres du consortium pour participer à la formation
ANF Web sémantique – TGIR Huma-Num	**Du 22 au 25 septembre 2014**	1	2 missions prises en charge par Cahier	Prise en charge par le consortium des frais d'inscription et de mission de membres du consortium pour participer à la formation
Semaine de co-working autour de SynopsX	**Du 3 au 6 novembre 2014**	1	1	Prise en charge par le consortium des frais de mission d'un partenaire pour une semaine de coworking à Constance coorganisée avec l'Atelier des humanités numériques AHN de l'ENS de Lyon

b) Communication à destination de la communauté et impact (3)

Type de valorisation	Indicateurs de production	Indicateurs de fréquentation	Commentaires	Liens
Blog (1)				
Carnet : https://cahier.hypotheses.org/	129 articles + 96 pages qui sont mises à jour régulièrement	317 122 visites (de octobre 2014 au 25 octobre 2021) 3 775 visites mensuelles en moyenne	Statistiques fournies par la plateforme Hypothèses. Un site a existé de 2011 à 2014 (aucune statistique de consultation n'est disponible pour cette version). Le carnet hypothèses a été ouvert en octobre 2014, toutes les pages du site antérieur ont	https://cahier.hypotheses.org/

			été importées dans le carnet.	
Réseaux sociaux (2)				
Compte twitter Consortium CAHIER @ccahier	301 Tweets (depuis la création du compte en décembre 2011)	566 abonnés (au 30/10/2021)		
Compte facebook Cahier @consortium.cahier	76 Posts (depuis la création de la page le 31/10/2016)	106 personnes aiment ça 113 personnes suivent ce lieu		
Listes (1)				
Liste CoPiL (liste réservée aux membres du comité de pilotage du consortium)	19 abonnés à la liste	19 abonnés peuvent communiquer sur la liste	Email de la liste : copil.cahier@listes.huma-num.fr	
Liste AG (liste réservée aux membres du consortium)	89 abonnés à la liste	89 abonnés peuvent communiquer sur la liste	Email de la liste : ag.cahier@listes.univ-tours.fr	
Liste ouverte aux personnes qui suivent les activités du consortium	253 abonnés à la liste	253 abonnés reçoivent les informations du consortium par mail	Email de la liste : cahier@groupes.renater.fr	
Vidéos (1)				
Vidéos youtube	1 vidéo réalisée par Huma-Num	Mise en ligne le 18/10/2021 32 vues le 25/10/2021	Statistiques de youtube	Massive data FAIRification https://www.youtube.com/watch?v=hygpiLsCJMY

c) Publications : guides méthodologiques (6)

Type de valorisation	Indicateurs de production	Indicateurs de fréquentation	Commentaires	Liens et métriques HAL
Guides (6)				
(2021) Plan de gestion des données	1 PGD de structure 10 PGD modèles pour les projets	10 projets ont préparé leurs plans de gestion des données	Les PGD modèles ont servi d'outils aux projets	Version 1 déposée le 29/10/2021 https://halshs.archives-ouvertes.fr/halshs-03409421 *Dépôt récent : statistiques non disponibles à ce jour*
(2021) Guide : « Décrire les textes dans le cadre d'une édition numérique. Le thésaurus "Typologie textuelle" du Consortium CAHIER »	1 Guide pour le thésaurus comportant 365 concepts décrivant 365 genres textuels	Thésaurus finalisé en juin 2021		Version 1 déposée le 25/10/2021 https://halshs.archives-ouvertes.fr/halshs-03402679 *Dépôt récent : statistiques non disponibles à ce jour*
(2020) Guide pour la FAIRisation des données des corpus d'auteurs préparé par le [Groupe de travail Data_Cahier]	Guide en téléchargement sur le carnet du consortium			Version 1 déposée le 22/07/2020 https://halshs.archives-ouvertes.fr/halshs-02889777 *Statistiques en date du 30/10/2021* *Consultations de la notice : 729* *Téléchargements de fichiers : 192* Version 2 déposée le 22/01/2021 https://halshs.archives-ouvertes.fr/halshs-03037748 *Statistiques en date du 30/10/2021* *Consultations de la notice : 424* *Téléchargements de fichiers : 232* Mémorandum déposé le 28/10/2021 https://halshs.archives-ouvertes.fr/halshs-03408209 *Dépôt récent : statistiques non disponibles à ce jour*

(2017) L'édition numérique de correspondances – guide méthodologique (novembre 2017)	Guide en téléchargement sur le carnet du consortium	4 025 visites (au 30/10/2021)	Statistiques fournies par la plateforme Hypothèses, n'incluent pas les lectures et téléchargements HAL	
(2017) L'édition numérique de corpus d'auteurs – aspects juridiques : Guide préparé par le groupe de travail "Questions juridiques" du consortium CAHIER (Corpus d'auteurs pour les humanités : informatisation, édition, recherche)	12 pages publiées sur le carnet du consortium depuis 2017	8 428 visites (au 30/10/2021)	Statistiques fournies par la plateforme Hypothèses	Version 1 déposée le 25/10/2021 https://halshs.archives-ouvertes.fr/halshs-03400177 *Dépôt récent : statistiques non disponibles à ce jour*
(2015) La publication des corpus d'auteurs, éditions de textes : Informations et recommandations (avril 2015), version PDF puis réédité et mis à jour sur le carnet d'hypothèses pour en faciliter la lecture	Guide en téléchargement sur le carnet du consortium et tiré à 500 exemplaires distribués par les membres du consortium,	2 244 visites (au 30/10/2021)	Statistiques fournies par la plateforme Hypothèses, n'incluent pas les lectures et téléchargements HAL	Version française publiée le 23/11/2018 : https://halshs.archives-ouvertes.fr/halshs-01932519 *Statistiques en date du 30/10/2021* *Consultations de la notice : 656* *Téléchargements de fichiers : 536* Version en espagnol déposée le 24/06/2019 https://halshs.archives-ouvertes.fr/halshs-02164065 *Statistiques en date du 30/10/2021* *Consultations de la notice : 513* *Téléchargements de fichiers : 406*

d) Outils développés (3)

Type de valorisation	Indicateurs de production	Indicateurs de fréquentation	Commentaires	Liens et métriques
Outils développés et objectifs				
WebOai : outil générateur d'URL OAI en vue de l'exposition et du moissonnage des données XML TEI	1 outil	11 projets ont utilisé le service	L'outil WebOai a été développé entre 2013 et 2016, avant la création de l'entrepôt Nakala. Il avait pour but de faciliter la création d'URL OAI. Les projets développant des ressources XML-TEI pouvaient ainsi exposer leurs données.	http://weboai.cahier.huma-num.fr/ 11 projets ont utilisé le service WebOai pour exposer leurs données. Toutefois, avec la création de Nakala, les ressources WebOai n'ont pas été moissonnées par Isidore.
Thésaurus des genres textuels « Typologie »	1 thésaurus comportant 365 concepts 365 genres textuels décrits 1 guide 1 article scientifique en cours de finalisation	Le thésaurus a été finalisé en juin 2021 Les *handle* sont d'ores et déjà disponibles via Nakala	Le thésaurus « Typologie » a été élaboré entre 2019 et 2021, il décrit 365 genres textuels (concepts). Chaque concept est pourvu d'un identifiant de type *handle*.	Typologie 43 https://opentheso.huma-num.fr/opentheso/ Les identifiants *handle* sont moissonnés par Nakala. Il reste à associer les mots clés des fichiers déposés dans Nakala aux concepts du thésaurus et à ces identifiants
Outils de dépôt en masse de données sur Nakala « *Mynkl* »	1 outil connecté à l'API de Nakala Facilite le dépôt de milliers de données	Les données accessibles dans Nakala et citées ci-dessous ont utilisé *Mynkl* pour déposer leurs données sur Nakala	Le dépôt en masse de données sur Nakala est un service qu'Huma-Num va développer à la fin de l'année 2021. CAHIER a répondu à ce besoin pour FAIRiser les données du consortium en développant cette application web.	http://mynakala.huma-num.fr/ *Mynkl* est une application web. L'utilisateur se connecte avec sa propre clé API pour utiliser le service. *Mynkl* ne stocke aucune donnée Le « mémo » d'utilisation de Mynkl a été déposé sur HAL le 28/10/2021 : https://halshs.archives-ouvertes.fr/halshs-03408209
CAHIER soutient et accompagne les développements de TXM et a soutenu le développement de Philologic. Il utilise et forme à BasicX, Métopes (développement d'un connecteur le connecteur Metopes-TXM), oXygen et Opentheso. La plupart des membres de CAHIER utilisent sharedocs pour la gestion de leurs projets numériques.				

e) Ressources publiées : données FAIR accessibles dans les entrepôts de données

Type de valorisation	Indicateurs de production	Indicateurs de fréquentation	Commentaires	Liens
Données accessibles sur Nakala				
Ressources FAIR du consortium	~55.302 images et ~16.030 fichiers de données à ce jour dont 2077 fichiers XML TEI (6 projets ont utilisé Nakala)	Données qui pourraient être fournies par Huma-Num pour suivre la réutilisation des données	Données déposées en masse à l'aide de l'outil *Mynkl*	Toutes les données sont pourvues d'un identifiant DOI
Données accessibles sur ORTOLANG				
1 projet a FAIRisé ses données à l'aide d'ORTOLANG	1604 images (cartes et lettres postales), au format JPEG 1793 textes au format XML (TEI P5)	Voir statistiques d'Ortolang : https://www.ortolang.fr/	https://repository.ortolang.fr/api/content/corpus14/8/Corpus14.xml et https://www.ortolang.fr/market/corpora/corpus14	
Données accessibles sur ZENODO				
1 projet a FAIRisé ses données à l'aide de Zenodo	167 images et 167 fichiers XML-TEI disponibles sur Zenodo	Voir statistiques de Zenodo	https://zenodo.org/record/5167263#.YVhje6SxWpo	
Données accessibles sur GITHUB				
2 projets utilisent GITHUB pour diffuser leurs données	439 fichiers XML-TEI disponibles sur github	Voir statistiques de fournies par github	https://github.com/ArchivesNationalesFR/editionTestamentsDePoilus https://github.com/Antonomaz/Corpus	

f) Données signalées et exposées dans Isidore

Type de valorisation	Indicateurs de production	Indicateurs de fréquentation	Commentaires	Liens et métriques
Données signalées sur Isidore				
Signalement dans ISIDORE	3 083 résultats (au 26/10/21)		Interrogation de ISIDORE le 26/10/21 avec « consortium cahier »	
Données consultables sur les sites web des projets membres	58 projets membres en 2021	NC	Voir ci-après tableau recensant la liste des projets	Voir ci-après tableau recensant la liste des projets et les liens URL fournis

g) Diffusion scientifique

Type de valorisation	Indicateurs de production	Indicateurs de fréquentation	Commentaires	Liens et métriques HAL
Colloques et conférences (23)				
Participations à des colloques au nom de CAHIER	15 soutiens financiers (limités à 600€) pour participation à des colloques	Public présent lors des interventions	Les membres soutenus ont indiqué dans les remerciements de leur publication et dans leur présentation le soutien obtenu du consortium	Sur HAL : 36 résultats avec une recherche fondée sur HAL structure ID : https://halshs.archives-ouvertes.fr/search/index/q/*/structId_i/545625/ Parmi ces résultats, on trouve des supports de présentation utilisés lors de colloques, des articles et des chapitres d'ouvrages
Invitations dans des colloques	Une invitation au minimum par an pour intervenir lors de colloques et présenter CAHIER (ADHO, DHnord, MSH, etc.)	Public présent lors des interventions	Invitations à des fins d'explication et de promotion du consortium	

				Sur Isidore, la recherche par mots clés « consortium CAHIER » et la sélection par type de documents « Colloques et conférences » donne **255 résultats** : https://isidore.science/s?type=http%3A%2F%2Fisidore.science%2Fontology%23conference&q=consortium+cahier#

h) Activités internationales

Type d'action	Rôle	Date - durée	Commentaires
Participation à un projet H2020	Participant EOSC Pillar	2021	Promotion des données FAIR : https://www.youtube.com/watch?v=hygpiLsCJMY
Participation à un WG (DARIAH, CLARIN, RDA, etc.)	Participant WG de Dariah sur « Standardization Survival Kit »	24 et 25 janvier 2019	A la suite de l'atelier, un modèle d'outil SSK a été étudié pour l'édition TEI. Il n'a pu être finalisé faute de temps. https://www.dariah.eu/2019/01/25/standardization-survival-kit-workshop-1-2019-textual-data-scenarios/
Animation d'un réseau	Participant	2017-2021	Des membres du consortium participent à des actions COST, notamment « Distant reading » (2 membres)
Participation à un webinaire, une conférence internationale	Participant	ADHO 2018 ADHO 2019 TEI 2016 TEI 2019 DH 2016 DH 2014	Participation aux tables rondes thématiques organisées par Huma-Num lors des conférences d'ADHO à Mexico (2018) et Utrecht (2019) Participation aux conférences TEI Rome en 2013 (M-L.Demonet, alors coordinatrice de CAHIER, était *keynote speaker*), Lyon en 2015 (CAHIER était l'un des « sponsors » de ce colloque : voir le site http://tei2015.huma-num.fr/fr/,), Vienne en 2016 et Tokyo en 2018 Participation aux conférences d'ADHO à Cracovie en 2016 et à Lausanne en 2014

i) Publications scientifiques propres au consortium

Type de valorisation	Indicateurs de production	Indicateurs de fréquentation	Commentaires	Liens et métriques HAL
Publications propres aux activités du consortiums				
Ouvrage collectif « Dix ans de corpus d'auteurs »	1	Nombre de livres achetés Téléchargements du livre des résumés	Livre à paraître en novembre 2021 chez l'éditeur Editions Archives Contemporaines. Le livre sera accessible en total OA, pourvu d'un DOI. Chaque article sera aussi pourvu d'un DOI. L'éditeur fournira des statistiques de consultation Le livre des résumés a été déposé sur HAL le 29/04/2021 https://halshs.archives-ouvertes.fr/halshs-03207669 *Statistiques en date du 30/10/2021* *Consultations de la notice : 290* *Téléchargements de fichiers : 191*	
Articles scientifiques	4	Galleron I., Idmhand F., Lavrentev A., Demonet A., *article scientifique en cours de finalisation sur la typologie des genres textuels* Galleron I., Idmhand F., « Why Go from Texts to Data, or The Digital Humanities as A Critique of the Humanities », *in* Word and Text - a journal of literary studies and Linguistics, 2020, http://jlsl.upg-ploiesti.ro/No_1_2020.html Galleron I., Idmhand F., « De l'interopérabilité à la réutilisabilité des éditions électroniques », *Humanités numériques* [En ligne], 1 | 2020, mis en ligne le 01 janvier 2020. URL : http://journals.openedition.org/revuehn/350 ; DOI : https://doi.org/10.4000/revuehn.350 ; https://isidore.science/document/10670/1.j76gme Galleron I., Idmhand F., Meynard C, « Que mille lectures s'épanouissent… Modélisation du personnage et expérience de « crowdreading » », DHQ: Digital Humanities Quarterly (2018) Déposé sur HAL le 14/06/2018 https://halshs.archives-ouvertes.fr/halshs-01815606 *Statistiques en date du 30/10/2021* *Consultations de la notice : 247* *Téléchargements de fichiers : 215*		

Compte HAL pour recenser la production scientifique de CAHIER et de ses membres (dont les articles scientifiques)	33	36 documents déposés *Date de la consultation : 30/10/2021*	Lien de consultation de la structure « CAHIER » sur HAL : https://halshs.archives-ouvertes.fr/search/index/q/*/structId_i/545625/ 36 documents déposés *Date de la consultation : 30/10/2021*

13. Liste consolidée des membres du consortium

a) Liste des 58 projets membres actifs à la date du 30-10-2021

Ces membres ont été successivement intégrés au consortium de 2011 à 2019

	Nom du projet Type de publications numériques	Nom et mail du responsable scientifique	Le cas échéant, nom de l'ingénieur responsable du volet numérique	Liste des membres du projet (nom + mails)	Site internet du projet	URL accès aux fichiers sources Disponibles à la date du 29/10/2021	Eventuellement, URL des entrepôts de données
1.	**EUROPOLENI** **AE**[31]	Marie-Therese Cam : marie-therese.cam@univ-brest.fr	Gwenaelle Patat : gwenaelle.patat@univ-rennes2.fr	Marie-Thérèse Cam : mcam@univ-brest.fr ; Eric Francalanza : Eric.francalanza@univ-brest.fr ; Geoffrey Williams ; Frédérique Plantevin : frederique.plantevin@univ-brest.fr ; Sofia Talas : sofia.talas@pd.infn.it	en cours	sur Humanum box	à venir
2.	**Édition électronique des œuvres de Louis de Boissy** **EE**	Ioana Galleron : ioana.galleron@gmail.com, Geoffrey Williams : williams@licorn-research.fr		Ioana Galleron : ioana.galleron@gmail.com, Geoffrey Williams : williams@licorn-research.fr	Site : http://www.licorn-research.fr/Boissy.html	81 fichiers XML-TEI https://nakala.fr/collection/11280/ce360734	81 fichiers XML-TEI https://nakala.fr/collection/11280/ce360734
3.	**Basnage** **EE**	Geoffrey Williams : williams@licorn-research.fr		Ioana Galleron : ioana.galleron@gmail.com, Geoffrey Williams : williams@licorn-research.fr, Joséphine Loterie, Estelle Allanic, Géraldine Le Bihan	Site : http://www.licorn-research.fr/Basnage.html	32 fichiers XML-TEI https://nakala.fr/collection/11280/f8aeea42	32 fichiers XML-TEI https://nakala.fr/collection/11280/f8aeea42
4.	**Bibliothèques Virtuelles Humanistes (BVH)** **EE**	Chiara Lastraioli : chiara.lastraioli@univ-tours.fr	Sandrine Breuil : sandrine.breuil@univ-tours.fr	Chiara Lastraioli : chiara.lastraioli@univ-tours.fr ; Marie-Luce Demonet : marie-luce.demonet@univ-tours.fr ; Toshinori Uetani : toshinori.uetani@univ-tours.fr ; Claire Sicard : claire.sicard@univ-tours.fr ; Rémi Jimenes : remi.jimenes@univ-tours.fr ; Anne-Laure Allain : anne.allain@univ-tours.fr ; Sandrine Breuil : sandrine.breuil@univ-tours.fr		Pour chaque édition TEI BVH-EPISTEMON, via la consultation de l'édition sous XTF > menu télécharger : TEI, PDF, HTML	

[31] Voir définitions supra en section n°7. AE = Archives éditorialisées / EE = Editions enrichies (EE)

5.	**Archives éLV(Archives Numériques de l'École de Lvov-Varsovie, Kasimir Twardowski Archives) AE**	Wioletta Miskiewicz : wioletta1miskiewicz@gmail.com			Site : http://www.elv-akt.net/		
6.	**Etudes sur la Renaissance d'Horace (ERHO) EE**	Nathalie Dauvois: ndauvois@gmail.com	Paul Gaillardon paul.gaillardon@ish-lyon.cnrs.fr	Paul Gaillardon: paul.gaillardon@ish-lyon.cnrs.fr, Astrid Quillien astrid.quillien@gmail.com, Tristan Vigliano tristan.vigliano@univ-amu.fr	Site : http://ihrim.huma-num.fr/nmh/Horatius/	Les TEI sont accessibles sur le site : http://ihrim.huma-num.fr/nmh/Horatius/ Et http://ihrim.huma-num.fr/nmh/Horatius/XML/fabricius-2.xml	
7.	**Les dossiers de Bouvard et Pécuchet EE**	Stéphanie Dord-Crouslé : stephanie.dordcrousle@ens-lyon.fr		Maud Ingarao : maud.ingarao@ens-lyon.fr	Site : http://www.dossiers-flaubert.fr/	Site : http://www.dossiers-flaubert.fr/	En cours sur Nakala
8.	**Base de français médiéval (BFM) EE**	Celine Guillot : celine.guillot@ens-lyon.fr	Alexey Lavrentev : alexei.lavrentev@ens-lyon.fr	Celine Guillot : celine.guillot@ens-lyon.fr ; Serge Heiden : slh@ens-lyon.fr ; Alexey Lavrentev : alexei.lavrentev@ens-lyon.fr ; Nadine Pontal : Nadine.Pontal@ens-lyon.fr	Site : http://txm.bfm-corpus.org/	170 fichiers XML-TEI https://nakala.fr/collection/10.34847/nkl.ea05m33f	170 fichiers XML-TEI https://nakala.fr/collection/10.34847/nkl.ea05m33f
9.	**Digital Theological Hobbes (DTH) EE**	Francesca Rebasti : francesca.rebasti@insa-lyon.fr	Serge Heiden : slh@ens-lyon.fr	Francesca Rebasti : francesca.rebasti@insa-lyon.fr ; Serge Heiden : slh@ens-lyon.fr	à venir	francesca.rebasti@insa-lyon.fr	en cours
10.	**Montesquieu, bibliothèque & éditions EE**	Catherine Volpihac : catherine.volpilhac@ens-lyon.fr		Catherine Volpihac : catherine.volpilhac@ens-lyon.fr , Maud Ingarao : maud.ingarao@ens-lyon.fr, Nadine Pontal : nadine.pontal@ens-lyon.fr	Site : http://montesquieu.huma-num.fr/accueil		
11.	**@mbrosius AE**	Aline Canellis : aline.canellis@univ-st-etienne.fr		Aline Canellis : aline.canellis@univ-st-etienne.fr	Site : https://auctorpatrum.hypotheses.org/author/auctorpatrum		
12.	**Charles Fontaine AE**	Elise Rajchenbach : elise.rajchenbach@univ-st-etienne.fr	Paul Gaillardon paul.gaillardon@ish-lyon.cnrs.fr	Elsa Kammerer; Marc Desmet marc.desmet@univ-st-etienne.fr; Nina Mueggler nina.mueggler@unifr.ch; Florence Bonifay flo.bonifay@laposte.net; Maud Lejeune maudlaetitia.lejeune@gmail.com; Pauline Dorio pauline.dorio4@gmail.com; Mathilde Vidal mathilde.vidal1@gmail.com; Jérémie Bichüe j.bichue@gmail.com; Grégoire Holtz volfony@hotmail.com; Ugo Pais upais@laposte.net; Sandra Provini sandra.provini@univ-rouen.fr; Sarah Delale sarah.delale@wanadoo.fr	Site: http://chfontaine.huma-num.fr/projet/presentation		
13.	**Correspondance de Pierre Bayle EE**	Anthony Mc Kenna : mckenna@univ-st-etienne.fr	Eric Olivier Lochard: Éric Olivier Lochard <eolochard@free.fr>	Anthony Mc Kenna : mckenna@univ-st-etienne.fr; Fabienne Vial-Bonacci: fabienne.vial@univ-st-etienne.fr	Site : http://bayle-correspondance.univ-st-etienne.fr/		

14.	**Marc-Michel Rey** **EE**	Christelle Bahier Porte : christelle.porte@univ-st-etienne.fr	Fabienne Vial-Bonacci fabienne.vial@univ-st-etienne.fr	Fabienne Vial : fabienne.vial@univ-st-etienne.fr liste complète : http://rey.huma-num.fr/equipe	Site : http://rey.huma-num.fr/home		
15.	**Digital Matteo Ricci** **EE**	Vito Avarello : vito.avarello@univ-st-etienne.fr	Fabienne Vial-Bonacci fabienne.vial@univ-st-etienne.fr				
16.	**SATELLITES – Les intellectuels « satellites ». Un autre regard sur la circulation des idées** **AE**	Fatiha Idmhand : fatihaidmhand@yahoo.es			Site internet : https://cortazar.nakalona.fr/	2395 fichiers de données (15350 images) https://nakala.fr/collection/10.34847/nkl.dcfeojtz En cours de dépôt : Archives de Fernando Aínsa : 4113 documents décrits à ce stade et 2719 images (2,61 Go en Jpeg)	2395 fichiers de données (15350 images) https://nakala.fr/collection/10.34847/nkl.dcfeojtz En cours de dépôt : Archives de Fernando Aínsa : 4113 documents décrits à ce stade et 2719 images (2,61 Go en Jpeg)
17.	**Projet Ichtya** **EE**	Brigitte Gauvin : brigitte.gauvin@unicaen.fr		Brigitte Gauvin ; Pierre-Yves Buard : pierre-yves.buard@unicaen.fr; Barbara Jacob : barabara.jacob@unicaen.fr; Marie Bisson : marie.bisson @unicaen.fr ; Thierry Buquet : thierry.buquet@unicaen.fr ; catherine Jacquemard : catherine.jacquemard@unicaen.fr ; Marie-Agnès Lucas-Avenel : marie-agnes.avenel@unicaen.fr	Site : http://www.unicaen.fr/puc/sources/depiscibus/accueil		
18.	**Projet Malaterra** **EE**	Marie-Agnes Avenel : marie-agnes.avenel@unicaen.fr	Pierre-Yves Buard : pierre-yves.buard@unicaen.fr		Site : https://www.unicaen.fr/puc/sources/malaterra/	Fichiers sources non téléchargeables	
19.	**Montedite** **EE**	Carole Dornier : carole.dornier@unicaen.fr	Pierre-Yves Buard : pierre-yves.buard@unicaen.fr	Carole Dornier : carole.dornier@unicaen.fr, Pierre-Yves Buard : pierre-yves.buard@unicaen.fr	https://www.unicaen.fr/services/puc/sources/Montesquieu/		
20.	**Les Écrits de l'abbé de Saint-Pierre** **EE**	Carole Dornier : carole.dornier@unicaen.fr, Claudine Poulouin : Claudine.Poulouin@univ-rouen.fr, Pascal Buelon : pascal.buleon@unicaen.fr	Julia Roger : julia.roger@unicaen.fr	Carole Dornier : carole.dornier@unicaen.fr, Claudine Poulouin : Claudine.Poulouin@univ-rouen.fr, Patrizia Oppici : patrizia.oppici@unimc.it; Pascal Buléon : pascal.buleon@unicaen.fr, Julia Roger : julia.roger@unicaen.fr,	Site : https://www.unicaen.fr/puc/sources/castel/accueil		
21.	**Nouvelle Édition Numérique de Fac-similés de Référence (Nénufar)** **EE**	Agnes Steuckardt : agnes.steuckardt@univ-montp3.fr	Hervé Bohbot : herve.bohbot@cnrs.fr	Giancarlo Luxardo : giancarlo.luxardo@univ-montp3.fr	Site : http://nenufar.huma-num.fr	9952 images https://nakala.fr/collection/11280/13816c8c	9952 images https://nakala.fr/collection/11280/13816c8c

22.	**Corpus 14** **EE**	Agnes Steuckardt : agnes.steuckardt@univ-montp3.fr	Giancarlo Luxardo : giancarlo.luxardo@univ-montp3.fr		https://www.univ-montp3.fr/corpus14/ Praxiling - UMR 5267 (2019). *Corpus 14* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v2, https://hdl.handle.net/11403/corpus14/v2	1604 images (cartes et lettres postales), au format JPEG 1793 textes au format XML (TEI P5) https://repository.ortolang.fr/api/content/corpus14/8/Corpus14.xml et https://www.ortolang.fr/market/corpora/corpus14	1604 images (cartes et lettres postales), au format JPEG 1793 textes au format XML (TEI P5) https://repository.ortolang.fr/api/content/corpus14/8/Corpus14.xml et https://www.ortolang.fr/market/corpora/corpus14
23.	**Édition électronique de la correspondance de Flaubert** **EE**	Yvan Leclerc : yvan-leclerc@wanadoo.fr			Site : http://flaubert.univ-rouen.fr/correspondance/edition		
24.	**Œuvres complètes de D'Alembert** **AE**	Alexandre Guilbaud : alexandre.guilbaud@imj-prg.fr, Irene Passeron : irene.passeron@wanadoo.fr			Sites : http://dalembert.academie-sciences.fr/ et http://dalembert.hypotheses.org/		
25.	**Enccre** **AE**	Alexandre Guilbaud : alexandre.guilbaud@imj-prg.fr, Irene Passeron : irene.passeron@wanadoo.fr			Site: http://enccre.academie-sciences.fr/		
26.	**Frénaud numérique** **AE**	Marianne Froye : marianne.froye@univ-fcomte.fr	Sébastien Jacquot : sebastien.jacquot@univ-fcomte.fr	Thomas Dandin : thomas.dandin@univ-fcomte.fr, Marianne Froye: marianne.froye@univ-fcomte.fr; , Sébastien Jacquot : sebastien.jacquot@univ-fcomte.fr		en cours sur Nakala	en cours sur Nakala
27.	**Fonds numérique Jean-Luc Lagarce** **AE**	Pascal Lécroart : pascal.lecroart@univ-fcomte.fr	Sébastien Jacquot : sebastien.jacquot@univ-fcomte.fr	Thomas Dandin : thomas.dandin@univ-fcomte.fr, Pascal Lécroart : pascal.lecroart@univ-fcomte.fr, Sébastien Jacquot : sebastien.jacquot@univ-fcomte.fr	Site : http://fanum.univ-fcomte.fr/lagarce/	en cours Nakala	en cours Nakala
28.	**Schola Rhetorica** **AE**	Christine NOILLE christine.noille-clauzade@wanadoo.fr	Joseph Fahey: fahey.joseph@gmail.com' 'ELAN : Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr'	Christine NOILLE christine.noille@sorbonne-universite.fr, Francis Goyet francis.goyet@gmail.com Joseph Fahey fahey.joseph@gmail.com, Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Cecile Lignereux cecile.lignereux@univ-grenoble-alpes.fr, Benoit Sans Benoit.Sans@ulb.be, Emmanuel Bury	Site : https://schola-rhetorica.org		

				emmanuel.bury@sorbonne-universite.fr, pascale.mounier@univ-grenoble-alpes.fr			
29.	**Manuscrits de Stendhal** **EE**	Françoise Leriche : francoise.leriche@univ-grenoble-alpes.fr		Cécile Meynard : cecile.meynard@univ-angers.fr ; Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr ; Jean-Jacques Labia jj.labia@gmail.com ; Hélène de Jacquelot helene.dejacquelot@unipi.it	Site : http://manuscrits-de-stendhal.org/		
30.	**Les deux évasions de Benoîte Groult (E2E)** **AE**	Cecile Meynard : cecile.meynard@univ-angers.fr	ELAN : Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr	Cecile Meynard : cecile.meynard@univ-angers.fr, Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr	Site : http://www.espace-transcription.org/corpus/description/3		
31.	**CORR-Proust** **AE**	Françoise Leriche : francoise.leriche@univ-grenoble-alpes.fr	ELAN : Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr	Françoise Leriche : francoise.leriche@univ-grenoble-alpes.fr, Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr ; Nathalie Mauriac Dyer : nathalie.mauriac@ens.psl.eu, Julie André : julie.andre@polytechnique.edu, Pyra Wise : Pyra.Wise@ens.fr, Caroline Szylowicz : szylowic@illinois.edu, François Proulx : fproulx@illinois.edu, Yuri Cerqueira Dos Anjos : yuri.cerqueiradosanjos@vuw.ac.nz	Site : http://proust.elan-numerique.fr/ 'https://gitlab.com/litt-arts-num/docker-proust/-/blob/master/README.md		
32.	**Édition Numérique des Cahiers d'Henri de Régnier (ENCHRE)** **AE**	Bernard Roukhomovsky : bernard.roukhomovsky@univ-grenoble-alpes.fr	ELAN : Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr	Bernard Roukhomovsky : bernard.roukhomovsky@univ-grenoble-alpes.fr, Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr + partenaires	Site : http://enchre.elan-numerique.fr/		
33.	**Projet La Réticence** **AE**	Brigitte Combe : brigitte.combe@gmail.com	ELAN : Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr	Brigitte Combe : brigitte.combe@gmail.com, Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr	Site : http://reticence.elan-numerique.fr/		
34.	**L'invention du théâtre antique dans l'Europe de la première modernité – Commentaires et**	Malika Bastin : malika.bastin@univ-grenoble-alpes.fr, Pascale Paré-Rey : pascale.pare-rey@univ-lyon3.fr	ELAN : Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-	Malika Bastin, Pascale Paré-Rey, Marc Douguet, Sabine Lardon, Laure Hermand-Schebat, Christian Nicolas, Elysabeth Hue-Gay, Emmanuelle Morlock, Christiane Louette, Jean-Yves Vialleton, Anne Garcia Fernandez, Elisabeth Greslou, Arnaud Bey,	Site : https://ithac.hypotheses.org/ http://ithac.elan-numerique.fr/		

	paratextes (ITHAC) **EE**		alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr	Alexia Dedieu, Sarah Gaucher, Louisa Laj			
35.	**Tacitus On Line** **EE**	Isabelle Cogitore : isabelle.cogitore@univ-grenoble-alpes.fr	ELAN : Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr	Elisabeth Greslou : elisabeth.greslou@univ-grenoble-alpes.fr, Anne Garcia-Fernandez : annegf@univ-grenoble-alpes.fr, Arnaud Bey : arnaud.bey@univ-grenoble-alpes.fr, Louis Autin (ls.autin@gmail.com) et liste sur le site http://tacitus.elan-numerique.fr/index.php?page=about	Site : http://tacitus.elan-numerique.fr/		
36.	**Archives des Traducteurs et des Écrivains de la Littérature Italienne : Éditions et Recherches (ATELIER)** **EE**	Filippo Fonio : filippo.fonio@univ-grenoble-alpes.fr	Belinda Missiroli : belinda.missiroli@univ-grenoble-alpes.fr			Recherche de financement toujours en cours	Recherche de financement toujours en cours
37.	**« Testaments de Poilus » – Une plateforme de transcription collaborative au service du patrimoine manuscrit** **AE**	Emmanuelle De Champs : edechamp@cyu.fr, Florence Clavaud : florence.clavaud@culture.gouv.fr	Florence Clavaud : florence.clavaud@culture.gouv.fr		Plateforme de transcription des Testaments de Poilus : https://testaments-de-poilus.huma-num.fr/ ; plateforme de restitution (édition) : https://edition-testaments-de-poilus.huma-num.fr/	https://github.com/ArchivesNationalesFR/editionTestamentsDePoilus (accès aux fichiers XML-TEI à l'unité depuis le site d'édition) 139 fichiers XML-TEI	
38.	**Écritures savantes au siècle des Lumières. La correspondance et les carnets de visiteurs de Jean-François Séguier** **AE**	Emmanuelle Chapron : emmanuelle.chapron@univ-amu.fr	Eric Carroll : eric.carroll@univ-amu.fr (ne s'occupe plus du projet depuis 2019)	Coresponsable scientifique du projet : François Pugnière. Contributeurs : Andrea Bruschi (andreabruschi@gmail.com), Lily Servel, Adeline Danerol, Florence Catherine, Véronique Chapron, Céline Buttin, Claire Torreilles, Etienne Stockland, Meike Knittel, Nicolas Rieucau, Jean Boutier, Gilles Montègre	Site : https://www.seguier.org/ (non maintenu depuis le départ d'Eric Carroll en 2019, inactif depuis 2020). Transfert des données en cours vers un site en construction.	https://www.seguier.org/correspondance/correspondance.aspx (idem : inaccessible). Données stockées sur le Sharedocs CAHIER	En cours de transfert vers Nakala. La collection est prête, mais encore privée. Elle sera publique en novembre 2021.
39.	**Création de corpus de livrets d'opéra sous l'ancien régime, l'Académie royale de**	Margareta Kastberg : margareta.kastberg@univ-fcomte.fr	Sebastien Jacquot : sebastien.jacquot@univ-fcomte.fr	Margareta Kastberg : margareta.kastberg@univ-fcomte.fr ; Sebastien Jacquot : sebastien.jacquot@univ-fcomte.fr	Site : http://fanum-txm.univ-fcomte.fr/txm/	en cours Nakala	En cours de transfert vers Nakala.

	musique de Paris (1671-1791) **AE**						
40.	**BIBLINDEX** **EE**	Laurence Mellerin : laurence.mellerin@mom.fr		Elysabeth Hue-Gay : elysabeth.hue-gay@mom.fr	Site : https://biblindex.org	https://www.biblindex.org/api	
41.	**Projet E-STAMPAGES** **EE**	Michele Brunet : michele.brunet@univ-lyon2.fr	Bruno Morandiere : Bruno.Morandiere@resefe.fr	Michele Brunet : michele.brunet@univ-lyon2.fr, Adeline Levivier : adeline.levivier@gmail.com , Emmanuelle Morlock : emmanuelle.morlock@mom.fr	Site : https://www.e-stampages.eu/		
42.	**Mauriac en ligne** **AE**	Jessica de Bideran : jessica.de-bideran@u-bordeaux-montaigne.fr; Caroline Casseville : caroline.casseville@u-bordeaux-montaigne.fr, Philippe Baudorre : philippe.baudorre@u-bordeaux-montaigne.fr	Jessica de Bideran : jessica.de-bideran@u-bordeaux-montaigne.fr (durant années post-doc)	Jessica de Bideran : jessica.de-bideran@u-bordeaux-montaigne.fr ; Caroline Casseville : caroline.casseville@u-bordeaux-montaigne.fr, Philippe Baudorre : philippe.baudorre@u-bordeaux-montaigne.fr	Site : https://mauriac-en-ligne.huma-num.fr/ Carnet de recherche : https://mauriacenligne.hypotheses.org		
43.	**Correspondance électronique d'Émile Zola (CorrELEZ)** **AE**	Jean-Sebastien Macke : jean-sebastien.macke@cnrs.fr			Site : http://www.archives-zoliennes.fr		
44.	**Thresors de la Renaissance** **AE**	Anne Reach Ngo : anne.reachngo@yahoo.fr	Richard Walter : richard.walter@ens.fr		Site : https://eman.hypotheses.org/198		
45.	**Natale Conti, Mythologia, 1567-1627** **AE**	Celine Bohnert : celine.bohnert@univ-reims.fr	Richard Walter : richard.walter@ens.fr		Site : https://eman-archives.org/Mythologia/		
46.	**« Renan Source » Une édition génétique numérique des manuscrits d'Ernest Renan** **AE**	Domenico Paone : domenico.paone@ens.fr	Richard Walter : richard.walter@ens.fr		Site : http://www.item.ens.fr/projet-renan-source/		
47.	**Archives Marguerite Audoux** **AE**	Bernard-Marie Garreau : bernard-marie.garreau@wanadoo.fr	Richard Walter : richard.walter@ens.fr		Site : https://eman-archives.org/Audoux/		

48.	**POR FAVOR** **AE**	Ludivine Thouverez : ludivine.thouverez@univ-poitiers.fr	Michael Nauge : michael.nauge@univ-poitiers.fr		Site : https://porfavor.nakalona.fr/		
49.	**Sociorama. Littérature panoramique internationale du XIXe siècle** **AE**	Nathalie Preiss : blaguezac@wanadoo.fr, Valerie Stienon : valerie.stienon@univ-paris13.fr	Laurent Simon et Benoît Morimont: l.simon@uliege.be et benoit.morimont@uliege.be		Site, en construction:http://web.philo.ulg.ac.be/sociorama/		
50.	**Les Archives d'Augustin Thierry (ArchAT)** **AE**	Aude Deruelle : aude.deruelle@univ-orleans.fr	Cyril Masset : cyril.masset@cnrs-orleans.fr ; Henri Seng : henri.seng@cnrs-orleans.fr		Site (en construction) : http://basex.irht.cnrs.fr/archat/accueil.html		On prévoit un dépôt sur Nakala
51.	**Base Louis Meigret** **AE**	Cendrine Pagani : cendrine.pagani@gmail.com			Site : https://meigret.j2p.fr/		
52.	**TransDiary-TEI** **EE**	Regis Schlagdenhauffen : regis.schlagdenhauffen@ehess.fr , rschlagd@ehess.fr					
53.	**ERabbinica** **EE**	Daniel Stoekl Ben Ezra : daniel.stoekl@ephe.psl.eu		Ron Naiweld : ron.naiweld@ehess.fr, Hayim Lapin : hlapin@umd.edu, Pawel Jablonski pawel.jablonski@etu.ephe.psl, Elena Lolli elena.lolli@ephe.psl.eu, Bronson Brown-DeVost bronson.brown-devost@ephe.psl.eu, Avigail Ohali, avigail.ohali@gmail.com	Site: https://editions.erabbinica.org/S07326.xml	https://github.com/umd-mith/mishnah-data 167 fichiers jpeg et 167 fichiers txt	https://zenodo.org/record/5167263#.YVhje6SxWpo
54.	**MERCURE GALANT** **AE**	Anne Piejus : anne.piejus@cnrs.fr	Thomas Bottini: thomas.bottini@cnrs.fr	Nathalie Berton-Blivet: nathalie.berton-blivet@cnrs.fr, Rebecca Bristow : rebecca.bristow@cnrs.fr	Site provisoire : https://obvil.sorbonne-universite.fr/corpus/mercure-galant/		En cours de transfert vers Nakala.
55.	**Antonomaz** **AE**	Karine Abiven : karine.abiven@sorbonne-universite.fr	Alexandre Bartz : alex.bartz@outlook.fr	Karine Abiven : karine.abiven@sorbonne-universite.fr; "Alexandre Bartz" <alex.bartz@outlook.fr>; Gael Lejeune : gael.lejeune@sorbonne-universite.fr, "Jean-Baptiste TANGUY" <jean-baptiste.tanguy@sorbonne-universite.fr>;	Documentation : http://stih-sorbonne-universite.fr/dokuwiki/doku.php?id=antonomaz; Outils et données : https://github.com/Antonomaz	~300 textes encodés en XML-TEI	https://github.com/Antonomaz/Corpus
56.	**Inventaire Condorcet** **AE**	Nicolas Rieucau : niko99@wanadoo.fr		Claire Bustarret : Claire.Bustarret@ehess.fr ; Nicolas Rieucau : niko99@wanadoo.fr. (Liste	www.inventaire-condorcet.com	https://www.inventaire-condorcet.com [onglet	

				complète des membres : https://www.inventaire-condorcet.com/Presentation/Equipe)		"Principales données"]	
57.	**Privilèges de librairie en France à l'époque moderne (XVIe XVIIe siècles)** **AE**	Edwige Keller : edwige.keller@univ-lyon2.fr	Pierre-Yves Jallud CNRS-ENS de Lyon	2020-2021 : Maxime Cartron (cartron.maxime@gmail.com) ; 2019- : Hélène Lannier (helenelannier@yahoo.fr) ; 2018 : Ghazi Eljorf (G.Eljorf@univ-lyon2.fr) ; 2018-2019 : Michèle Clément (michele.clement@univ-lyon2.fr), Rémi Jimenes (remi.jimenes@univ-tours.fr), Sabine Juratic (sabine.juratic@ens.psl.eu), Henriette Pommier (henriette.pommier@ens-lyon.fr), Daniel Régnier-Roux (Daniel.Regnier-Roux@ens-lyon.fr)	Site : https://privileges-librairie.huma-num.fr/	https://privileges-librairie.huma-num.fr/fiches-privileges	Entrepôt BnF - Fichiers pourvus d'URL ark (exp: https://catalogue.bnf.fr/ark:/12148/cb333694104)
58.	**Espace Afrique-Caraïbe** **AE**	RIFFARD Claire, claire.riffard@cnrs.fr		Nicolas MARTIN-GRANEL (yanikos@aol.com) ; Sonia Le Moigne-Euzenot (sonialemoigneeuzenot@gmail.com) ; Serge Meitinger (serge.meitinger@gmail.com) ; karolina Resztak (karolina_resztak@o2.pl) ; Céline Gahungu (cgahungu@hotmail.fr) ; Élisabeth DEGON (babeth2gon@gmail.com)	https://eman-archives.org/francophone/		

b) Impact de l'adhésion au consortium CAHIER sur les 58 membres actifs à la date du 30-10-2021

Pour mesurer cet impact, nous avons interrogé les membres sur deux aspects : l'obtention de financements grâce aux formations, travaux et inclusion dans le réseau de CAHIER, et les publications. Ce tableau est fondé sur les déclarations des membres, tous n'ont pas répondu à l'enquête.

	Nom du projet	**Nom et mail du responsable scientifique**	**Financements obtenus grâce à la participation du projet à CAHIER (ANR, CollEx, Région, Université, MSH, Européen, IUF, etc.)**	**Publications liées au projet et à la participation dans CAHIER (avec URL le cas échéant)**
1.	**EUROPOLENI**	Marie-Therese Cam : marie-therese.cam@univ-brest.fr	Aide précieuse et indispensable de la MSHB et de Gwenaelle Patat. Le projet s'arrête en décembre. Il a reçu l'aide du département du Finistère avec un an de post doc, un financement de l'Institut des sciences de l'homme et de la société-UBO, un financement de la MSHB, l'aide d'une infographie de l'UFR Lettres, un financement de l'UMR de mathématiques. La référence est toujours Europoleni. Une ANR élargissant le corpus et intégrant la fabrique des instruments de physique de Poleni sera déposée en 2022 : dépôt de la pré-soumission fin octobre. Les originaux déposés sur site seront transcrits et encodés, traduits du latin (pour la moitié d'entre eux). Le projet sera porté par E. Francalanza, Prof. de littérature du XVIIIe s., spécialiste de l'épistolaire. Sa dimension internationale et interdisciplinaire sera confortée. Financement CAHIER	Un article à paraître en histoire des mathématiques ; un millier de lettres bientôt en ligne et le reste sera déposé dans Nakala, en attente de tri. un master a été soutenu, deux thèses sont en cours ; le musée Poleni de l'université de Padoue a été inauguré le 1er septembre dernier. Les webinaires organisés par Gwenaelle permettent d'avancer. Le site sera ouvert fin décembre, mais il y a encore beaucoup de travail.

2.	**Édition électronique des œuvres de Louis de Boissy**	Ioana Galleron : ioana.galleron@gmail.com, Geoffrey Williams : williams@licorn-research.fr	Financement CAHIER	Galleron, Ioana ; Idmhand, Fatiha, « Why Go from Texts to Data, or the Digital Humanities as A Critique of the Humanities », Word and Text, no. X/ 2020, p. 53-69, http://jlsl.upg-ploiesti.ro/site_engleza/No_1_2020.html. Galleron, Ioana ; Idmhand, Fatiha, « 'Réutilisabilité' : L'utilisateur dans l'édition électronique », revue Humanistica, numéro 1, 2019, https://revues.univ-lyon3.fr/humanites-numeriques/ Galleron, Ioana ; Idmhand, Fatiha ; Meynard, Cécile, « Que mille lectures s'épanouissent… Modélisation du personnage et expérience de 'crowdreading' », Digital Humanities Quaterly, volume 12, no. 1, 2018, http://www.digitalhumanities.org/dhq/vol/12/1/000363/000363.html; 2. Ioana Galleron, Fatiha Idmhand, Marie-Luce Demonet, Cécile Meynard, Elena Pierazzo, et al. LES PUBLICATIONS NUMERIQUES DE CORPUS D'AUTEURS - Guide de travail, grille d'analyse et recommandations (V1-Novembre 2018). [Rapport de recherche] Huma-Num ; identifiant : halshs-01932519
3.	**Basnage**	Geoffrey Williams : williams@licorn-research.fr	Projet ANR BasNum https://anr.fr/Projet-ANR-18-CE38-0003 https://basnage.hypotheses.org/ Financement CAHIER	
4.	**Bibliothèques Virtuelles Humanistes (BVH)**	Chiara Lastraioli : chiara.lastraioli@univ-tours.fr	Financement Formation TEI et développement projet EPISTEMON Financement CAHIER	Rapport Ioana Galleron, Marie-Luce Demonet, Cécile Meynard, Idmhand Fatiha, Elena Pierazzo et al. Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations (V1-Novembre 2018) [Rapport de recherche] Huma-Num. 2018, 19 p https://halshs.archives-ouvertes.fr/halshs-01932519 ; halshs-01337873v1 Ouvrage (y compris édition critique et traduction) Marie-Luce Demonet, Alain Legros, Mathieu Duboc, Lauranne Bertrand, Alexei Lavrentiev. Michel de Montaigne, Essais, 1588 (Exemplaire de Bordeaux), édition numérique génétique (XML-TEI/ PDF) Marie-Luce Demonet. 2016, Marie-Luce Demonet https://halshs.archives-ouvertes.fr/halshs-01337873 ; halshs-02132956v1 Communication dans un congrès Toshinori Uetani, Guillaume Porte, Sandrine Breuil, Mathieu Duboc. « Bibliotheques françoises » or A Virtual Workshop for the Literary History of Early Modern France TEI Conference 2018, Text Encoding Initiative Consortium, Sep 2018, Tokyo, Japan https://halshs.archives-ouvertes.fr/halshs-02132956 ; halshs-01225079v1 Autre publication Marie-Luce Demonet, Toshinori Uetani, Lauranne Bertrand. François Rabelais, La Sciomachie, Lyon, Sébastien Gryphe, 1549. Edition numérique en XML/TEI 2015 https://halshs.archives-ouvertes.fr/halshs-01225079 ; hal-02068085v1 Chapitre d'ouvrage

				Marie-Luce Demonet. La confiscation des données issues de l'humanisme numérique Véronique Ginouvès; Isabelle Gras. La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques, Presses universitaires de Provence, 2018, Digitales, 9791032001790 https://halshs.archives-ouvertes.fr/hal-02068085
5.	**Archives éLV(Archives Numériques de l'École de Lvov-Varsovie, Kasimir Twardowski Archives)**	Wioletta Miskiewicz : wioletta1miskiewicz@gmail.com	Financement CAHIER	
6.	**Etudes sur la Renaissance d'Horace (ERHO)**	Nathalie Dauvois: ndauvois@gmail.com	Financement CAHIER : Financement de formation TEI et achat documents à numériser	
7.	**Les dossiers de Bouvard et Pécuchet**	Stéphanie Dord-Crouslé : stephanie.dordcrousle@ens-lyon.fr	Financement CAHIER	medihal-01322398v1 Vidéo Stéphanie Dord-Crouslé, Christian Dury. Les dossiers de Bouvard et Pécuchet 2016 halshs-01083474v1 Chapitre d'ouvrage Stéphanie Dord-Crouslé. Les "seconds volumes" possibles de Bouvard et Pécuchet : l'avènement d'un lecteur-auteur ? Dominique Pety. Patrimoine littéraire en ligne : la renaissance du lecteur ?, Éditions de l'université de Savoie, pp.117-131, 2016, Corpus, 978-2-919732-44-9 halshs-00736015v1 Article dans une revue Stéphanie Dord-Crouslé, Emmanuelle Morlock, Raphaël Tournoy. Nouveaux objets éditoriaux. Le site d'édition des dossiers documentaires de Bouvard et Pécuchet (Flaubert) Les Cahiers du numérique, Lavoisier, 2012, 7 (3-4/2011 "Empreintes de l'hypertexte. Rétrospective et évolution", sous la dir. de Caroline Angé), pp.123-145. ⟨10.3166/LCN.7.3-4.123-145⟩ halshs-00957031v1 Article dans une revue Stéphanie Dord-Crouslé. Le creuset flaubertien : l'édition des dossiers documentaires de Bouvard et Pécuchet, prologue à l'avènement de "seconds volumes possibles" Revue Flaubert, Centre Flaubert, 2014, pp.1-12 halshs-00957029v1 Direction d'ouvrage, Proceedings, Dossier Stéphanie Dord-Crouslé. Revue Flaubert, n° 13 - "Les dossiers documentaires de Bouvard et Pécuchet": l'édition numérique du creuset flaubertien, Actes du colloque de Lyon des 7-9 mars 2012 Université de Rouen, pp.300, 2014 halshs-00441286v2 Chapitre d'ouvrage Stéphanie Dord-Crouslé, Emmanuelle Morlock. L'édition électronique des dossiers de Bouvard et Pécuchet de Flaubert : des fragments textuels en quête de mobilité

				Catherine Bougy, Carole Dornier et Catherine Jacquemard. Le patrimoine à l'ère du numérique : structuration et balisage, Presses universitaire de Caen, pp.79-89, 2011, Schedae halshs-00760914v1 Rapport Stéphanie Dord-Crouslé. Compte-rendu de fin de projet -Projet ANR-07-CORP-009 BOUVARD - Les Dossiers de Bouvard et Pécuchet de Flaubert. Enrichissement, valorisation, documentation d'un corpus multi supports [Rapport de recherche] ANR (Agence Nationale de la Recherche - France). 2012 halshs-00602644v1 Article dans une revue Stéphanie Dord-Crouslé, Emmanuelle Morlock. Sur le modèle du kaléidoscope : concevoir l'édition électronique du "second volume" de Bouvard et Pécuchet Nouveaux cahiers François Mauriac, Société des éditions Grasset (1993-), 2011, pp.169-183 halshs-00838143v1 Communication dans un congrès Stéphanie Dord-Crouslé. Fragments textuels et catégories de classement. Un cas d'utilisation de XML-TEI dans le dispositif éditorial du corpus BOUVARD Éditions critiques et génétiques en Rhône-Alpes, Jun 2013, Grenoble, France halshs-01390741v1 Communication dans un congrès Stéphanie Dord-Crouslé. Le bêtisier agricole de Bouvard et Pécuchet 42nd Annual Nineteenth-Century French Studies Conference - « La terre », Gretchen Schultz (Brown University) et Pratima Prasad (University of Massachusetts Boston), Oct 2016, Providence (RI), Brown University, États-Unis
8.	**Base de français médiéval (BFM)**	Celine Guillot : celine.guillot@ens-lyon.fr	Financement CAHIER	Lavrentiev, Alexei, Bourdot, Charles et Heiden, Serge (2018) « Métopes + TXM: Integrating Text Publishing and Text Analysis Tools Based on TEI Encoding », in TEI as a Global Language. Book of Abstracts. The 18th Annual TEI Conference and Members' Meeting, Tokyo, Japon, p. 255-256. https://halshs.archives-ouvertes.fr/halshs-03363491 Lavrentiev, Alexei et Guillot-Barbance, Céline (2021) « La Base de français médiéval et le consortium CAHIER : dix ans d'échanges et de collaborations », in 10 ans avec CAHIER. Des corpus d'auteurs pour les humanités à leur exploitation numérique, 7-10 juin 2021, Bordeaux, France. https://halshs.archives-ouvertes.fr/halshs-03363517
9.	**Digital Theological Hobbes (DTH)**	Francesca Rebasti : francesca.rebasti@insa-lyon.fr	AAP Labex COMOD 2019-2020, projet Digital Theological Hobbes (DTH), https://comod.universite-lyon.fr/digital-theological-hobbes-dth--112843.kjsp Financement CAHIER	Rebasti, Francesca; Heiden, Serge, 2019, "The Problem of Hobbes and the Bible: A Textometric Approach", https://doi.org/10.34894/I4U4TH
10.	**Montesquieu, bibliothèque & éditions**	Catherine Volpihac : catherine.volpilhac@ens-lyon.fr	Financement CAHIER	
11.	**@mbrosius**	Aline Canellis : aline.canellis@univ-st-etienne.fr		
12.	**Charles Fontaine**	Elise Rajchenbach : elise.rajchenbach@univ-st-etienne.fr	AAP de l'Université Jean Monnet Saint-Étienne 2019	Élise Rajchenbach, article "Charles Fontaine, passeur du De Asse", in "Les Noces de Philologie et de Guillaume Budé (septembre 2021)

			IUF junior 2021: "Pratiques et pensée du réseau à la Renaissance (1500-1550)	https://chfontaine.hypotheses.org/page/4; Élise Rajchenbach, "Penser, révéler, exploiter l'humanisme en réseaux: le cas de l'édition numérique des œuvres complètes de Charles Fontaine", école d'été du CERCOR "Humanités numériques et corpus", 6-8 septembre 2021 Paul Gaillardon: "Interroger l'œuvre de Charles Fontaine par requêtes XPath", école d'été du CERCOR; Élise Rajchenbach, "Charles Fontaine Parisien: une enfance à l'ombre de Notre-Dame (sur quelques documents d'archives récemment exhumés", dans la _BHR_: https://chfontaine.hypotheses.org/1551 Élise Rajchenbach, communication et article dans "Autour de Clément Marot et des recueils collectifs. Configuration des champs poétiques français (1536-1537)": https://chfontaine.hypotheses.org/1516 Elise Rajchenbach, article sur Nicole Le Jouvre dans la BHR: https://chfontaine.hypotheses.org/1465 Paul Gaillardon, article dans _Le Français Pré-classique_: https://chfontaine.hypotheses.org/1500/ Jérémie Bichüe, thèse de doctorat: https://chfontaine.hypotheses.org/1471/ Élise Rajchenbach, article dans la BHR sur Nicole Le Jouvre: https://chfontaine.hypotheses.org/1465 Élise Rajchenbach, article dans "Chacun son Horace": https://chfontaine.hypotheses.org/1076 Projet pédagogique: création de l'article Wikipedia de Charles Fontaine: https://fr.wikipedia.org/wiki/Charles_Fontaine Communication au séminaire spécifique de l'IHRIM, 2018: https://chfontaine.hypotheses.org/726
13.	**Correspondance de Pierre Bayle**	Anthony Mc Kenna : mckenna@univ-st-etienne.fr		La Correspondance de Pierre Bayle, éd. E. Labrousse et A. McKenna et al., Oxford, Fondation Voltaire, 1999-2017, 15 vol. (et de nombreux articles) dans des revues spécialisées)
14.	**Marc-Michel Rey**	Christelle Bahier Porte : christelle.porte@univ-st-etienne.fr	Financement CAHIER	C. Bahier-Porte et F. Vial-Bonacci, « Le Commerce du livre à la lumière de la correspondance – M. M. Rey, P. Rousseau et Ch. Weissenbruch », in Trois siècles d'histoire du livre et de la pensée à travers le Fonds Weissenbruch. Du 'Journal encyclopédique' aux humanités numériques, Bruxelles, Archives générales du Royaume, 2020, p. 207-222; https://hal-ujm.archives-ouvertes.fr/hal-02298356v1F Vial-Bonacci, « L'Edition numérique de la Correspondance de Marc-Michel Rey, libraire du XVIIIe siècle », in Inventorier les correspondances des Lumières, Londres, Andrew Brown, (à paraître 2021). https://hal-ujm.archives-ouvertes.fr/hal-02298361v1 A. Mckenna et F. Vial-Bonacci, « Les Manuscrits clandestins dans les papiers de Marc-Michel Rey », La Lettre Clandestine, Paris, PUPS, 2015, n° 23, p. 25-45. https://hal-ujm.archives-ouvertes.fr/ujm-01490359v1
15.	**Digital Matteo Ricci**	Vito Avarello : vito.avarello@univ-st-etienne.fr	Financement CAHIER	
16.	**SATELLITES – Les intellectuels « satellites ». Un autre regard sur la**	Fatiha Idmhand : fatihaidmhand@yahoo.es	Financement CAHIER pour le recrutement d'un stagiaire. Travail qui a permis de préparer le dossier présenté ensuite à CollEx	Fatiha Idmhand "Para un estudio "computacional" de los Intelectuales satélites" *in* Quiroga. Revista de patrimonio iberoamericano, Núm. 14 (Julio-Diciembre **2018**), ISSN 2254-7037, https://revistaquiroga.andaluciayamerica.com/index.php/quiroga/index

	circulation des idées			Galleron, Ioana ; Idmhand, Fatiha, « Why Go from Texts to Data, or the Digital Humanities as A Critique of the Humanities », Word and Text, no. X/ 2020, p. 53-69, http://jlsl.upg-ploiesti.ro/site_engleza/No_1_2020.html. Galleron, Ioana ; Idmhand, Fatiha, « 'Réutilisabilité' : L'utilisateur dans l'édition électronique », revue Humanistica, numéro 1, 2019, https://revues.univ-lyon3.fr/humanites-numeriques/ Galleron, Ioana ; Idmhand, Fatiha ; Meynard, Cécile, « Que mille lectures s'épanouissent… Modélisation du personnage et expérience de 'crowdreading' », Digital Humanities Quaterly, volume 12, no. 1, 2018, http://www.digitalhumanities.org/dhq/vol/12/1/000363/000363.html; 2. Ioana Galleron, Fatiha Idmhand, Marie-Luce Demonet, Cécile Meynard, Elena Pierazzo, et al. LES PUBLICATIONS NUMERIQUES DE CORPUS D'AUTEURS - Guide de travail, grille d'analyse et recommandations (V1-Novembre 2018). [Rapport de recherche] Huma-Num ; identifiant : halshs-01932519
17.	**Projet Ichtya**	Brigitte Gauvin : brigitte.gauvin@unicaen.fr	Financement CAHIER	Mise en ligne de la Bibliothèque Ichtya : https://ichtya.unicaen.fr/lab/bibliotheque/ Mise en ligne du thesaurus noms latins de poissons et de créatures aquatiques figurant dans les textes latins d'ichtyologie antique et médiévale : https://ichtya.unicaen.fr/lab/thesaurus/ Mise en ligne de la bibliographie d'Ichtya sur Zotero.org : https://www.zotero.org/groups/ichtya/items Lucas-Avenel, Marie-Agnès, « À propos d'un monstre marin inédit de Thomas de Cantimpré », dans Inter litteras & scientias. Recueil d'études en hommage à Catherine Jacquemard, éd. Brigitte Gauvin et Marie-Agnès Lucas-Avenel, Caen, Presses Universitaires de Caen, 2019 (coll. Miscellanea), p. 97-116 Gauvin, Brigitte, « Carolus Figolus, Ichtyologia », dans Inter litteras & scientias. Recueil d'études en hommage à Catherine Jacquemard, éd. Brigitte Gauvin et Marie-Agnès Lucas-Avenel, Caen, Presses Universitaires de Caen, 2019 (coll. Miscellanea), p. 33-64 Buard, Pierre-Yves, « Le réseau de la baleine ou la visualisation de l'histoire d'un texte », dans Inter litteras & scientias. Recueil d'études en hommage à Catherine Jacquemard, éd. Brigitte Gauvin et Marie-Agnès Lucas-Avenel, Caen, Presses Universitaires de Caen, 2019 (coll. Miscellanea), p. 185-198 Buquet T., Clavel B., Gauvin B., Jacquemard C. & Lucas-Avenel M.-A. (éd.), Animaux aquatiques et monstres des mers septentrionales (imaginer, connaître, exploiter, de l'Antiquité à 1600) (actes du colloque de Cerisy, 31 mai-3 juin 2017), in Anthropozoologica 53/2. http://anthropozoologica.com/53/fasc2 Lucas-Avenel Marie-Agnès, « Les "monstres marins" sont-ils des "poissons" ? Le livre VI du Liber de natura rerum de Thomas de Cantimpré », RursuSpicae, 11. En ligne : http://journals.openedition.org/rursus/132 ; DOI : 10.4000/rursus.1320 Jacquemard Catherine, Gauvin Brigitte, Lucas-Avenel Marie-Agnès (ed.), avec la collaboration de C. Février et F. Lecocq, HORTVS SANITATIS, Livre IV, Les poissons, Collection Fontes & Paginæ, Caen, Presses universitaires de Caen, 2013, 496p.

18.	**Projet Malaterra**	Marie-Agnes Avenel : marie-agnes.avenel@unicaen.fr	Financement CAHIER	Marie-Agnès Lucas-Avenel. Geoffroi Malaterra, Histoire du Grand Comte Roger et de son frère, Robert Guiscard, vol. 1, Livres I & II. Presses universitaires de Caen, 2016 Marie-Agnès Lucas-Avenel. Les silences de l'Anonyme du Vatican dans sa réécriture de l'Histoire de Geoffroi Malaterra. Jouanno, Corinne. Les Silences de l'historien, Oublis, omissions, effets de censure dans l'historiographie antique et médiévale, Brepols, pp.275-299, 2019, Giornale Italiano di Filologia - Bibliotheca, 20, ⟨10.1484/M.GIFBIB-EB.5.117913⟩ Marie-Agnès Lucas-Avenel. Chronologie et organisation narrative dans les œuvres historiographiques de Geoffroi Malaterra et Guillaume de Pouille. Bourgain, Pascale Tilliette, Jean-Yves. Le sens du temps : actes du VIIe Congrès du Comité international de latin médiéval (Lyon, 10-13 septembre 2014) = The sense of time : proceedings of the 7th Congress of the International Medieval Latin commitee (Lyon, 10-13 september 2014), Droz, pp.619-636, 2017 Marie-Agnès Lucas-Avenel. Écrire la conquête : une comparaison des récits de Guillaume de Poitiers et Geoffroi Malaterra. Bates, David; van Houts, Elisabeth; D'Angelo, Edoardo. People, Texts and Artefacts : Cultural Transmission in the Norman Worlds of the Eleventh and Twelfth Centuries, Institute of Historical Research, pp.161-178, 2017.
19.	**Montedite**	Carole Dornier : carole.dornier@unicaen.fr	Financement CAHIER	Dornier C. Montesquieu et la tradition des recueils de lieux communs. Revue d'Histoire littéraire de la France, 2008, n° 4, p. 809-820 ; « Lectures portables et recueils : les cahiers de travail de Montesquieu », XIIe Congrès international des Lumières(SIEDHS), juillet 2007, Montpellier, Table ronde « Bibliothèques d'écrivains au 18ème siècle » . « Un manuscrit de Montesquieu en ligne : le projet Montedite », XIIe Congrès international des Lumières(SIEDHS), juillet 2007, Montpellier, Table ronde « l'édition électronique d'œuvres complètes I », organisée par Alexandre Guilbaud et Guillaume Jouve;. id. Les "Pensées" de Montesquieu comme espace de constitution de l'auteur, Studi francesi, n° 161, 2010, p. 304-314 Dornier C., « De la compilation de fragments au texte intégral : histoire de l'édition des Pensées de Montesquieu, Revue française d'histoire du livre, n° 132, 2011, p. 231-250 ; id. "L'histoire du manuscrit des Pensées de Montesquieu", Revue d'Histoire littéraire de la France, Paris, PUF, juillet 2012, n° 3, p. 593-600 « Les Pensées, laboratoire intellectuel de Montesquieu », Revue Montesquieu, n° 7, 2003-2004, parue en février 2005, p.7 (introduction au volume) « La Mise en archive de la réflexion dans les Pensées de Montesquieu », Revue Montesquieu, n° 7, 2003-2004, parue en février 2005, p. 25-39 Dornier C., Buard P.-Y. « L'édition électronique de cahiers de travail : l'exemple de Mes Pensées de Montesquieu », Kodikologie und Paläographie im Digitalen Zeitalter 2 / Codicology and Palaeography in the Digital Age 2. Norderstedt : BoD, 2010, p. 361-374 ; "Œuvres perdues, pratiques oubliées : l'heuristique de la ruine textuelle chez Montesquieu", Que faire avec les ruines?Poétique et politique des vestiges, Presses universitaires de Rennes, 2015, p. 83-93 'Métamorphoses du recueil de notes chez Montesquieu : du prélèvement documentaire au travail auctorial', intervention dans : Launching RASCIO – Workshop (European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement n° 749180) 24 mai 2018, Université Ca' Foscari, Venise (Italie)

20.	**Les Écrits de l'abbé de Saint-Pierre**	Carole Dornier : carole.dornier@unicaen.fr, Claudine Poulouin : Claudine.Poulouin@univ-rouen.fr, Pascal Buelon : pascal.buleon@unicaen.fr	CPER Nom du projet: R25_P7 "Valorisation technologique du Patrimoine" ; Université de Caen projet NUMNIE (STC/SHS) ; Région Normandie Réseaux d'intérêt normands, Contenus et Corpus numériques (CORNUM) Financement CAHIER	Les projets de l'abbé Castel de Saint-Pierre (1658-1743), pour le plus grand bonheur du plus grand nombre, C. Dornier et C. Pouloin (dir.), Caen, Presses universitaires de Caen, 2011 Les idées de l'abbé Castel de Saint-Pierre (1658-1743), «Toutes les parties de la bienfaisance», S. Gregori et P. Oppici (dir.), Macerata, EUM, 2014 (colloque de Macerata, 2011 C. Dornier, « Le progrès de la raison, de l'éloquence et des lettres au service de l'utilité publique : l'abbé de saint-Pierre, Descartes de la politique », Écrire et penser en moderne (1687-1750), C. Bahier-Porte, C. Poulouin (dir.), Paris, Champion, 2015, p. 277-287 « L'abbé de Saint-Pierre : rationalité politique et écriture du système », L'Esprit de système au XVIIIe siècle, Paris, éditions Hermann, "Les collections de la République des lettres", 2017, p. 33-42 5 articles sur Castel de Saint-Pierre dans Christian-Muslim Relations. A Bibliographical History, Volume 13 Western Europe (1700-1800), David Thomas and John Chesworth (dir.), Leyde, Brill, 2019 C. Dornier, « L'abbé Castel de Saint-Pierre : l'homme de lettres, conseiller politique d'une monarchie rationalisée », Lumières et identités, 15e Congrès international sur les Lumières (SIEDS 2019) (Édimbourg, Écosse, 14-19 juillet 2019) C.Dornier, La Monarchie éclairée de l'abbé de Saint-Pierre (1658-1743), Oxford University Studies in the Enlightenment, Liverpool U.P., 2020.
21.	**Nouvelle Édition Numérique de Fac-similés de Référence (Nénufar)**	Agnes Steuckardt : agnes.steuckardt@univ-montp3.fr	Financement CAHIER	
22.	**Corpus 14**	Agnes Steuckardt : agnes.steuckardt@univ-montp3.fr	Financement CAHIER	
23.	**Édition électronique de la correspondance de Flaubert**	Yvan Leclerc : yvan-leclerc@wanadoo.fr	Financement CAHIER	
24.	**Œuvres complètes de D'Alembert**	Alexandre Guilbaud : alexandre.guilbaud@imj-prg.fr, Irene Passeron : irene.passeron@wanadoo.fr	Financement CAHIER	
25.	**Enccre**	Alexandre Guilbaud : alexandre.guilbaud@imj-prg.fr, Irene Passeron : irene.passeron@wanadoo.fr		
26.	**Frénaud numérique**	Marianne Froye : marianne.froye@univ-fcomte.fr	Amorçage : POenum (financement région), Chrysalide (financement université) Financement CAHIER	« Frénaud numérique : une expérience d'encodage plurigénérique », communication dans le cadre du séminaire « Humanités numériques » de la Maison des sciences de l'homme en Bretagne, Rennes, 16 mars 2020 Élaborer, numériser, mettre en ligne et exploiter un corpus d'auteur : exemples de deux cas pratiques en littérature, « Projet Frénaud 2018 : constitution et exploitation d'une base de données », atelier Dahlia (Digital humanities and Cultural Heritage data and knowledge management and analysis), EGC (Association internationale francophone d'extraction et de gestion des connaissances, Metz, 2019, http://dahlia.egc.asso.fr/atelier_DAHLIA2019_actes.pdf, pp. 70-82., « Frénaud numérique : des prémices du projet à sa concrétisation », école de

				printemps de l'IDHN (Institut des humanités numériques), 4 mars 2020, L'Isle-Adam, présentation des enjeux du projet « Frénaud numérique » et de ces enjeux informatiques.
27.	**Fonds numérique Jean-Luc Lagarce**	Pascal Lécroart : pascal.lecroart@univ-fcomte.fr		
28.	**Schola Rhetorica**	Christine NOILLE christine.noille-clauzade@wanadoo.fr	Financement CAHIER	
29.	**Manuscrits de Stendhal**	Françoise Leriche : francoise.leriche@univ-grenoble-alpes.fr	Financement CAHIER	Stendhal, Journaux et Papiers, volume 1 : 1797-1804, éd. C. Meynard, H. de Jacquelot, M-R. Corredor, Grenoble, ELLUG, 2013 ;
30.	**Les deux évasions de Benoîte Groult (E2E)**	Cecile Meynard : cecile.meynard@univ-angers.fr		
31.	**CORR-Proust**	Françoise Leriche : francoise.leriche@univ-grenoble-alpes.fr	ANR-21-CE27-0002-01 (2021-2024) : Corr-Proust_1907-1914 Financement CAHIER	F. Leriche et E. Greslou, « Des éditions imprimées à l'édition numérique de la Correspondance : enjeux éditoriaux, scientifiques, solutions d'encodage », Bulletin d'informations proustiennes, 2019, n° 49, p. 43-55 N. Mauriac Dyer, « Revisiter l'annotation génétique de la Correspondance », Bulletin d'informations proustiennes, 2019, n° 49, p. 77-91 C. Szylowicz, « Identification et référencement des lettres dans l'édition numérique de la Correspondance », Bulletin d'informations proustiennes, 2019, n° 49, p. 57-62 P. Wise, « Jean Bénac dans l'Enfer de la Grande Guerre : une source de Robert de Saint-Loup au front », Quaderni Proustiani, no 12, « Les Enfers de la Recherche », 2018, p. 113-140, URL : http://quaderniproustiani.padovauniversitypress.it/system/files/papers/qp2018-7.pdf P. Wise, « Les poupées russes de l'annotation. Des maîtres aux valets : quelques nouveautés », Bulletin d'Informations proustiennes, 2019, n° 49, p. 63-76 F. Proulx, « Proust's Epistolary Diplomacy », Diplomacy and the Modern Novel : France, Britain, and the Mission of Literature, dir. A. Hepburn et I. Daunais, Univ. of Toronto Press, 2019, p. 137-57 F. Proulx, « Un autre moi-même : Between the Self and the Other in Proust's Correspondence », Encounters in Philosophy, Literature, and the Arts, dir. J. Brillaud et V. Greene, Londres, Bloomsbury Press, 2021, p. 153-164 F. Leriche, « Quels réseaux épistolaires définir dans la correspondance de Proust ? », in Lettres dans la Toile, F. Dubosson éd. (à paraître fin 2021)
32.	**Édition Numérique des Cahiers d'Henri de Régnier (ENCHRE)**	Bernard Roukhomovsky : bernard.roukhomovsky@univ-grenoble-alpes.fr		
33.	**Projet La Réticence**	Brigitte Combe : brigitte.combe@gmail.com	Financement CAHIER	Brigitte Ferrato-Combe, "Les brouillons de La Réticence, terrains d'expérimentations", dans Isabelle Roussel-Gilet et Evelyne Thoizet (dir.), "Jean-Philippe TOUSSAINT en coulisses : making of, expérimentations, décalages", ntre-deux, n° 9, juin 2021, p.
34.	**L'invention du théâtre antique dans l'Europe de la première modernité – Commentaires et paratextes (ITHAC)**	Malika Bastin : malika.bastin@univ-grenoble-alpes.fr, Pascale Paré-Rey : pascale.pare-rey@univ-lyon3.fr	IThAC a bénéficié du soutien de l'IDEX (IRS contrat doctoral) Et du soutien de l'ANR 2019 : https://anr.fr/Projet-ANR-19-CE27-0009 https://ithac.hypotheses.org	M. Bastin-Hammou, « Aemilius Portus, Greek Scholar and Latin Humanist. Some Reflexions on Aemilius Portus's Edition of Aristophanes (1607) », dans Vaios Vaiopoulos, Ioannis Deligiannis & Vasileios Pappas (éd.) Post-Byzantine Latinitas. Latin in post-Byzantine Scholarship (1453-1821), Brepols, Turnhout, 2020, p. 77-92 P. Paré-Rey, « L'œuvre tragique de Sénèque au XVIIIème siècle : lectures, relectures et controverses », Anabases, 33 « La tragédie de Sénèque (XVIIIe-XIXe

				siècles) : éclipse et résistance d'un modèle théâtral » S. Humbert-Mougin (éd.), 2021, p. 55-75. P. Paré-Rey, « Virilité et virginité dans la Médée de Sénèque, dans quelques traductions et illustrations modernes » Actes du colloque « SENECA TRAGICVS : VIR, VIS, VIOLENTIA, VIRTUS, VIRAGO. La virilité et ses déclinaisons dans le théâtre de Sénèque et chez ses émules de Mussato à nos jours », organisé en mars 2019 par AMU, Presses Universitaires de Provence, collection « Textuelles Théâtre », sous presse. Ferrand, « Fabrique de l'argumentum dans les premières éditions de Plaute et de Térence », Pensée et pratique de l'intrigue comique (France-Italie, XVIe-XVIIIe siècles), éd. C. Deloince-Louette et J.-Y. Vialleton, Fabula / Les colloques, 2020 (https://www.fabula.org/colloques/document6523.php) M. Ferrand, « Lire Plaute en temps de guerre. Les leçons inaugurales de Jean Passerat au Collège royal (1572-1597) », Lectures du théâtre français des XVIe et XVIIe siècles, éd. S. Berrégard, Strasbourg, Presses universitaires de Strasbourg, à paraître
35.	**Tacitus On Line**	Isabelle Cogitore : isabelle.cogitore@univ-grenoble-alpes.fr	Financement CAHIER	
36.	**Archives des Traducteurs et des Écrivains de la Littérature Italienne : Éditions et Recherches (ATELIER)**	Filippo Fonio : filippo.fonio@univ-grenoble-alpes.fr	Pas encore, plusieurs tentatives n'ont pas donné de résultat positif pour l'instant	
37.	**« Testaments de Poilus » – Une plateforme de transcription collaborative au service du patrimoine manuscrit**	Emmanuelle De Champs : edechamp@cyu.fr, Florence Clavaud : florence.clavaud@culture.gouv.fr	L'adhésion à Cahier est venue après le financement ANR	
38.	**Écritures savantes au siècle des Lumières. La correspondance et les carnets de visiteurs de Jean-François Séguier**	Emmanuelle Chapron : emmanuelle.chapron@univ-amu.fr	Financement CAHIER	E. Chapron, F. Pugnière (dir.), Écriture épistolaire et production des savoirs au XVIIIe siècle. Les réseaux de Jean-François Séguier, Classiques Garnier, pp.9-19, 2019, Les Méditerranées, 978-2-406-08359-7. ⟨10.15122/isbn.978-2-406-08359-7.p.0009⟩. ⟨hal-02542051⟩
39.	**Création de corpus de livrets d'opéra sous l'ancien régime, l'Académie royale de musique de Paris (1671-1791)**	Margareta Kastberg : margareta.kastberg@univ-fcomte.fr	Financement CAHIER	La langue de la tragédie en musique, France 1674-1732, Honoré Champion, Collection « Lettres numériques, Champion, Paris, en cours de publication, Jacquot Sébastien, Kastberg Sjöblom Margareta, « Le livret d'opéra : établissement et exploration textométrique d'un corpus patrimonial de l'époque classique » in (éds. D. Mayaffre, C. Poudat, L. Vanni) JADT 2016 Proceedings, Presses de Université de Nice-Sophia Antipolis, 2016. http://lexicometrica.univ-paris3.fr/jadt/jadt2016/es Numériques », Paris, en cours de publication

40.	**BIBLINDEX**	Laurence Mellerin : laurence.mellerin@mom.fr		M. Büchler, L. Mellerin (éd.), Computer-Aided Processing of Intertextuality in Ancient Languages, (JDMDH, 2017): https://jdmdh.episciences.org/page/intertextuality-in-ancient-languages -- L. Mellerin (éd.), Le puits des eaux vives, Cahiers de BiblIndex 3, Coll. "Biblia Patristica" 22, Strasbourg 2021 : http://www.brepols.net/Pages/ShowProduct.aspx?prod_id=IS-9782503596174-1 --
41.	**Projet E-STAMPAGES**	Michele Brunet : michele.brunet@univ-lyon2.fr		
42.	**Mauriac en ligne**	Jessica de Bideran : jessica.de-bideran@u-bordeaux-montaigne.fr; Caroline Casseville : caroline.casseville@u-bordeaux-montaigne.fr, Philippe Baudorre : philippe.baudorre@u-bordeaux-montaigne.fr		Jessica de Bideran. Numérisation et extension du patrimoine littéraire. Réflexions à propos de « Mauriac en ligne ». Fabienne Henryot. La fabrique du patrimoine écrit Objets, acteurs, usages sociaux, Presses de l'enssib, 2021, p.115-126, 2020, ⟨10.4000/books.pressesenssib.10587⟩. ⟨hal-02470121⟩ Philippe Baudorre, Jessica de Bideran. Mauriac en une, Mauriac en livre, Mauriac en ligne. Réflexion sur les dispositifs éditoriaux. Humanités numériques, Bruxelles: Humanistica, 2020, ⟨10.4000/revuehn.325⟩. ⟨hal-03196099⟩ Jessica de Bideran. Du fragment daté au corpus patrimonialisé : Numérisation et muséalisation de l'article de presse mauriacien. Etudes digitales, Classiques Garnier, 2016, Le texte à venir, pp.125-142. ⟨hal-01855780⟩
43.	**Correspondance électronique d'Émile Zola (CorrELEZ)**	Jean-Sebastien Macke : jean-sebastien.macke@cnrs.fr		Dossier « Naturalismes du monde », Les Cahiers naturalistes, 2020.
44.	**Thresors de la Renaissance**	Anne Reach Ngo : anne.reachngo@yahoo.fr	IUF	AAnne Réach-Ngô. Le site des Thresors de la Renaissance : De la bibliothèque numérique du chercheur à la bibliothèque imaginaire de l'utilisateur. Réforme, Humanisme, Renaissance, Association d'Études sur la Renaissance, l'Humanisme et la Réforme, 2019, Penser, décrire, communiquer. Les bibliothèques de la Renaissance aujourd'hui, 1 (88), pp.149-179. ⟨10.3917/rhren.088.0149⟩. ⟨halshs-02882760⟩ : https://www.cairn.info/revue-reforme-humanisme-renaissance-2019-1-page-149.htm
45.	**Natale Conti, Mythologia, 1567-1627**	Celine Bohnert : celine.bohnert@univ-reims.fr	IUF Financement CAHIER	
46.	**« Renan Source » Une édition génétique numérique des manuscrits d'Ernest Renan**	Domenico Paone : domenico.paone@ens.fr		D. Paone, M. C. Sabouret, « La transmission des archives renaniennes de Corrie Siohan à Renan Source », Études Renaniennes, n. 120, décembre 2020, p. 29-42. -- D. Paone, « Les archives à l'ère du numérique : le projet Renan Source », Colligere. Le carnet des bibliothèques et archives du Collège de France, 2019 (https://archibibscdf.hypotheses.org/3982) -- D. Paone, « Les archives, de l'analogique au numérique : le projet Renan Source », in D. Gambarara, F. Reboul (dir.), Travaux des colloques Le cours de linguistique générale, 1916-2016. L'émergence, le devenir, Cercle Ferdinand de Saussure, 2018 (https://www.clg2016.org/documents/CLG2016-Paone.pdf)
47.	**Archives Marguerite Audoux**	Bernard-Marie Garreau : bernard-marie.garreau@wanadoo.fr		
48.	**POR FAVOR**	Ludivine Thouverez : ludivine.thouverez@univ-poitiers.fr		Ludivine THOUVEREZ, "Humor político en la Transición: las juntas militares del Cono Sur vistas por los humoristas de Por Favor". Atlante, Numéro 13-Automne 2020. URL: https://atlante.univ-lille.fr/13-humor-politico-en-la-transicion-las-juntas-

				militares-del-cono-sur-vistas-por-los-humoristas-de-por-favor.html Ludivine Thouverez,« Le dessin de presse face à la violence terroriste de l'ETA (1974-2004) : contre-pouvoir ou outil de propagande politique ? ». In C. Decobert, M. S. Rodriguez, Construction et déconstruction du politique par les médias européens depuis 1975, à paraître en 2021.
49.	**Sociorama. Littérature panoramique internationale du XIXe siècle**	Nathalie Preiss : blaguezac@wanadoo.fr, Valerie Stienon : valerie.stienon@univ-paris13.fr	Financement CAHIER	
50.	**Les Archives d'Augustin Thierry (ArchAT)**	Aude Deruelle : aude.deruelle@univ-orleans.fr		
51.	**Base Louis Meigret**	Cendrine Pagani : cendrine.pagani@gmail.com		
52.	**TransDiary-TEI**	Regis Schlagdenhauffen : regis.schlagdenhauffen@ehess.fr , rschlagd@ehess.fr		
53.	**eRabbinica**	Daniel Stoekl Ben Ezra : daniel.stoekl@ephe.psl.eu		Daniel STOEKL BEN EZRA. (2021). Medieval Hebrew manuscripts version 1.0. Zenodo. https://doi.org/10.5281/zenodo.5468286. Daniel STOEKL BEN EZRA. (2021). Medieval Hebrew manuscripts in Sephardi bookhand version 1.0. Zenodo.— https://doi.org/10.5281/zenodo.5468665 Daniel STOEKL BEN EZRA. (2021). Medieval Hebrew manuscripts in Italian bookhand version 1.0. Zenodo. https://doi.org/10.5281/zenodo.5468573 Daniel STOEKL BEN EZRA. (2021). Medieval Hebrew manuscripts in Ashkenazi bookhand. Zenodo. https://doi.org/10.5281/zenodo.5468478 Daniel Stökl Ben Ezra, Bronson Brown-DeVost, Pawel Jablonski, Benjamin Kiessling, Elena Lolli, Hayim Lapin, "BiblIA – a General Model for Medieval Hebrew Manuscripts and an Open Annotated Dataset" HIP@ICDAR 2021. Daniel Stökl Ben Ezra, Bronson Brown-DeVost and Pawel Jablonski "Exploiting Insertion Symbols for Marginal Additions in the Recognition Process to Establish Reading Order" IWCP@ICDAR 2021 Daniel Stökl Ben Ezra, Hayim Lapin, "Z-profile: Holistic Preprocessing Applied to Hebrew Manuscripts for HTR with Ocropy and Kraken" Manuscript Cultures 15 (2021) 25-36. T. Kuflik, M. Lavee, D. Stökl Ben Ezra, A. Ohali, V. Raziel-Kretzmer, U. Schor, A. Wecker, E. Lolli, P. Signoret, 'Tikkoun Sofrim – Combining HTR and Crowdsourcing for Automated Transcription of Hebrew Medieval Manuscripts', DH2019 Daniel Stökl Ben Ezra, « The Mishnah into French : translation issues », in E. Bar-Asher Siegal & A. Koller (eds.), Studies in Mishnaic Hebrew and Related Dialects : Proceedings of the Yale Symposium, May 2014 (New Haven: Program in Judaic Studies Yale University & Jerusalem: Magnes 2018) 349-367. M. Seuret, D. Stökl Ben Ezra, M. Liwicki, Robust Heartbeat-based Line Segmentation Methods for Regular Texts and Paratextual Elements. HIP@ICDAR 2017: 71-76.— G. Sadeh, L. Wolf, T. Hassner, D. Stökl Ben Ezra & N. Dershowitz « Viral Transcript Alignment » ICDAR 2015: 711-5.
54.	**MERCURE**	Anne Piejus : anne.piejus@cnrs.fr		

	GALANT			
55.	**Antonomaz**	Karine Abiven : karine.abiven@sorbonne-universite.fr		2021a : Karine Abiven, Gaël Lejeune, “Des données au corpus : l’exploitation numérique des mazarinades”, Dix ans de Corpus d’auteurs, Editions des Archives contemporaines, accepté 2021b : Karine Abiven, Jean-Baptiste Tanguy et Gaël Lejeune, “Exploiter en corpus des données textuelles océrisées : l’écriture burlesque de la Fronde (1648-1652)”, accepté, revue Humanités numériques, n°4 - Humanistica. 2021 c, K. Abiven, A. Bartz, J.-B. Tanguy, G. Lejeune, "Vers une collection numérique des imprimés de la Fronde : exposer, fouiller, relier des mazarinades", dans A. Réach-Ngô (dir.), "Circulation des écrits littéraires de la Première Modernité et Humanités numériques", Le Verger.
56.	**Inventaire Condorcet**	Nicolas Rieucau : niko99@wanadoo.fr	ANR CONDOR (2016-2020) https://anr.fr/Project-ANR-16-CE27-0003 Financement CAHIER	https://www.inventaire-condorcet.com/Actualites https://cahier.hypotheses.org/guide-correspondance
57.	**Privilèges de librairie en France à l’époque moderne (XVIe XVIIe siècles)**	Edwige Keller : edwige.keller@univ-lyon2.fr	Soutien financier du GIS Institut du Genre dans le cadre de l’APP 2018-2019	« Equality in the printed book: the case of book privileges in France in the seventeenth century », Derval Conroy (éd.), Towards an Equality of the Sexes in Early Modern France (1600-1700), SRS Renaissance and Early Modern Worlds of Knowledge series, Routledge Studies, London, 2021, p. 184-209. URL : https://www.routledge.com/Towards-an-Equality-of-the-Sexes-in-Early-Modern-France/Conroy/p/book/9780367224929 (avec Miriam Speyer) : « Les privilèges d’impression du recueil Barbin et des recueils de vers polygraphiques au XVIIe siècle. Législation et pratiques éditoriales », Le Recueil Barbin (1692), « Une histoire de la poésie par les ouvrages même des poètes », Actes du Colloque international IHRIM, Lyon, 3-4 mai 2018 Mathilde Bombart, Maxime Cartron et Michèle Rosellini (dir.), Cahiers du Gadges, Pratiques & formes littéraires 16-18, 2020, p. 21-60, [En ligne], mis en ligne le 26 novembre 2019, URL : https://publications-prairial.fr/pratiques-et-formes-litteraires/index.php?id=79
58.	**Espace Afrique-Caraïbe**	RIFFARD Claire, claire.riffard@cnrs.fr	Financement CAHIER	

c) Liste des projets associés au consortium

Il s’agit soit de projets qui ont été membres de CAHIER et qui, ayant terminé leurs travaux, sont devenus membres associés, soit de projets qui ont préféré le statut d’associé pour suivre les travaux du consortium sans contribuer directement aux travaux (groupes de travail, etc.)

	Nom du projet	Nom et mail du responsable scientifique	Site internet du projet
1.	**Atelier André Breton**	Constance KREBS constance@giantchair.com	http://andrebreton.fr
2.	**Encyclopédie des Littératures en Langues africaines (ELLAf)**	Ursula BAUMGARDT	http://ellaf.huma-num.fr/

3.	**EDITEF**	Chiara LASTRAIOLI	http://www.editef.univ-tours.fr/
4.	**Corpus Descartes (ProDescartes)**	Vincent Carraud	http://www.unicaen.fr/puc/sources/prodescartes/accueil
5.	**Section de l'Humanisme**	Marie-Elisabeth Boutroue	http://www.irht.cnrs.fr/
6.	**Œuvres complètes de Giono**	Véronique Magri	
7.	**Corpus Boris Tchitchérine**	Sylvie Martin	http://wiki-tchitcherine.ens-lyon.fr/index.php/Accueil
8.	**Routes du livre italien ancien en Normandie**	Silvia Fabrizio-Costa	http://www.unicaen.fr/recherche/mrsh/rdli/
9.	**Cours d'Antoine Desgodets**	Robert Carvais, Emmanuel Château	http://www.desgodets.net/
10.	**Les guides de Paris**	Marianne Cojannot-Le Blanc, Emmanuel Château	http://passes-present.eu/fr/guides-de-paris-les-historiens-des-arts-et-les-corpus-numeriques-363
11.	**Correspondance Nalèche**	Odile Gaultier-Voituriez	https://naleche.hypotheses.org/
12.	**Ponge**	Pauline Flepp	http://obvil.paris-sorbonne.fr/corpus/ponge/
13.	**Bibliothèques Privées à l'Âge Moderne (BIPrAM)**	Raphaële Mouren	http://www.enssib.fr/les-poles-thematiques/histoire-des-bibliotheques/dossiers/e-collections-et-collectionneurs/bipram
14.	**Édition électronique de contes populaires français**	Anne Garcia-Fernandez	
15.	**Journaux d'Alexandre Dumas**	Sarah Mombert	http://alexandredumas.org/
16.	**Archive Numérique Desanti**	David Wittmann, Maud Ingarao	http://archive.desanti.huma-num.fr/
17.	**NaviLog (AnaLog + VariaLog)**	Marie-Hélène Lay	http://forell.labo.univ-poitiers.fr/
18.	**Marivaux sur la scène européenne : archives numériques multimédia**	Paola Ranzini	
19.	**Notes de cours de l'ENS**	Emmanuelle Sordet, Charlotte Dessaint	
20.	**NUMERISLAV : Edition numérique des archives scientifiques de l'Institut d'Etudes slaves**	Archaimbault Sylvie	
21.	**Chispa : Création d'outils pour l'exploitation numérique de corpus de manuscrits HISPAniques**	Fatiha Idmhand	https://chispa.hypotheses.org/ ; https://guarnido.nakalona.fr/ ; https://molina.nakalona.fr/ Projet terminé en 2017. La mise à jour des dépôts est en cours sur Nakala. Les collections Collection José Mora Guarnido ID : 10.34847/nkl.8f18ifc3 https://nakala.fr/collection/10.34847/nkl.e57911x0 à peu près 1700 fichiers de données pour 15000 images Collection Carlos Denis Molina ID : 10.34847/nkl.e57911x0 https://nakala.fr/collection/10.34847/nkl.8f18ifc3 à peu près 1700 fichiers de données pour 15000 images
22.	**Holographical-Lee (HoL)**	Sophie Geoffroy	
23.	**READ-IT**	Brigitte OUVRY-VIAL, François VIGNALE	https://readit-project.eu/
24.	**Le Règne d'Astrée**	Delphine Denis	http://astree.huma-num.fr/
25.	**Hyperdonat, édition numérique de commentaires anciens**	Christian Nicolas	http://hyperdonat.huma-num.fr/editions/index.html
26.	**Édition génétique numérique du « Robinson » de Paul Valéry**	Franz Johansson	http://www.item.ens.fr/index.php?id=13898
27.	**Édition de la correspondance du géomètre Gaspard Monge 1795-1799**	Marie Dupond	http://eman-archives.org/monge/
28.	**DYRIN**	Thierry Buquet	http://www.unicaen.fr/craham/spip.php?article1159

14. Liste des annexes et annexes

a) Annexe n°1 : Cartes du réseau depuis 2014 ….

b) Annexe n°2 : Extraits du Memo mynkl

c) Annexe n°3a : Plan de gestion des données de CAHIER
Annexe n°3b : Plans de gestion des données de dix projets membres

d) Annexe n°4 : Introduction et sommaire du livre « Dix ans de Corpus d'auteurs »

e) Annexe n°5 : Projet de consortium de réseau « REST CAHIER »

f) Annexe n°6 : Projet de consortium dédié aux outils « OLIO »

g) Annexe n°7 : Projet de consortium provisoirement appelé « CAHIER après CAHIER » (lettre d'intention)

a) Annexe n°1 : Cartes du réseau depuis 2014

Carte du réseau en 2014

Carte du réseau en 2017

Carte du réseau en 2021

b) Annexe n°2 : Extraits du Memo mynkl

Le mémorandum complet a été mis en ligne sur HAL : https://halshs.archives-ouvertes.fr/halshs-03408209 . Les pages qui suivent constituent une sélection issue de ce mémo.

Il s'inspire des guides pour la FAIRisation des données également accessibles en ligne sur HAL : Voir « Guides pour la FAIRisation des données » : V1 https://halshs.archives-ouvertes.fr/halshs-02889777 et V2 https://halshs.archives-ouvertes.fr/halshs-03037748

Les tutoriels décrivant l'utilisation de l'outil sont en ligne sur : http://mynakala.huma-num.fr/docs

Les codes de l'outil seront déposés sur un git (gitlab d'Huma-Num) à la fin de l'année 2021.

FAIRiser vos données : mémorandum

consortium cahier

Andrés ECHAVARRÍA
Ala Eddine LAOUIR
Fatiha IDMHAND
Ioana GALLERON
Laurent PASSION

Guide pour la FAIRisation des données des corpus d'auteurs

Les principes FAIR (Findability, Accessibility, Interoperability and Reusability) définissent un ensemble minimal de principes qui permettent aux machines et aux humains de trouver, d'accéder, d'interopérer et de réutiliser les données et métadonnées de recherche. Les principes FAIR doivent être considérés comme de bonnes pratiques destinées à faciliter la réutilisation des données et les résultats de la recherche.

Le consortium CAHIER, et ses membres, recommandent et mettent en œuvre ces bonnes pratiques.

Dans le cycle de vie d'un projet, le dépôt des données (4) en vue de leur archivage arrive généralement en fin de chaîne, après la collecte (1), le traitement (2) et l'analyse (3). Le dépôt contribue à rendre pérennes et à mettre en valeur les résultats. Pour chaque phase du cycle de vie d'un projet, l'infrastructure HumaNum propose différents outils.

Le consortium CAHIER utilise principalement Nakala pour stocker ses données afin qu'elles soient trouvables, accessibles, interopérables et réutilisables. Ce petit guide décrit le processus à mettre en œuvre pour rendre les données du Consortium CAHIER « FAIR ». Celui-ci est organisé en quatre étapes :

1° Évaluer le degré d'ouverture de ses projets.

2° Confronter ses métadonnées aux attentes du consortium.

3° Compléter et corriger les métadonnées si nécessaire.

4° Déposer sur Nakala, obtenir un identifiant pérenne et l'associer aux documents publiés sur son propre site web (ou sur un site web institutionnel).

1° Mon projet est-il FAIR ?

Pour être FAIR, les données produites dans le cadre du projet doivent être trouvables (Findable), accessibles et téléchargeables librement (Accessible), interopérables (Interoperable) et réutilisables (Reusable). Voici quelques éléments concrets liés à cet objectif, et auxquels le dépôt sur Nakala apporte une réponse.

Services Huma-Num
par étapes

on des données

F

Pour que les données et métadonnées soient trouvables, il faut les pourvoir d'un identifiant pérenne et unique au niveau mondial de type DOI, Handle ou ARK par exemple. Il convient également d'être certain que ses métadonnées puissent être moissonnées par les agrégateurs des grandes bibliothèques numériques.

Si votre institution de rattachement ou un partenaire clé de votre projet (bibliothèque, service d'archives ou informatique, etc.) offre de tels services et qu'elle donne à vos données et métadonnées un identifiant de type DOI, Handle ou ARK, alors vos données sont trouvables (Findable) et la première condition est remplie. Toutefois, il reste utile d'effectuer le dépôt sur Nakala et d'équiper vos données de deux identifiants car l'outil offre une série de services complémentaires.

Dans la majorité des cas, les institutions hébergeant des projets (ou les sites web de projets) ne proposent pas de DOI, Handle ou ARK. Ainsi, même si vos données sont visibles via le site web hébergé sur le serveur de votre institution, vos données ne sont pas FAIR en l'absence d'identifiants uniques et pérennes. Nous vous recommandons de doubler cette publication numérique d'un dépôt sur un entrepôt de données comme Nakala. C'est même une nécessité de la science ouverte.

A

Un site web ne répond que partiellement à la question de l'Accessibilité et cela, même si le site est « hébergé par » ou « chez » Huma-Num car vos données ne sont pas, pour autant, accessibles au même niveau que des données versées dans un entrepôt ouvert. Force est également de constater que dans de nombreux cas, une grande partie des ressources n'est pas accessible sur le site web : conditionner l'accès aux données par une demande d'inscription préalable ou imposer la consultation par l'intermédiaire exclusif d'une interface va à l'encontre de l'objectif de l'Accessibilité.

Si la non-exposition des données peut être nécessaire dans la phase de préparation des documents, il est essentiel qu'au terme d'un projet, une partie significative des données soit accessible. Lorsque des images ne peuvent être rendues, un embargo est possible sur ces données, toutefois, cette restriction peut ne pas concerner les métadonnées ou les transcriptions. Des données peuvent donc être mises en ligne.

Les membres de CAHIER sont invités à engager un dialogue avec les auteurs,

leurs ayants-droits, ou avec les institutions qui leur fournissent les images : une explication de l'intérêt de l'open access permet parfois d'obtenir leur accord pour une plus ample exposition des données.

Le consortium CAHIER utilise Nakala pour remplir l'ensemble des conditions FAIR pour ses données.

Guide pour la FAIRisatio

I

La préparation des données selon des pratiques et des référentiels mondialement connus et partagés est essentielle pour assurer leur interopérabilité. Dans votre propre projet, vous avez probablement déjà veillé à cet aspect, par exemple en utilisant un CMS pour saisir et exposer vos données. Le dépôt sur Nakala vient dans le prolongement de cet effort : après l'exposition, il vous reste à stocker vos données.

R

La réutilisabilité a concerné, jusqu'à présent, la qualité du fichier numérique et son format : ouvert, standardisé, etc. Néanmoins, les conditions de cette réutilisabilité ont été moins étudiées et pensées. Le dépôt d'un grand volume de données sur une même plateforme est à même de stimuler cette réutilisabilité, en donnant plus de visibilité à votre projet.

2° Confronter ses métadonnées aux attentes du consortium

Le RDA FAIR Data Maturity Model Working Group a publié en avril 2020 un guide comportant des indicateurs des données FAIR. Après avoir consulté ces indicateurs, et en vue d'harmoniser les pratiques de dépôt, le consortium a défini un modèle minimal commun des métadonnées qui doivent accompagner les fichiers produits dans le cadre du consortium.

Ce socle commun de métadonnées descriptives peut être étendu ad libitum mais il est utile de faire converger les pratiques d'encodage.

Il est important d'aborder la question de la structuration des métadonnées : l'identification du format utilisé pour les structurer permettra d'organiser le dépôt en masse des données dans l'entrepôt.

Le consortium CAHIER a développé une application qui communique avec la plateforme Nakala à travers son API.

FAIR Data Maturity Model: specification and guidelines

Plus d'information sur Nakala

Au mois de décembre 2020, la nouvelle version de la plateforme de stockage de données Nakala a été présentée par Huma-Num. Cette nouvelle version a intégré une interface utilisateur plus facile à utiliser, ce qui a été très apprécié par les utilisateurs. De plus, l'API de Nakala est devenue plus riche, facilitant ainsi l'intégration de toutes les fonctionnalités de Nakala dans d'autres outils.

Nakala utilise un triplestore pour enregistrer les données fournies par les utilisateurs. Cette méthode de sauvegarde facilite la publication des données sur le Web et dans le monde de l'Open Data (ou Linked Data). Ces services seront très intéressants pour les chercheurs à l'avenir.

Nakala n'a pas encore développé le service de dépôt en masse des données. Comme ce manque pouvait limiter les projets de CAHIER, le consortium a créé une application web, *mynkl*, qui communique avec l'API de Nakala et qui facilite le dépôt de grands volumes de données. L'appilcation *mynkl* comble, provisoirement, un manque.

Plus d'information sur ISIDORE

mynkl

L'outil mynkl

Services fournis par nakala

Interface Nakala

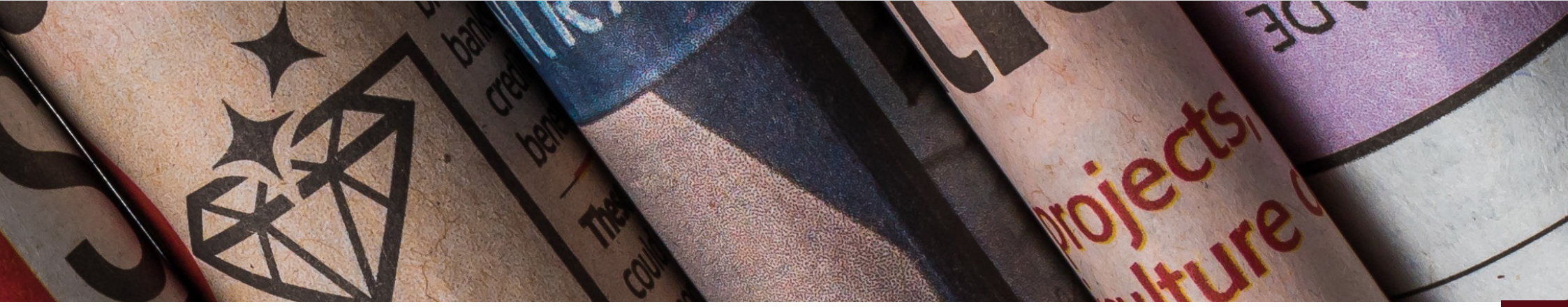
Base de données RDF ou Triplestore

Point d'accès SPARQL Nakala

API Nakala

Chercheur

Outil MyNkl développé par consortium cahier

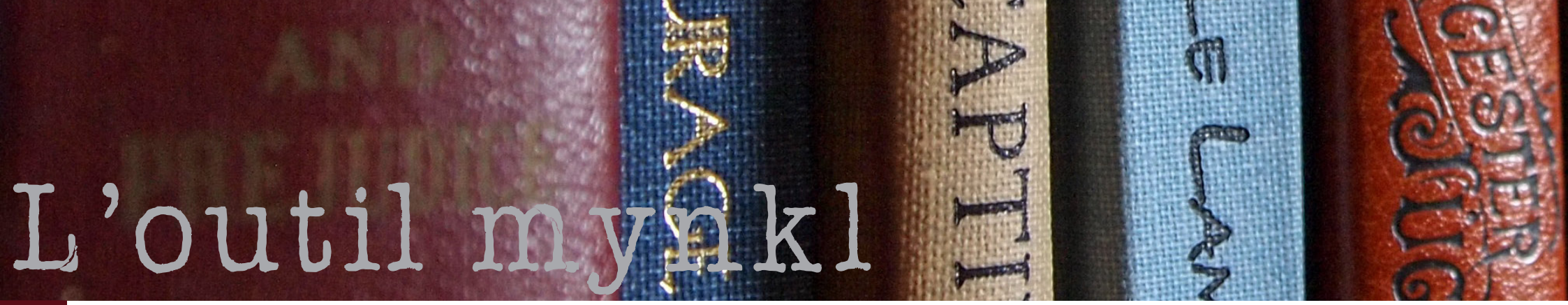


Quelles métadonnées utilise l'outil mynkl ?

Les projets du Consortium CAHIER ont principalement utilisé le Dubln Core, le XML-TEI ou des tableurs (csv) pour structurer leurs métadonnées.
L'entrepôt de stockage de données Nakala utilise le standard DCTerms pour structurer les métadonnées des objets déposés.
Pour faciliter le dépôt des données, le Consortium a puisé dans les teiHeader et les balises Dublin Core des projets (ou dans les tableurs), les informations nécessaires au dépôt dans Nakala.

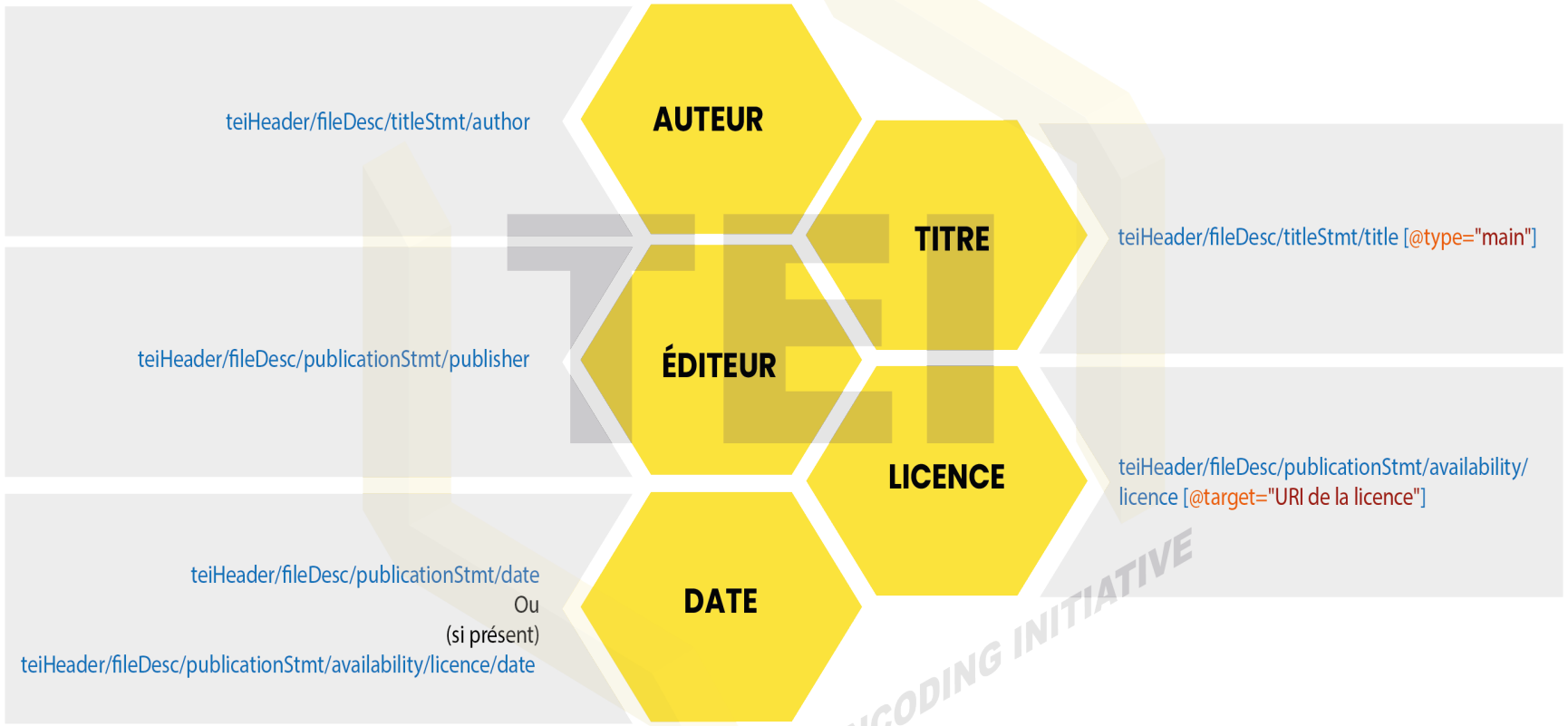
API Nakala

Point d'accès SPARQL Nakala

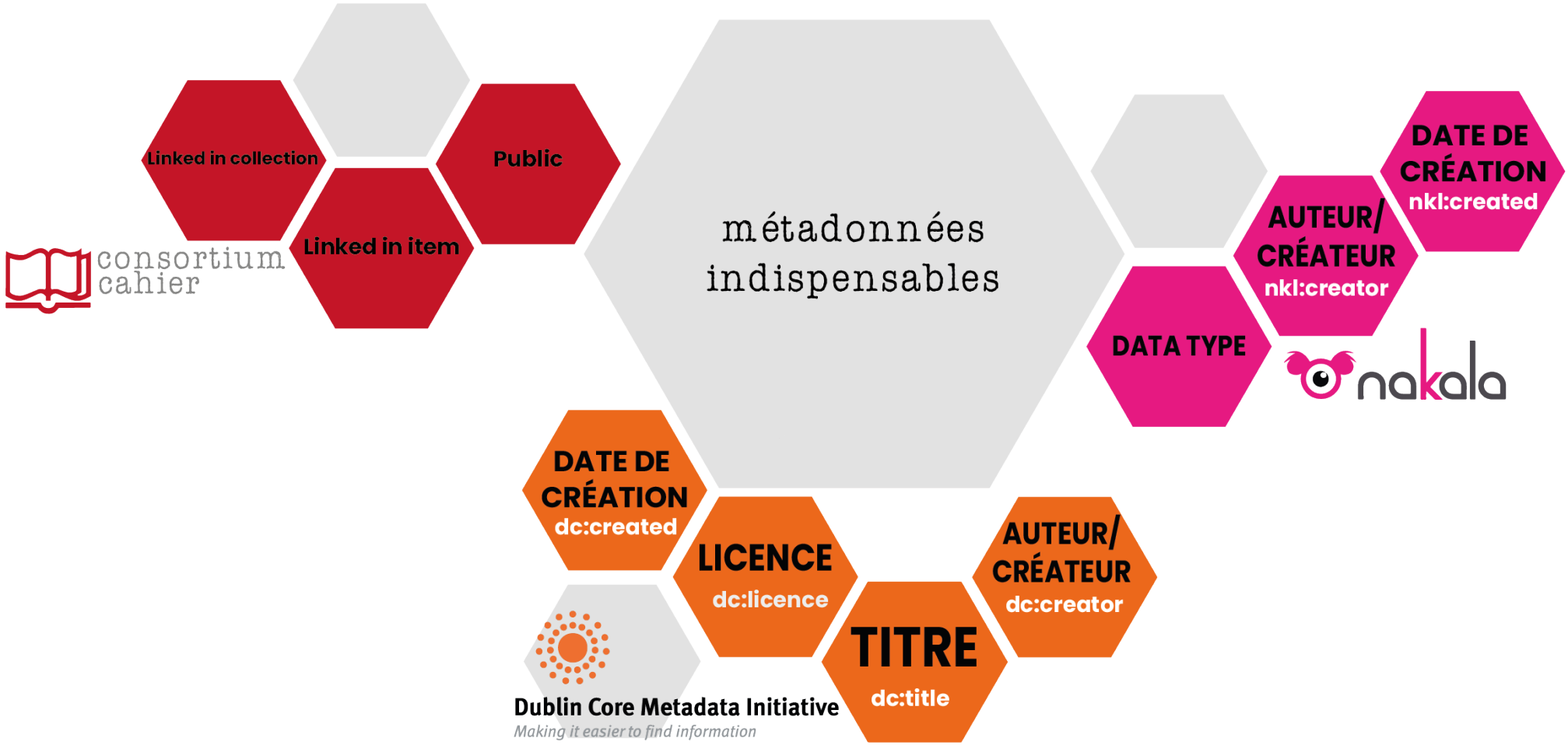


L'outil mynkl

L'outil mynkl organise le mapping de vos données en s'appuyant sur la sélection de balises TEI suivantes :

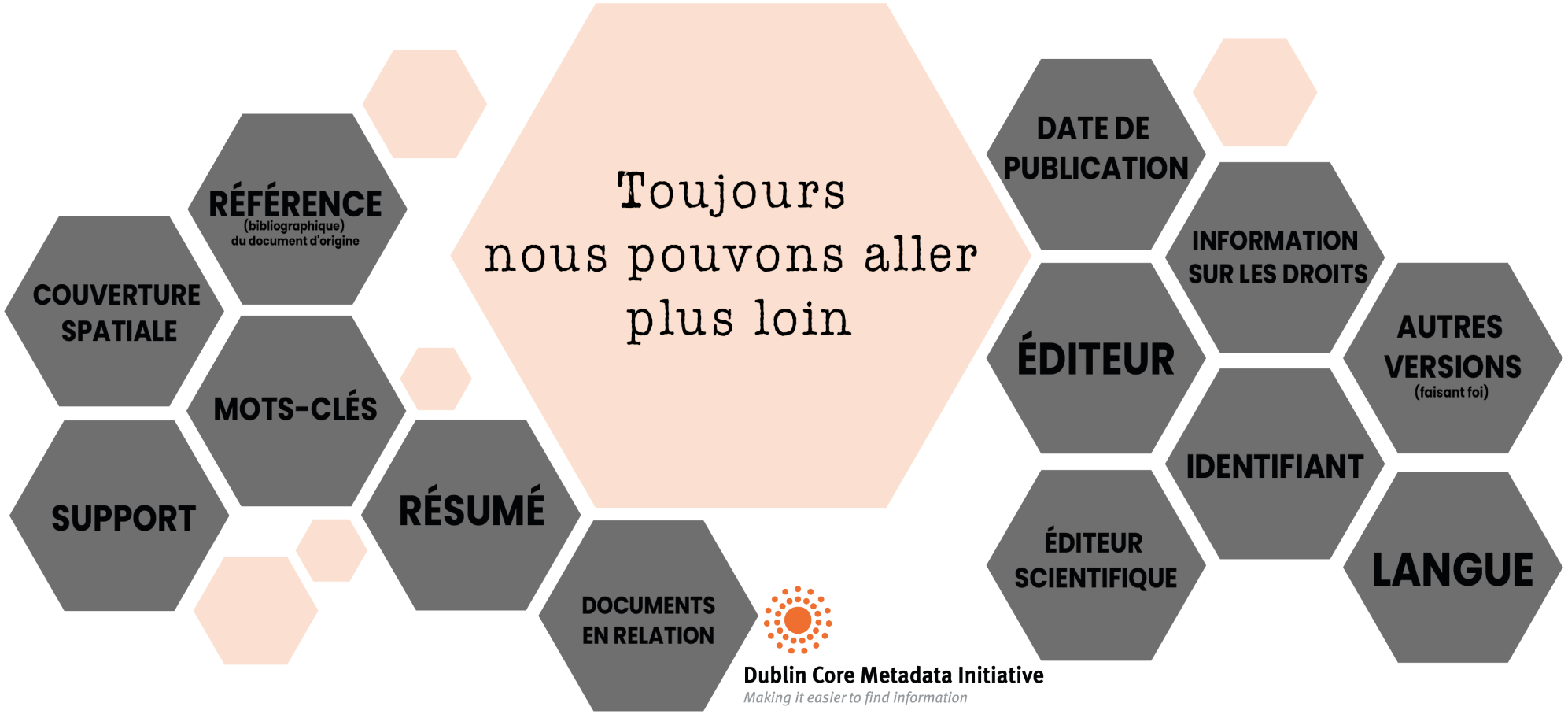


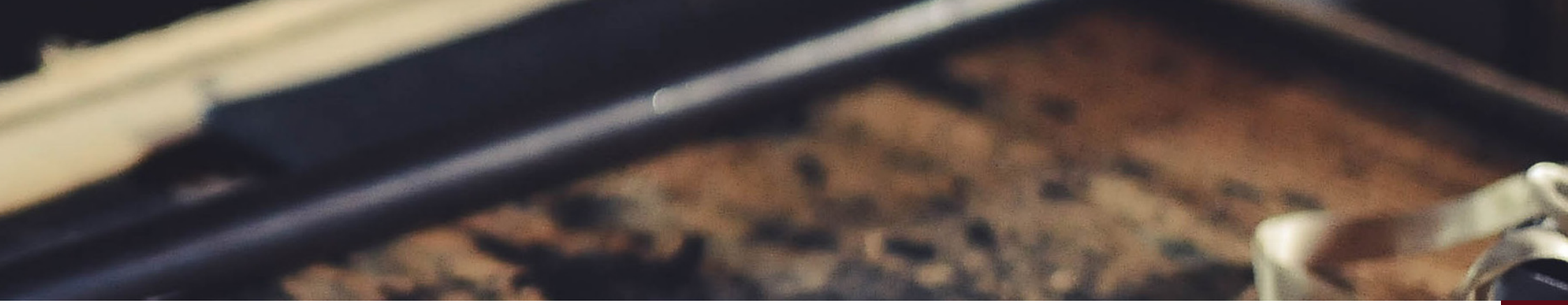
L'outil mynkl organise le mapping de vos données en s'appuyant sur la sélection des champs DC suivants :



L'outil mynkl

Si vous utilisez un tableur et les métadonnées suivantes, l'outil mynkl organise le mapping de vos données





Si vous utilisez un tableur, voici une liste de champs de métadonnées que nous vous suggérons

Linked in collection
Public
Linked in item
Format dc-rdf:format
DataType
Titre dc-rdf:title
Créateur nkl:creator
Créateur dc-rdf:creator
Couverture spatiale dc-rdf:spatial
Date de création dc-rdf:created
Date de création nkl:created
Licence : dc-rdf-license
Langue dc-rdf:language
Sujet dc-rdf:subject
Description dc-rdf:description
Contributeur dc-rdf:contributor
Éditeur dc-rdf:publisher
Ayants droit dc-rdf:rightsHolder
Relation dc-rdf:relation
Résumé dc-rdf:abstract
Date de disponibilité dc-rdf:available
Date de modification dc-rdf:modified
Est une version de dc-rdf:isVersionOf
Support dc-rdf:medium
Référence bibliographique dc-rdf:bibliographicCitation

Les métadonnées

À propos de quelques métadonnées

Titre dc-rdf:title

Titre dc-rdf:title : sert à déterminer le nom donné à la ressource
Exemple 1 : Le Rouge et le violet
Exemple 2 : Le Rouge et le violet : édition électronique
Exemple 3 : Lettre de Pierre à Paul
Exemple 4 : [Ceci est un titre forgé. Le document n'a pas de titre, j'en donne un et je le mets entre crochets. L'usage veut que l'on reprenne, normalement, la première phrase du manuscrit ou de la lettre dans le cas de correspondances.]

Langue dc-rdf:language

Ce champ indique la (ou les) langue.s de la ressource. Il existe différentes représentations des langues possibles reconnues par Nakala. On peut retrouver cette liste sur le site de l'API de Nakala à partir du lien lié au QR code Langues Nakala.

Langues Nakala

Public

Il est nécessaire de prêter une attention particulière à ce point, car une fois que les données sont publiques dans Nakala, elles ne peuvent plus être retirées du serveur, à moins d'une communication directe entre les chercheurs et les administrateurs de Nakala.

Lorsque le dépôt est effectué et que la donnée est publique, elle devient citable, c'est pour cela qu'elle ne peut plus être supprimée.

Date de création nkl:created - Date de création dc-rdf:created

Ce champ détermine la date de création de la ressource. La pratique recommandée est de décrire la date, la date/heure ou la période. Il s'agit de la date de création du document source (appelée date première dans le catalogue OAI du consortium CAHIER). La date doit être inscrite suivant la norme ISO 8601, c'est-à-dire AAAA-MM-JJ.

Les métadonnées

DataType

Ce champ sert à déterminer le type de données que le chercheur dépose dans les serveurs de Nakala. Pour bien renseigner cet espace il faut coller dans le champ dédié le lien URI correspondant au type de donnée parmi la liste suivante :

Image	"http://purl.org/coar/resource_type/c_c513"
Vidéo	"http://purl.org/coar/resource_type/c_12ce"
Son	"http://purl.org/coar/resource_type/c_18cc"
Articles de journaux	"http://purl.org/coar/resource_type/c_6501"
Poster dans une conférence	"http://purl.org/coar/resource_type/c_6670"
Objet présenté dans une conférence	"http://purl.org/coar/resource_type/c_c94f"
Objet d'apprentissage	"http://purl.org/coar/resource_type/c_e059"
Ouvrage	"http://purl.org/coar/resource_type/c_2f33"
Carte ou une carte géographique	"http://purl.org/coar/resource_type/c_12cd"
Jeu ou une série de données	"http://purl.org/coar/resource_type/c_ddb1"
Logiciel ou un développement informatique	"http://purl.org/coar/resource_type/c_5ce6"
Ressource qui n'est pas incluse parmi la liste de controlled vocabulaires for repositories (COAR)	"http://purl.org/coar/resource_type/c_1843"

Matériel d'archive	"http://purl.org/library/ArchiveMaterial"
Collection [URI DublinCore]	"http://purl.org/ontology/bibo/Collection"
Bibliographie	"http://purl.org/coar/resource_type/c_86bc"
Séries [URI DublinCore]	"http://purl.org/ontology/bibo/Series"
Note de lecture, une critique de livre ou une recension d'ouvrage	"http://purl.org/coar/resource_type/c_ba08"
Manuscrit	"http://purl.org/coar/resource_type/c_0040"
Lettre	"http://purl.org/coar/resource_type/c_0857"
Rapport	"http://purl.org/coar/resource_type/c_93fc"
Periodique [c'est un concept rendu obsolète par la COAR mais avec une URI existante et utilisé par Nakala pour l'instant	"http://purl.org/coar/resource_type/c_2659"
Prépublication	"http://purl.org/coar/resource_type/c_816b"
Synthèse, un article de synthèse ou une recension	"http://purl.org/coar/resource_type/c_efa0"
Partition de musique	"http://purl.org/coar/resource_type/c_18cw"
Ensemble des données collectées enquête en considérant les complétions effectuées par les participants.	"https://w3id.org/survey-ontology#SurveyDataSet"

Les métadonnées

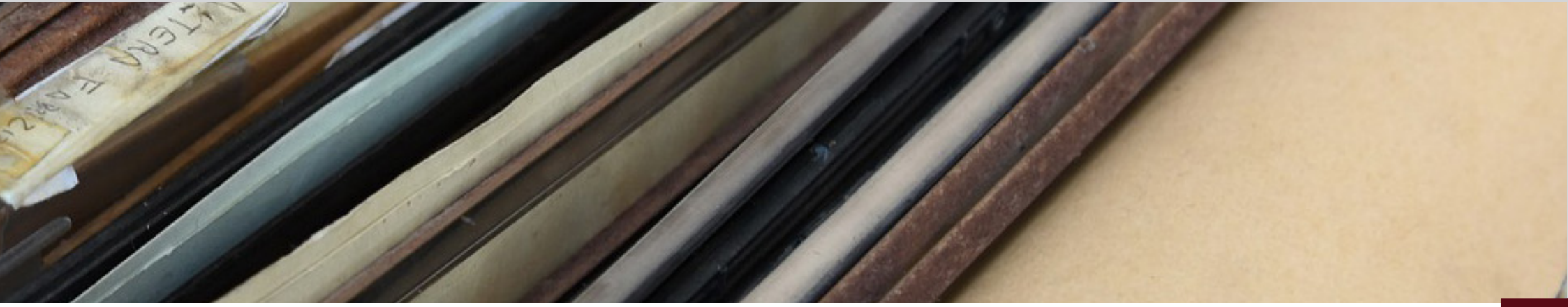
Texte simple	“http://purl.org/coar/resource_type/c_18cf”
Thèse, une mémoire de thèse ou un rapport de thèse.	“http://purl.org/coar/resource_type/c_46ec”
Site web	“http://purl.org/coar/resource_type/c_7ad9”
Data paper ou article sur les données	“http://purl.org/coar/resource_type/c_beb9”
Ressource interactive	“http://purl.org/coar/resource_type/c_e9a0”

Créateur dc-rdf:creator - Créateur nkl:creator

Ce champ détermine l’entité principalement responsable de la fabrication de la ressource. Creator peut comprendre une personne, une organisation ou un service. Les noms des auteurs dans cet espace s’écrivent [prénom][virgule][espace][nom]

Exemple 1 : Jean-Paul, Sartre
Exemple 2 : Emile, Zola
Exemple 3 : Julio, Cortazar

Liste DataTypes API Nakala



Référence bibliographique dc-rdf:bibliographicCitation

Ce champ détermine s'il s'agit d'un livre, d'un article, ou d'une autre ressource documentaire.

Couverture spatiale dc-rdf:spatial

Ce champ détermine les caractéristiques spatiales de la ressource.

Ayants droit dc-rdf:rightsHolder

Ce champ détermine la personne ou l'organisation qui possède ou gère les droits sur la ressource. La pratique recommandée est de faire référence au détenteur des droits avec une URI. Si cela n'est pas possible ou faisable, une valeur littérale qui identifie le titulaire des droits peut être fournie.

Les métadonnées

Licence : dc-rdf-license

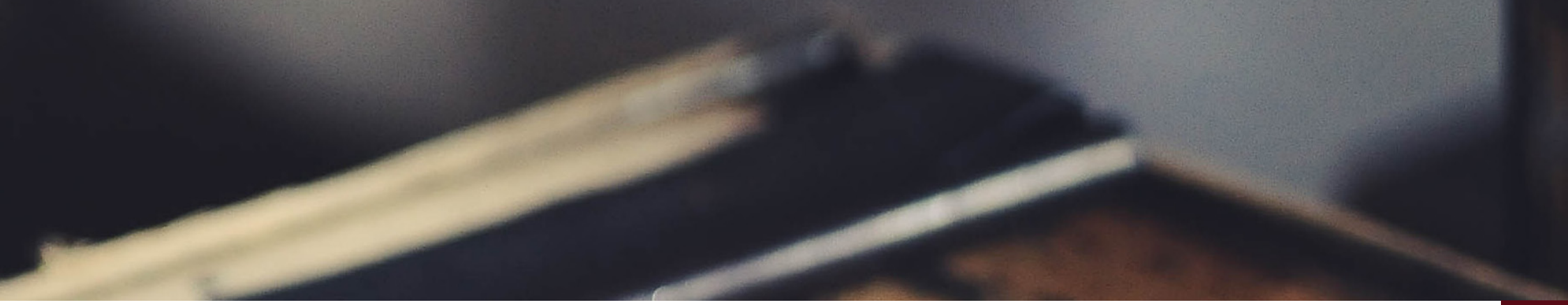
Ce champ détermine le document légal ou le type d'autorisation délivrée avec la ressource. La pratique recommandée est d'identifier la licence avec une URI. Si cela n'est pas possible ou faisable, une valeur littérale qui identifie la licence peut être fournie. Les différentes licences possibles peuvent se retrouver sur le site de l'API de Nakala en suivant le lien lié au QR code Licences Nakala.

Licences Nakala

Source dc-rdf:source

Ce champ indique la référence ou côte de la ressource connexe à partir de laquelle la ressource numérique décrite est dérivée. La pratique recommandée est d'identifier la ressource connexe au moyen d'une chaîne conforme à un système d'identification formel. On mentionnera ici les données bibliographiques du document source (date, lieu et année de publication de la source, côte du document dans l'institution).

Exemple 1 : NAF 10266
Exemple 2 : Cote : NAF 10266



Éditeur dc-rdf:publisher

Ce champ détermine l'entité responsable de la mise à disposition de la ressource.

Exemple 1 : Projet Région n°12345 | Consortium CAHIER TGIR Huma-Num
Exemple 2 : Projet FLG – AAP MSH Centre Sud

Contributeur dc-rdf:contributor

Ce champ détermine l'entité responsable des contributions à la ressource. Les directives relatives à l'utilisation de noms de personnes ou d'organisations en tant que créateurs s'appliquent également aux contributeurs. En général, le nom d'un contributeur doit être utilisé pour indiquer l'entité.

Exemple 1 : Dupont, Jeanne (Professeur des Universités)
Exemple 2 : Dupont, Jeanne (Professeur des Universités) | Itterom, Ocnar (Chercheur CNRS)
Exemple 3 : Ghog, Nav (Critique d'art)

Les métadonnées

Droits dc-rdf:rights

Ce champ détermine les informations sur les droits détenus dans et sur la ressource. En général, les informations comprennent une déclaration sur les divers droits de propriété associés à la ressource, y compris les droits de propriété intellectuelle.

Exemple 1 : Fonds Jean-Charles Carpo – Bibliothèque Jacques Tecuod
Exemple 2 : Archives familiales Jean-Sol Ertrap

Relation dc-rdf:relation

Ce champ désigne une ressource connexe. La pratique recommandée est d'identifier la ressource connexe au moyen d'une URI. Si cela n'est pas possible ou faisable, une chaîne conforme à un système d'identification formel peut être fournie.

Exemple 1 : Le rose et le jaune – Manuscrit 1 | Le rose et le jaune – Manuscrit 3 | Le rose et le jaune – Carnet 1
Exemple 2 : NAF 10266 | NAF 10267 | NAF 10268

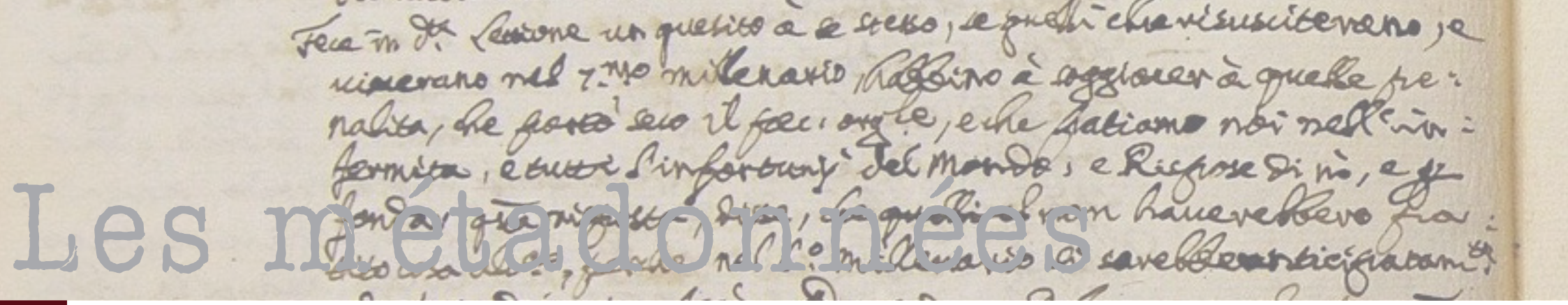
Format dc-rdf:format

Ce champ détermine le format du fichier. La pratique recommandée est d'utiliser un vocabulaire contrôlé lorsqu'il est disponible. Par exemple, pour les formats de fichiers, on peut utiliser la liste des Internet Media Types (MIME).

Résumé dc-rdf:abstract

Ce champ comporte quelques lignes et informations à propos de la ressource. Elles sont destinées au public.

Exemple 1 : Le roman parle de ....
Exemple 2 : Ce document est le premier de...

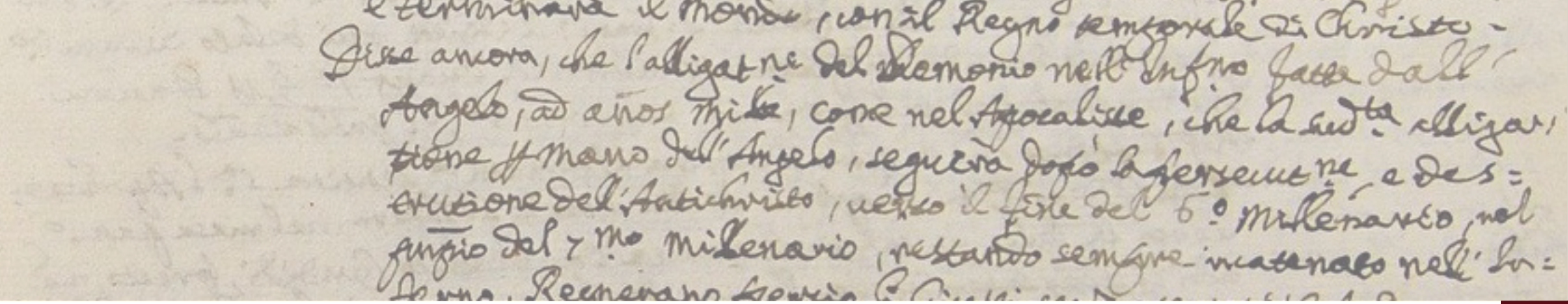


Date de disponibilité dc-rdf:available

Ce champ détermine la date à laquelle la ressource est devenue ou deviendra disponible.
Comme recommandé pour la propriété Date, dont Available est une sous-propriété, la date doit être inscrite suivant la norme ISO 8601, c'est-à-dire AAAA-MM-JJ. Il est donc possible de mettre des données sous embargo.

Date de modification dc-rdf:modified

Ce champ détermine la date à laquelle la ressource a été modifiée. La pratique recommandée est de donner la date, la date/heure ou la période. La date doit être inscrite suivant la norme ISO 8601, c'est-à-dire AAAA-MM-JJ.



Est une version de dc-rdf:isVersionOf

Ce champ indique la ressource connexe dont la ressource décrite est une version, une édition ou une adaptation. Les changements de version impliquent des modifications substantielles du contenu plutôt que des différences de format. Cette propriété est destinée à être utilisée avec des valeurs non-littérales. Cette propriété est une propriété inverse de Has Version. La ressource décrite est une version, une nouvelle édition, ou une adaptation de la ressource référencée. Des changements de versions impliquent de réels changements du contenu plutôt que des différences dans le format.

Exemple : Le rose et le jaune – Manuscrit 2

Support dc-rdf:medium

Ce champ indique le support matériel ou physique de la ressource.

FAIRiser vos données : mémorandum

c) Annexe n°3a : Plan de gestion des données Du consortium CAHIER

Plan de Gestion de Données du Consortium CAHIER

PGD V1: 30/10/2021, PGD Consortium CAHIER de Huma-Num

Ceci est le Plan de Gestion des Données du consortium CAHIER de l'Infrastructure Huma-Num. Il s'agit d'un plan de gestion des données de structure. De 2011 à 2021, CAHIER a travaillé à la numérisation et à la construction de données sur des corpus d'auteurs. Ce plan de gestion des données offre une vue générale des actions menées par le consortium sur les données au bout de dix ans.

CAHIER a donc rédigé deux types de plans de gestion des données : un Plan dit « de structure » et un plan par projet volontaire.

Les plans proposés par CAHIER s'inspirent des recommandations de l'Agence Nationale de la Recherche et du guide publié en 2016 par le programme Horizon Europe[1][1]. Les modèles mis en ligne par la plateforme DMP OPIDoR (https://opidor.fr/), l'outil d'aide à la création en ligne de plans de gestion de données (Data Management Plan ou DMP), ont été étudiés et adaptés aux besoins de la communauté scientifique du Consortium CAHIER. Le modèle de PGD proposé par CAHIER sera mis en ligne sur Opidor.

Auteurs du plan de gestion des données :

IDMHAND, Fatiha, IdHAL : fatiha-idmhand ; ORCID : 0000-0001-7135-9182, Université de Poitiers, Institut des textes et Manuscrits Modernes (ITEM, UMR8132), Paris, France

Rôle dans le consortium CAHIER : Coordinatrice adjointe

L'HERMITE, Laurène, Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France

Rôle dans le consortium CAHIER : Chargée de PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD Consortium CAHIER de Huma-Num
Une seule version de ce PGD est actuellement prévue

[1] [1] H2020 Programme « Guidelines on FAIR Data Management in Horizon 2020 », https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf#page=10

SOMMAIRE

Table des matières

PARTIE I

PGD V1: 30/10/2021, PGD du Consortium CAHIER de Huma-Num

PLAN DE GESTION DES DONNÉES DE STRUCTURE

LE CONSORTIUM CAHIER

1) Informations sur le plan de gestion de données (PGD)

- ***Présentation de la section***

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

- ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

IDMHAND, Fatiha, IdHAL : fatiha-idmhand ; ORCID : 0000-0001-7135-9182, Université de Poitiers, Institut des textes et Manuscrits Modernes (ITEM, UMR8132), Paris, France
Rôle dans le consortium CAHIER : Coordinatrice adjointe

L'HERMITE, Laurène, Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le consortium CAHIER : Chargée de PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD Consortium CAHIER de Huma-Num
Une seule version de ce PGD est actuellement prévue

2) Présentation du projet et responsabilités

- ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...).
Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France

Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

Consortium CAHIER"Corpus d'auteurs pour les humanités : informatisation, édition, recherche" de l'infrastructure Huma-Num (2011-2021)

Responsable du projet (principal researcher) et unité de rattachement

- Responsable du Consortium de 2011 à 2015 :

DEMONET, Marie-Luce, Université de Tours, Centre d'études Supérieures de la Renaissance (UMR7323)

Rôle dans le consortium CAHIER : Coordinatrice de 2011 à 2015

- Co-responsables du Consortium de 2015 à 2021 :

LEBARBE, Thomas, Université de Grenoble-Alpes, (UMR5316), Grenoble, France

Rôle dans le consortium CAHIER : Coordinateur de 2015-2021

IDMHAND, Fatiha, IdHAL : fatiha-idmhand ; ORCID : 0000-0001-7135-9182, Université de Poitiers, Institut des textes et Manuscrits Modernes (ITEM, UMR8132), Paris, France

Rôle dans le consortium CAHIER : Coordinatrice adjointe de 2015-2021

Rôle dans le consortium CAHIER : Coordinatrice du 09/09/2021-31/12/20212021

Financeur(s) du projet et type de financement

Infrastructure Huma-Num, financement de consortiums
CNRS UAR 3598 Huma-Num
Bâtiment de recherche Nord, 14, cours des humanités, 93322 Aubervilliers cedex

Référence de la convention de financement

Infrastructure Huma-Num, financement de consortiums

Institution / organisme / unité porteuses du projet

Centre national de la recherche scientifique, CNRS

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Maison des Sciences de l'Homme Val de Loire (MSH VdL) - USR CNRS 3501 -
33, Allée Ferdinand de Lesseps , 37204 TOURS CEDEX 03

Descriptif et objectif(s) du projet

CAHIER est un consortium interdisciplinaire de projets numériques menés principalement dans les domaines des ”corpus d’auteurs”. Constitué en fédération en 2011 dans le cadre de l’infrastructure”CORPUS”il a ensuite été intégré à la TGIR Huma-Num.
L’ensemble des projets sur corpus d’auteurs membres du consortium ont une activité éditoriale numérique ou double support (sur papier et en ligne). Ces éditions comportent différents degrés de description et d’analyse allant de l’édition du texte brut avec apparat critique minimal jusqu’à l’édition critique complète, l’édition génétique associant aux textes les images des documents originaux (manuscrits ou imprimés), en passant par des données encyclopédiques (dictionnaires, pièces d’archives, illustrations, bases de données). Ce consortium se définit par rapport à l’existence d’une œuvre (incluant les documents préparatoires) ou de plusieurs œuvres identifiées, dont la cohérence mérite d’être soulignée, publiée et outillée pour donner lieu à de nouvelles recherches. Il a pour but :

- d’augmenter l’acquisition de données de qualité (image et texte) tout en tenant compte des limites de taille
- de proposer et partager des normes de transcription en suivant les objectifs éditoriaux clairement énoncés
- de permettre l’indexation des corpus et des images
- d’offrir des données et des métadonnées compatibles avec les standards internationaux qui permettent l’exploitation des données (catalogage, archivage, identification, protection des données, moissonnage, etc.)
- d’adapter les solutions juridiques et les modèles de convention de partenariat (établir un dialogue efficace et collectif avec les éditeurs) en lien avec l’évolution des pratiques numériques
- d’offrir les moyens d’évoluer vers le web sémantique, la visualisation, les entrepôts de données, les modes de représentation d’ensembles documentaires, et vers l’annotation collaborative
- de favoriser l’appropriation des outils numériques en organisant des formations et des ateliers.

Le consortium permet de tester, voire d’élaborer, des outils prototypes (édition, traitement des données, etc.) dans le domaine des humanités numériques. Il favorise, en contribuant à leur financement, la participation aux ateliers et congrès internationaux et aux manifestations organisées par les structures européennes.

Dates et durée

Date de début de financement et de début des travaux : 2011
Date de fin de financement et de fin des travaux : 2021

Mots clés du projet

Corpus, Auteurs, Humanités, Informatisation, Édition, Recherches

Publications (articles, pré-proposition, site web, …)

Site web du consortium : https://cahier.hypotheses.org

Listes des articles publiés par le Consortium : https://halshs.archives-ouvertes.fr/search/index/q/*/structId_i/545625/
Autres livrables (guides, recommandations, etc.) : https://cahier.hypotheses.org/guides

3) Présentation et description du corpus

- ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doivent être décrits. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

Consortium CAHIER

Corpus

CAHIER est un consortium interdisciplinaire qui associe des projets numériques du domaine des "corpus d'auteurs». Qu'ils relèvent de la littérature, de la philosophie ou d'une thématique liée à une école ou à une pratique, les corpus étudiés par ce consortium sont, le plus souvent, associés à une activité éditoriale numérique, le plus souvent sur double support (papier et en ligne).

Période couverte par le corpus, auteur(s) concerné(s)

Le consortium se définit par rapport à l'existence d'une œuvre (incluant les documents préparatoires) ou de plusieurs œuvres identifiées, dont la cohérence mérite d'être soulignée, publiée et outillée pour donner lieu à de nouvelles recherches. Les corpus du consortium couvrent toutes les périodes depuis l'Antiquité, tous les genres littéraires et toutes les spécialités des sciences des textes.

Organisation du corpus

Les corpus numériques sont créés par les projets membres, ils comportent différents degrés de description et d'analyse allant de l'édition du texte brut avec apparat critique minimal, du corpus d'images accompagné de métadonnées jusqu'à l'édition critique complète ou l'édition génétique.

Mode de collecte et origine des données

Les projets numériques membres du consortium CAHIER collectent et constituent leurs données numériques de la façon suivante:

- lorsqu'il s'agit d'images: numérisations des sources primaires par le projet ou achat d'images numériques auprès d'opérateurs ayant déjà numérisé ces sources
- lorsqu'il s'agit des métadonnées : production automatique de métadonnées lors de la numérisation, saisie manuelle de métadonnées destinées à enrichir les données, saisie manuelle de métadonnées enrichissant l'édition (lors d'édition XML-TEI)
- lorsqu'il s'agit de données disponibles en ligne : réutilisation de données publiées, collecte semi-automatisée de données disponibles en ligne, traitement semi-automatisé des données, réutilisation

Les données constituées, collectées et étudiées par le consortium sont décrites selon les standards et normes actuelles:

- les archives sont décrites en XML EAD ou en Dublin Core selon les normes et recommandations des domaines
- les images sont nommées par les institutions qui hébergent les documents numérisés, lorsqu'il s'agit d'équipe de recherches, elles respectent les recommandations *du domaine: https://doranum.fr/stockage-archivage/comment-nommer-fichiers/*
- les éditions numériques sont produites et les données décrites en XML TEI

Etat du corpus numérique

Le consortium CAHIER a fondamentalement produit trois "types" de publications numériques : des "archives éditorialisées", des PGD V1: 30/10/2021, PGD du Consortium CAHIER de Huma-Num (2011-2021) des "éditions de lecture" et des "éditions enrichies"[2].
A la différence du guide des recommandations d'autres consortiums comme celui de la TEI et des *Best Practices for TEI in Libraries*, les publications numériques prises en compte par le consortium CAHIER ne sont pas uniquement codées en XML TEI, d'autres options éditoriales et techniques, susceptibles d'être utilisées par les chercheurs, ont été prises en compte, comme par exemple la publication via un CMS. Toutefois, pour garantir l'accessibilité et l'évaluation des publications, CAHIER a fortement recommandé d'utiliser des standards de description de données partagés par les communautés académiques internationales.

- le métalangage de description de documents XML permet la structuration poussée des contenus manipulés ainsi que leur exploitation dans des contextes variés. Ainsi, le consortium CAHIER a très largement recommandé, encouragé et utilisé le vocabulaire de référence TEI pour l'encodage de documents textuels ou nativement numériques et suggéré l'EAD pour l'encodage des objets de recherche archivistiques ;
- dans le cadre d'un projet de publication ayant vocation à être largement diffusé dans un contexte d'Open Access, l'usage de vocabulaires standardisés pour décrire les documents ca été particulièrement encouragé en lieu et place de la création de vocabulaires ad hoc potentiellement équivoques et abscons, le Dublin Core a fait partie des préconisations de CAHIER.

[2] Voir à ce propos : Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations (V1-Novembre 2018), https://halshs.archives-ouvertes.fr/halshs-01932519

Types de données:

Les données numériques produites par les projets membres du Consortium CAHIER sont de quatre types:

- images : JPEG, PDF et TIFF
- textes : txt, xml tei, JPEG, TIFF et PDF
- audio et son : mp3, mp4
- vidéos : avi, flv

Volumétrie

La production des données des 65 projets membres a été estimée à :

- 50 sites webs
- 30 URL permettant d'accéder aux téléchargements direct des fichiers
- 327450 fichiers annoncés comme disponibles mais 13201 fichiers réellement disponibles, pas d'informations précises sur la présence ou non d'océrisations
- 500000 images sources potentiellement disponibles mais moins de 35000 réellement disponibles, impossible d'estimer le poids en Go à ce stade

Modifications effectuées sur les données, versions, …

CAHIER n'a réalisé aucun traitement sur les fichiers produits par ses membres. Les projets membres sont les seuls à modifier leurs fichiers.
CAHIER a largement encouragé la FAIRisation des données et le dépôt de celles-ci dans l'entrepôt fourni par l'infrastructure Huma-Num : Nakala.
Les membres de CAHIER ont également utilisé Zenodo et Ortolang pour FAIRiser leurs données

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

A ce stade, CAHIER n'a réalisé aucun enrichissement des données produites par les projets membres.

Métadonnées, créées et standards et formats utilisés

Les métadonnées produites par les projets membres du Consortium sont destinées à décrire les ressources produites, les classer, organiser et à caractériser leur contenu. CAHIER a produit, de façon quasi-systématique, au moins quatre types de métadonnées pour ses publications numériques :

- des métadonnées descriptives permettant l'identification non ambiguë de la source (analogique et/ou numérique) ;
- des métadonnées administratives apportant des informations sur les caractéristiques des fichiers, les droits d'accès et d'usage, et sur le processus de création des données ;
- des métadonnées structurelles expliquant la composition ou l'organisation de la ressource : pages, chapitres, table des matières ou autres éléments constitutifs. Ces métadonnées facilitent la caractérisation, la navigation, la présentation et la compréhension de la structure des sources ;
- des métadonnées techniques précisant les caractéristiques techniques des données, les logiciels utilisés pour leur production et manipulation, leurs versions ;

- et des annotations permettant d'analyser et d'interpréter la source ou métadonnées d'enrichissement (balisage) faites au fil du texte, au moyen d'un jeu réfléchi et prédéterminé d'étiquettes et d'attributs.

Les métadonnées descriptives, administratives et techniques

CAHIER a recommandé qu'un certain nombre de champs soient complétés et notamment les champs mentionnés par la norme Dublin Core non-étendu. Ont été préconisés, en tant que minimum souhaitable, les champs qui apportent des informations sur le texte (titre, éditeur, date), la propriété intellectuelle (auteur, droits), l'instanciation, la gestion, la ressource (formats, dimensions), son contenu (type, mots-clés) et les modalités de préservation des documents. Dans une édition XML/ TEI, ces informations se retrouvent dans la structure minimale de l'en-tête (header TEI).

Lors de la saisie des métadonnées, CAHIER a incité les projets membres à porter une attention soutenue à la normalisation de la présentation de celles-ci, ce qui implique, par exemple, le respect des recommandations internationales pour la saisie des dates (AAAA-MM-JJ), des noms de lieux (PAYS, Ville), des noms de personnes (NOM, Prénom), etc. Le cas échéant, il pourra être utile de recourir à des thésaurus pour faciliter l'exploitation postérieure des données.

Dans le cas d'une publication de type "Édition enrichie" (type 3 du guide publié en 2018[3]), on présentera obligatoirement un jeu de métadonnées plus étendu en apportant toutes les informations nécessaires à la description du témoin de départ et à la caractérisation de la publication effectuée. Dans ce cas, un standard reconnu d'encodage des métadonnées, tel que METS, MIX, UNIMARC, XML-EAD, Dublin Core simple/étendu est recommandé, et lorsqu'il s'agit d'une édition XML/TEI, plusieurs éléments et sections du header permettent d'atteindre un très haut niveau de précision et de finesse dans la description de la source de départ et de l'édition produite. CAHIER a préconisé ainsi, qu'en plus de la section obligatoire <fileDesc>, que les sections <encodingDesc>, <profileDesc> et <revisionDesc> soient renseignées.

Les métadonnées structurelles et l'annotation sémantique

Dans une édition papier, les notes constituent traditionnellement l'espace privilégié dédié à l'apport des informations scientifiques. Tout en permettant d'insérer des notes, l'édition électronique dispose de systèmes plus élaborés pour enrichir le texte que CAHIER a recommandé d'utiliser. La présence d'enrichissements sémantiques, à l'aide de balises par exemple, constitue un élément discriminant entre les différents types de publications numériques.

L'enrichissement de la publication peut être réalisé à l'aide de multiples technologies et outils (traitement ou éditeur de texte, éditeur xml ou html, etc.) mais selon les projets de publication numériques, CAHIER a encouragé ses projets membres, selon les cas, à :

- coder, dans le cas des "éditions de lecture", les grandes sections du texte (, dans la TEI), les titres (lorsqu'ils existent, <head> dans la TEI) et les paragraphes (à l'aide des balises dédiées) de façon générale dans le cas ; également les actes, scènes et tours de parole pour le théâtre ; les strophes et les vers pour la poésie ; les et les paragraphes, les destinataires, les expéditeurs, les lieux de rédaction et d'expédition et les dates pour les correspondances ; les sections et les articles, les

[3] Cf. : https://halshs.archives-ouvertes.fr/halshs-01932519

rédacteurs, les dates de parution des textes pour les journaux, revues et gazettes ; et et de séparer le mot vedette du texte de l'entrée à proprement parler pour les entrées des dictionnaires et des encyclopédies.

- coder, dans le cas des "éditions enrichies", les divisions et les différents éléments de structure (scène et actes ; chapitres ; vers ou groupes de vers ; articles, etc.) avec une granularité aussi fine que possible. Dans le cas du théâtre, on pourra trouver des attributs (comme @who) permettant d'identifier de façon non ambiguë le locuteur de chaque réplique, et des didascalies clairement typées ; dans le cas de la poésie, des éléments annotant la rime et le rythme ; dans le cas des dictionnaires, une annotation de l'information grammaticale, étymologique et d'usage (si possible), l'identification des sources des citations et leur équipement avec des liens hypertexte pointant vers des éditions extérieures, etc.

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Le Consortium CAHIER a laissé chaque projet membre utiliser les vocabulaires contrôlés propres au champ du corpus d'auteurs ou du projet de recherche. Le plus souvent, les membres ont utilisé les vocabulaires contrôlés de thésaurus tels que le RAMEAU (référentiel d'autorité - matière en usage dans le milieu de la documentation et des bibliothèques d'enseignement supérieur et par la BnF).

A partir de 2015, et pour combler un manque concernant le vocabulaire permettant de décrire les concepts du domaine des textes littéraires et plus particulièrement les "genres" ou "types" de textes, CAHIER à travaillé à l'élaboration d'un référentiel pour les genres textuels : le thésaurus "Typologie textuelle" (https://opentheso.huma-num.fr/opentheso/ : ouvrir "Typologie 43"). CAHIER a utilisé le logiciel opensource Opentheso pour l'élaborer. Les thésaurus Opentheso sont édités et consultés en ligne et peuvent être importés et exportés sous plusieurs formats, et notamment SKOS[4] (SKOS Reference 2009). Un identifiant pérenne de type Handle ou Ark peut être attribué à chaque concept.[5]

Dans ce thésaurus, 365 concepts ont été décrits et tous ont été pourvus d'un identifiant unique de type DOI. CAHIER recommande aux projets membres d'utiliser les URL de ces identifiants, de les insérer dans les métadonnées pour décrire le genre textuel ou type de leurs documents.

[4] Langage de représentation de schémas de concepts mis au point par le W3C. Le SKOS permet de gérer des modèles sémantiques relationnels (type thésaurus, index d'autorités matière, taxonomies, folksonomies, …) de façon simple, dans la perspective du web sémantique.

[5] La documentation complète d'Opentheso est disponible en ligne sur le site https://opentheso.hypotheses.org.

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être important de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Les projets membres du Consortium ont fait appel à leurs institutions pour organiser le stockage de leurs données pendant la durée de leurs projets.
Dans quelques cas épars, ces institutions ont également la capacité de fournir des services de stockage à long terme et des identifiants pérennes aux données de façon à faciliter le stockage et l'accessibilité de celles-ci.
De façon générale, CAHIER a préconisé l'utilisation des services d'Human-Num pour assurer :

- la préservation des données pendant la durée des projets (le service Sharedocs a été largement utilisé)
- le stockage des données à long terme via l'entrepôt de stockage Nakala permettant de rendre les données des projets FAIR.

En 2021, les préconisations du Consortium en vue d'accompagner et de simplifier la FAIRisation des données des projets grâce à l'outil Nakala[6] ont essentiellement concerné les métadonnées descriptives. CAHIER a recommandé le respect des métadonnées suivantes et indiqué aux projets les champs attendus par Nakala :

Champ requis par Nakala	Données décrites en Dublin Core par les projets / Données attendues par Nakala en DCterms	Données décrites en TEI par les projets / Données attendues par Nakala en DCterms
Titre	dc:title / dcterms:title	teiHeader/fileDesc/titleStmt/title @type="main" / dcterms:title

[6] Voir le guide FAIR du Consortium CAHIER : https://halshs.archives-ouvertes.fr/halshs-02889777

Auteur	dc:creator / dcterms:creator	teiHeader/fileDesc/titleStmt/author / dcterms:creator
Éditeur scientifique	dc:contributor / dcterms:contributor	teiHeader/fileDesc/titleStmt/editor / dcterms:contributor
Éditeur	dc:publisher / dcterms:publisher	teiHeader/fileDesc/publicationStmt/publisher / dcterms:publisher
Date de publication du fichier électronique	dc:issued / dcterms :dateIssued Ou dcterms:date available (date à laquelle la ressource est devenue ou deviendra disponible.)	teiHeader/fileDesc/publicationStmt/date / dcterms :dateIssued Ou teiHeader/fileDesc/publicationStmt/availability/licence/date (si présent) / Ou dcterms:date available
Date	dc:date / dcterms:dateCreated Date de création du document source (appelée date première dans le catalogue OAI du consortium CAHIER)	teiHeader/profileDesc/creation/date @when="1857" / dcterms:dateCreated Ou teiHeader/profileDesc/creation/date @notBefore="1700" @notAfter="1750" / dcterms:dateCreated
Informations sur les droits d'utilisation	dc:rights / dcterms:rights	teiHeader/fileDesc/sourceDesc/bibl/note[@type='settlement'] / dcterms:rights ou teiHeader/filedesc/sourceDesc/biblStruct/note[@type='settlement'] / dcterms:rights ou, pour les manuscrits, teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/ settlement / dcterms:rights
Identifiant	dc:identifier / dcterms:identifier	teiHeader/fileDesc/publicationStmt/idno[@type="URL"] / dcterms:identifier Dans un second temps on ajoutera teiHeader/fileDesc/publicationStmt/idno @type="DOI" / dcterms:identifier
URI de licence	dc:rights / dcterms:license	teiHeader/fileDesc/publicationStmt/availability/licence[@target="URI de la licence"] / dcterms:license

Référence (bibliographique) du document d'origine	dc:source / dcterms:source	teiHeader/fileDesc/sourceDesc/bibl/note[@type='identifier'] / dcterms:source ou pour une bibliographie plus détaillée : teiHeader/fileDesc/sourceDesc/biblStruct/note[@type='identifier'] / dcterms:source Pour les manuscrits teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/[différents champs pertinents, don't idno] / dcterms:source
Langue	dc:language / dcterms:language	teiHeader/profileDesc/langUsage/language @ident="fr" / dcterms:language
Résumé	dc:description / dcterms:description	teiHeader/profileDesc/abstract / dcterms:description
Mots-clés	dc:subject / dcterms:subject	teiHeader/profileDesc/textClass/keywords/term @type="subject" / dcterms:subject

Accès, partage et limites des données

Les données produites par les projets membres du Consortium CAHIER sont d'emblée entièrement accessibles sans restriction. Dès sa fondation en 2011, CAHIER a fait de l'ouverture des données une priorité. Lorsque certaines données posent des contraintes aux ou des modalités spécifiques de diffusion, elles sont publiées sous embargo, limité dans le temps ou sous une Licence qui limite le partage des et explicite clairement les conditions et raisons particulières de consultation ou de réutilisation (nécessité d'un logiciel par exemple, d'un mot de passe, etc.).

La majorité des données diffusées par CAHIER sont rendues publiques en accès libre, à partir de textes et d'images intégralement libres de droits ou avec autorisation des ayants-droits.

En revanche, les métadonnées sont rendues disponibles sous la forme d'apparat critique et de notes explicatives sans restrictions, sous réserve du respect de la propriété intellectuelle de leurs auteurs et des documents auxquels elles se réfèrent.

5) Responsabilités et ressources pour la gestion des données

- ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Le consortium CAHIER ne dispose pas de personne ressource dédiée à la gestion des données. Chaque projet membre gère ses données dans le cadre de son laboratoire, de sa MSH ou de son université.
En revanche, CAHIER assure la mission de relais avec l'infrastructure à l'étape du stockage pérenne des données. Les données déposées dans Nakala sont gérées par l'infrastructure Huma-Num qui assure le stockage pérenne de celles-ci, leur délivre des identifiants pérennes (DOI) et propose des services permettant de les rendre interopérables et moissonnables par l'intermédiaire du protocole OAI-PMH.
Une fois sur Nakala, le responsable de la gestion des données est Huma-Num.
Avant ce dépôt, le :

- responsable de la qualité des données (traitement, anonymisation, format, nettoyage,…) est : le responsable du projet membre
- responsable de la collecte des données est : le responsable du projet membre
- responsable du stockage et de la sécurité des données (s'il ne s'agit pas d'un service d'Huma-Num) : le responsable du projet membre

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

6) Archivage des données

- ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

- ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

CAHIER utilise et recommande l'entrepôt de données Nakala

Durée de conservation des données

A ce stade, CAHIER recommande le stockage illimité dans le temps dans l'entrepôt de données Nakala.

Volume des données à conserver

L'ensemble des données déposées sur Nakala est voué à être conservé sur le long terme, il représente actuellement, et à ce stade des travaux du Consortium plus de 60Go.

Coûts alloués à la conservation

Les coûts d'archivage sont assumés par Huma-Num via Nakala. L'économie est estimée à près de ~7000€ / an pour chaque projet membre

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

Les outils utilisés par les membres du Consortium pour produire leurs données sont normalement libres, ouverts et pérennes. Aucun logiciel spécifique ne devrait être nécessaire pour accéder aux données.

7) Partage des données à l'issue du projet

- ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

- ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR : les données doivent être Trouvables (Findables), Accessibles, Interopérables et Réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentation, par des formats ouverts et non propriétaires et par une disponibilité garantie par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

Les données produites par le Consortium CAHIER concernent des corpus issus du domaine des sciences des textes littéraires, de tous les genres textuels de ce domaine et couvrent un empan large allant de l'Antiquité à nos jours. Le potentiel de réutilisation des données produites est élevé, bien que ces corpus numériques s'adressent, dans leur grande majorité, à une communauté experte et spécialiste du champ disciplinaire, de la période et/ou de l'auteur. Les données publiées par CAHIER visent essentiellement à documenter la recherche internationale sur les corpus concernés, la publication de ces données apporte des informations permettant d'éclaircir des zones d'ombre..
Si ces données s'adressent avant tout à un public de chercheurs en sciences humaines et sociales, elles peuvent néanmoins intéresser un public d'amateurs.

Eléments d'accompagnement qui permettent la réutilisation des données.

Comme indiqué précédemment, le consortium CAHIER a veillé à accompagner et recommander un certain nombre de bonnes pratiques facilitant la réutilisation des données produites par le consortium:

- les données produites respectent les standards et bonnes pratiques de numérisation des domaines ;
- les jeux de données sont accompagnés d'au moins quatre à cinq types de métadonnées afin d'en faciliter la contextualisation, la compréhension et la réutilisation à long terme ;
- les données produites sont documentées ;
- les données produites sont ouvertes, libres, réutilisables ;
- les données sont stockées dans un entrepôt ouvert (Nakala notamment) et sûr.

Publications sur les données pour en améliorer l'exposition

Les membres du consortium CAHIER ont publié de nombreux travaux en vue de disséminer leurs résultats scientifiques et de promouvoir et donner à connaître les données produites. Certains de ces travaux peuvent être lus et consultés sur HAL : https://halshs.archives-ouvertes.fr/search/index/q/*/structId_i/545625/ et sur les sites web des projets.

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

Chaque projet membre a assorti ses publications numériques et données de la licence la plus adaptée en fonction de son projet.
Toutefois, CAHIER a recommandé l'usage de Licence Creative Commons, et tout particulièrement, des licences suivantes afin d'encourager la réutilisation et la réexploitation des données :

Conditions et modes d'accès	**Licence ou contrat**	**Embargo / confidentialité**
Lorsque l'accès est totalement libre et la donnée réutilisable et modifiable sans restriction (CC) La citation de l'origine de la donnée ou source est obligatoire (BY)	Licence CC - BY	Pas d'embargo
Lorsque l'accès est totalement libre et la donnée réutilisable (CC) mais sous les conditions suivantes : - pas d'utilisation commerciale = NC - pas de modification de la source (ND) - obligation de rediffuser selon les mêmes conditions (SA) La citation de l'origine de la donnée ou source est obligatoire (BY)	Licence CC - BY - NC Licence CC - BY - NC - ND ou Licence CC - BY - NC - SA	Pas d'embargo mais si embargo il y a, obligation de préciser la durée des restrictions

PARTIE II

PGD V1: 30/10/2021, PGD du Consortium CAHIER de Huma-Num

MODÈLE ET RECOMMANDATIONS POUR LA RÉDACTION D'UN PLAN DE GESTION DES DONNÉES SUR DES CORPUS D'AUTEURS

1) Plan de gestion de données (PGD) du projet XXXXXXXX

- ***Présentation de la section***

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

- ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

NOM, Prénom, IdHAL : xxxxxxxxx ; ORCID : xxxxxxxxx, Université de xxxxxxxxx, xxxxxxxxx (sigle, EA ou UMR n°xxxx), xxxxxxxxx, France
Rôle dans le projet : xxxxxxxxx

NOM, Prénom, IdHAL : xxxxxxxxx ; ORCID : xxxxxxxxx, Université de xxxxxxxxx, xxxxxxxxx (sigle, EA ou UMR n°xxxx), xxxxxxxxx, France
Rôle dans le projet : xxxxxxxxx

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD projet xxxxxxxxx
xxxxxxxxx version de ce PGD est actuellement prévue

2) Présentation du projet et responsabilités

- ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...).
Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

xxxxxxxxx

Responsable du projet (principal researcher) et unité de rattachement

NOM, Prénom, IdHAL : xxxxxxxxx ; ORCID : xxxxxxxxx, Université de xxxxxxxxx, xxxxxxxxx (sigle, EA ou UMR n°xxxx), xxxxxxxxx, France
Rôle dans le projet : xxxxxxxxx

Financeur(s) du projet et type de financement

xxxxxxxxx

Référence de la convention de financement

xxxxxxxxx

Institution / organisme / unité porteuses du projet

xxxxxxxxx

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

xxxxxxxxx

Descriptif et objectif(s) du projet

xxxxxxxxx

Dates et durée

Date de début de financement et de début des travaux : xxxxxxxxx
Date de fin de financement et de fin des travaux : xxxxxxxxx

Mots clés du projet

xxxxxxxxx

Publications (articles, pré-proposition, site web, …)

Site web du projet : xxxxxxxxx
Listes des articles publiés par le projet : xxxxxxxxx
Autres livrables (guides, recommandations, etc.) : xxxxxxxxx

3) Présentation et description du corpus

- ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

xxxxxxxxx

Présenter et décrivez le corpus

xxxxxxxxx

Période couverte par le corpus, auteur(s) concerné(s)

xxxxxxxxx

Organisation du corpus

xxxxxxxxx

Mode de collecte et origine des données

xxxxxxxxx

Etat du corpus numérique

xxxxxxxxx

Types de données:

xxxxxxxxx

Volumétrie

xxxxxxxxx

Modifications effectuées sur les données, versions, …

xxxxxxxxx

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

xxxxxxxxx

Métadonnées, créées et standards et formats utilisés

xxxxxxxxx

Les métadonnées descriptives, administratives et techniques

xxxxxxxxx

Les métadonnées structurelles et l'annotation sémantique

xxxxxxxxx

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

xxxxxxxxx

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

xxxxxxxxx

Accès, partage et limites des données

xxxxxxxxx

5) Responsabilités et ressources pour la gestion des données

- ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

xxxxxxxxx

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont

pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

6) Archivage des données

- ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

- ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

xxxxxxxxx

Durée de conservation des données

xxxxxxxxx

Volume des données à conserver

xxxxxxxxx

Coûts alloués à la conservation

xxxxxxxxx

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

xxxxxxxxx

7) Partage des données à l’issue du projet

- ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s’il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

- ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR: les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d’accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l’accompagnement d’une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

xxxxxxxxx

Eléments d’accompagnement qui permettent la réutilisation des données.

xxxxxxxxx

Publications sur les données destinées à en améliorer l’exposition

xxxxxxxxx

Conditions de réutilisation : licences et contrats pour l’ensemble du projet

xxxxxxxxx

Annexe

RECOMMANDATIONS DÉTAILLÉES DU CONSORTIUM CAHIER POUR L'ÉLABORATION DES PLANS DE GESTION DES DONNÉES DES PROJETS SUR CORPUS D’AUTEURS

1) Informations sur le plan de gestion de données

Responsabilités (rédacteur du PGD, relecteurs, autres intervenants assurant la gestion du PGD et ses mises à jour)

Recommandations :
Désignez un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Si possible, associez-le à son identifiant ORCID, IdRef, ISNI, IdHal. Nommez ensuite l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Modèle / exemple :
Rôle : Nom, Prénom, Identifiant, Institution, Laboratoire, Unité de rattachement, Ville, Pays

Versions du document, historique des mises à jour et nombre de versions prévues

Recommandations :
Un PGD est un document qui évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet. Indiquez les versions dans l'ordre antéchronologique en commençant par l'actuelle.
Pour H2020 par exemple il est demandé 3 versions du PGD dont la première doit être envoyée dans les 6 mois après le début du projet.
Modèle / exemple :
PGD V2 : 08/09/2021, P. Nom, P. Nom, P. Nom, …
PGD V1 : 15/06/2021, P. Nom, P. Nom,...
Trois versions du PGD sont prévues.

2) Présentation du projet et responsabilités

Nom du projet

Recommandations :
Si le nom du projet est un acronyme, indiquez également la version développée.
Modèle / exemple :
Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")

Responsable du projet (principal researcher) et unité de rattachement

Recommandations :
Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Modèle / exemple :

Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France

Financeur(s) du projet et type de financement

Référence de la convention de financement

Institution / organisme / unité porteuses du projet

Recommandations :
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Organismes partenaires, ressources et co-financeurs du projet

Modèle / exemple :
Projet dans le cadre de l'IUF, en partenariat avec la Bibliothèque Mazarine, et cofinancé par l'OBVIL (SU), le DIM STCN, le consortium CORLI.

Descriptif et objectif(s) du projet

Recommandations :
Description de la nature du projet, ses objectifs et son déroulement. Il s'agit de comprendre le contexte et les types de données qui seront produites ou collectées au cours du projet.

Dates et durée

Recommandations :
Dates de début de financement ou date de début des travaux pour les projets sans financement.
Date de fin de financement ou durée prévue des travaux pour les projets sans financement.

Mots clés du projet

Recommandations :
Utilisez dans la mesure du possible des vocabulaires contrôlés, des thésaurus - préciser lesquels - et indiquer si les termes disposent d'identifiants types DOI, URI ou permaliens et les lier.

Publications (articles, pré-propositions, site web, …)

Recommandations :
Publications dans le cadre du projet et listées dans l'ordre antéchronologique.
Modèle / exemple :

En 2020 : Mise en ligne du Thésaurus des poissons et créatures aquatiques. En 2021 il comprend 2290 entrées.
En 2020 : Mise en ligne de la Bibliothèque Ichtya : https://ichtya.unicaen.fr/lab/bibliotheque/

3) Présentation et description du corpus

Présentation et description du corpus

Mode de constitution du corpus, collecte et origine des données

Recommandation :
Précisez le mode de collecte et l'origine des données
Modèle / exemple :
Collecte de données primaires avec la numérisation de sources papier - précisez les centres d'archives, bibliothèques, centres d'études supérieures, …
Moissonnage de ressources en ligne - précisez l'origine des données (Gallica, Europeana, autre bibliothèque numérique)
Collecte et extraction de passages et extraits de sources (textes, images) - précisez les sources...

Période couverte par le corpus, auteur(s) concerné(s) et organisation du corpus

Recommandations :
Décrivez l'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers.
Vous pouvez vous référer aux recommandations en ligne sur https://doranum.fr/stockage-archivage/comment-nommer-fichiers/. Y sont spécifiées les 5 règles de nommage essentielles des fichiers et documents.

Etat du corpus numérique

Recommandations :
Indiquez la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état (le corpus corpus est-il complet ? Si non pour quelles raisons ? L'état physique des sources était-il altéré, ce qui peut impacter sur la qualité de leur version numérique ?)
Pour que les données soient réutilisables sur le long terme, assurez-vous d'utiliser des formats ouverts, non propriétaires et d'un usage répandu au sein de la communauté de recherche. Pour vérifier l'éligibilité de vos formats vous pouvez utiliser l'outil https://facile.cines.fr/
Modèle / exemple :
Nom du jeu de données : Environ 16 000 feuillets en bon état de conservation, numérisés et"océrisés", qui représentent un poids d'environ 1,7Go (sur un corpus de 100 000 pages au total qui sera numérisé pour un total estimé de 16Go). Les données sont disponibles en mode texte au format PDF.
Nom du jeu de données : 950 images au format jpeg (env. 7Go)

Modifications effectuées sur les données, versions

Recommandations :
Les données ont-elles subi des traitements ? Si oui lesquels, par quels moyens et avec quels outils ? Retracez l'historique de ces modifications.
Soyez"fair": pour garantir l'accessibilité et la réutilisation des données il est recommandé de privilégier des logiciels libres de droits et open source.

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Recommandations :
Précisez le type et la nature de ces données. Il peut s'agir de notices descriptives ou bibliographiques, données d'analyse quantitative effectuées sur des textes, données issues de modélisation ou de simulation, …
Indiquez le mode de création ou de collecte (en précisant leur origine pour les données réutilisées) de ces données. L'objectif de leur création ou de leur collecte. Leur volumétrie.

Métadonnées préexistantes, métadonnées créées et standards et formats utilisés

Recommandations :
Il est recommandé d'utiliser un schéma de description courant, à l'instar du Dublin Core qui est également un standard interdisciplinaire. Il est également possible d'utiliser des standards spécifiques au type de données et à la discipline (EAD pour la description d'archives, MARC ou UniMARC pour des données bibliographiques, …).
Les formats d'échanges les plus courants sont l'XML et le CSV.
Pour s'informer sur les métadonnées, standards et formats : https://doranum.fr/metadonnees-standards-formats/fiche-synthetique/

Référentiels d'indexation et vocabulaires contrôlés, thésaurus ou ontologies disciplinaires utilisés

Documentation destinée à accompagner les métadonnées en vue de la réutilisation des données

Recommandations :
L'ANR spécifie dans son modèle de PGD que "la documentation accompagnant les données permet aux utilisateurs de les repérer facilement et apporte les informations nécessaires à un bon usage et une bonne interprétation". Il peut s'agir a minima d'un fichier "read me" rassemblant les informations générales sur les données, de la documentation destinée à la description et à la compréhension de l'organisation des données, un glossaire pour le vocabulaire spécifique et les acronymes, etc.

4) Stockage, sauvegarde et sécurité des données

Documentation numérique ou papier décrivant et renseignant le lieu de stockage final, les lieux et infrastructures de stockage des données pendant le projet

Recommandations *:*
Décrivez ici le matériel physique et les lieux de stockage des données.
Stockez vos données dans au moins 2 endroits. Évitez le stockage externe et privilégiez les outils mis à disposition par l'institution.
Les bonnes questions à se poser pour organiser au mieux la sauvegarde des données[7] *:*

- *Quel volume approximatif de données devrons-nous sauvegarder ? L'espace de stockage est-il suffisant ?*
- *Quelle sera la périodicité des sauvegardes ? quotidienne ? hebdomadaire ? mensuelle ?*
- *Quel service utiliser, fourni par quelle institution ?*
- *Les données sont-elles sous contrat de maintenance ?*
- *Faudra-t-il accéder fréquemment aux données ? en temps réel ?*
- *etc.*

Volumétrie des données stockées. Modalités de sauvegarde et de protection des données

Recommandations :
Indiquez ici les procédures de sauvegarde mises en place (fréquence des sauvegardes ? Automatisée ou non ?), les personnes en charge de la protection des données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…
Indiquez également si l'accès aux données est sécurisé ou non.

Risques

Recommandations :
Différents facteurs sont susceptibles de menacer l'intégrité, la disponibilité et la confidentialité des données produites au cours du projet. Les risques peuvent être de différentes natures : des risques naturels pesant sur les infrastructures (zones sismiques, inondables etc.), des risques techniques (corruptions ou pertes de données, problèmes de serveurs etc.), des risques de confidentialité (accès non autorisés, fuites de données sensibles, etc.)[8]

[7] D'après : Atelier Données,"Guide de bonnes pratiques sur la gestion des données de la recherche", janv. 2021. https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html (consulté le juill. 13, 2021).
[8] D'après A.CARTIER, R.DELEMONTEZ, M.MOYSAN, N.REYMONET. Réaliser un plan de gestion de données"FAIR": guide de rédaction (v2, 2018)

5) Accès et partage des données

Modalités d'accès et de partage des données pendant la durée du projet

Recommandations :
Précisez si des données seront d'emblée entièrement accessibles sans restriction.
Existe-t-il des contraintes ou des modalités de diffusion particulières pour certains jeux de données (période limitée de partage, embargo) ?
Indiquer si les données sont accessibles selon des restrictions ou des conditions particulières : sont-elles sur un serveur local, un intranet, sur internet, en libre accès ou avec un accès authentifié. Préciser, le cas échéant, les différents niveaux d'habilitations et les règles qui les régissent.
Modèle / exemple :
Les données primaires seront rendues publiques en accès libre, à partir de textes et d'images intégralement libres de droits.
Les métadonnées seront disponibles sous la forme d'apparat critique et de notes explicatives (éditions scientifiques intégrales, dossiers d'études, données bibliographiques, ressources informatives historiques). Elles seront rendues publiques sous la réserve du respect de la propriété intellectuelle de leurs auteurs (professeurs, chercheurs, spécialistes universitaires).

Limites éventuelles à l'accès aux données

Recommandations :
Utilisation nécessaire de logiciels propriétaires par exemple.

Partage des données

Recommandations :
Indiquez par exemple si elles sont stockées sur un entrepôt de données, indexées dans un catalogue, accessibles par demande directe…
Dans le cadre du consortium CAHIER et afin de respecter les principes FAIR, le consortium recommande le stockage des données sur Nakala, le service de dépôt et de stockage sécurisé proposé par la TGIR HumaNum pour les SHS. Un second dépôt sur un autre entrepôt de votre choix ou sur une plateforme institutionnelle reste possible dans tous les cas.
Modèle / exemple :
Toutes les données sont libres de droits. Elles seront accessibles sur une interface dédiée et déposées sur Nakala. Les données et métadonnées seront interopérables et moissonnables par l'intermédiaire du protocole OAI-PMH.

6) Responsabilités et ressources pour la gestion des données

Identifiez et décrivez les rôles de responsabilité des données dans votre projet, et nommez si possible les personnes impliquées.

Recommandations :
Nommez la/les personnes responsables de la saisie des données, de la production des métadonnées, du traitement, de l'analyse, du stockage et de la sauvegarde des données ainsi que de leur partage et éventuellement leur archivage. Un responsable de la coordination du système de gestion des données pourra être nommé. Celui-ci devra idéalement être impliqué dans le pilotage du projet.
Modèle / exemple *:*
Responsable de la gestion des données : Nom prénom, institution, ville, adresse mail.
Responsable de la qualité des données (traitement, anonymisation, format, nettoyage, …) : Nom prénom, institution, ville, adresse mail
Responsable de la collecte des données : Nom prénom, institution, ville, adresse mail
Responsable du stockage et de la sécurité des données : Nom prénom, institution, ville, adresse mail

Évaluez les coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Recommandations :
Moyens matériels (serveurs, machines virtuelles, logiciels dédiés, licences supplémentaires, …)
Moyens humains (recrutement évalué en ETP pour des tâches spécifiques relevant de la gestion des données, actions de formation, …)

7) Archivage des données

Quelles sont les données à conserver sur le moyen et le long terme ? Quelles sont les données à détruire ?

Recommandations :
A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire destructibles pour des raisons légales et de confidentialité. Expliquez le choix des données à conserver et à détruire.

Sur quelle plateforme est prévu l'archivage pérenne des données ? Si un autre moyen est envisagé, précisez lequel et décrivez les outils et méthodes.

Recommandations :
Indiquez quelle plateforme pourrait accueillir les données et à quel coût.
Certaines institutions proposent un service d'archivage des données de la recherche. Actuellement le service de référence mandaté par le Ministère de l'enseignement supérieur et de la recherche est le CINES.

Durée de conservation des données

Recommandations :
Elle est à déterminer en fonction des jeux de données, des coûts alloués à l'archivage et au service de conservation choisi.

Volume des données à conserver

Modèle / exemple :
L'ensemble des données est voué à être conservé sur le long terme. Ce qui représente environ 50 Go.

Coûts alloués à la conservation

Recommandations :
Les coûts nécessaires de l'archivage sont à déterminer en fonction de la durée de conservation de vos données, du volume à conserver et de la plateforme sélectionnée.

Quels outils, méthodes, procédures seront nécessaires pour accéder à ces données archivées et les réutiliser ? (logiciels spécifiques, identification et droits pour accéder à la plateforme, ...)

8) Partage des données à l'issu du projet

Politique de dissémination des données

Recommandations :
Précisez s'il existe des limites à la diffusion des données. Expliquez également comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public. En effet, les données devront dans la mesure du possible respecter des principes FAIR, c'est-à-dire être trouvables (findables), accessibles, interopérables et réutilisables : les données devront être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI), leur usage est facilité a minima par l'accompagnement d'une description et de documentation, des formats ouverts et non propriétaires, un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

Recommandations :
Expliquez ici le potentiel de réutilisation des données et à qui elles pourraient être utiles. Il est notamment possible de suggérer d'autres axes de recherche à partir du projet
Modèle / exemple :
Les données publiées visent à documenter la recherche internationale sur la circulation de la pensée durant la première moitié du vingtième siècle. En raison du contexte de ces exils (guerres d'Espagne, guerres mondiales), il reste encore des zones d'ombres sur l'itinéraire et le parcours de certains intellectuels de l'époque. La publication de ces données vise à apporter des informations permettant d'éclaircir ces zones d'ombre.
Ces données s'adressent avant tout à un public de chercheurs en sciences humaines et sociales, mais peuvent également intéresser un public d'amateurs.

Eléments d'accompagnement qui permettent la réutilisation des données.

Recommandations :
Indiquez si de la documentation accompagne les jeux de données afin d'en faciliter la compréhension et réutilisation à long terme : des documents permettant de les décrire, d'en expliquer l'usage et les potentialités d'usage, la manière dont les données ont été produites ou collectées, quelles sont les contraintes d'usage, … voire de les enrichir.
Modèle / exemple :
Thésaurus, ontologie, codes informatiques, algorithmes, document explicatif (read me), inventaire, bibliographie,...

Publications sur les données pour en améliorer l'exposition

Recommandations :
Publications faites sur le projet et l'exploitation de ces jeux de données.
La rédaction d'un data paper est généralement recommandée. On peut également citer des articles publiés sur le le projet et faisant des liens avec les données.

Conditions de réutilisation (licences et contrats pour l'ensemble du projet et sur chaque jeu de données)

Recommandations :

Donnez les conditions de réutilisation des données pour le projet.

Listez pour chaque jeu de données : les conditions et modes d'accès aux données (accès libre ou restreint, sur autorisation, selon le statut, …), s'il est sous licence ou contrat[9], s'il est contraint par un embargo ou d'accès restreint pour des questions de confidentialité ou de sensibilité des données.

Modèle */* exemple :

Les données initiales sont dans le domaine public. Certaines données créées dans le cadre du projet sont librement accessibles par téléchargement sous une licence Attribution – Pas d'Utilisation Commerciale – Partage dans les Mêmes Conditions 4.0".

[9] Voir **A.CARTIER, R.DELEMONTEZ, M.MOYSAN, N.REYMONET**. *Réaliser un plan de gestion de données"FAIR": guide de rédaction* (v2, 2018), p31.
Disponible en ligne : https://archivesic.ccsd.cnrs.fr/sic_01690547v2/document
"La licence précise les conditions de partage et de réutilisation des données diffusées dans le cadre du projet, ainsi les éventuelles contreparties intellectuelles ou économiques qui y sont associées. Il est important de préciser qu'une diffusion en libre accès ne signifie pas nécessairement qu'une œuvre est libre de droit. La licence a notamment pour objectif de clarifier le statut juridique d'une œuvre et de préciser les conditions d'usage. De trop nombreux contenus ne sont pas réutilisés au maximum de leur potentialité en raison des ambiguïtés juridiques dues à l'absence de licences explicites. Il existe de nombreuses licences libres à utiliser selon les législations, les formats de données, les souhaits de protection des auteurs, les exigences des financeurs, etc."
Pour les jeux de données (liste non exhaustive)

- CC-by 4.0 : https://creativecommons.org/licenses/by-sa/4.0/deed.fr
- CC0 : https://creativecommons.org/publicdomain/zero/1.0/legalcode.fr
- La licence ETALAB : https ://www.etalab.gouv.fr/licence-ouverte-open-licence

Pour les bases de données (liste non exhaustive)

- OKF – Open Database License (OdbL) : http://opendatacommons.org/licenses/odbl/1.0/
- OKF – ODC-By : http://opendatacommons.org/licenses/by/1.0/
- OKF – ODC-PDDL (Public domain) : http://opendatacommons.org/licenses/pddl/1.0/

Pour les logiciels (liste non exhaustive)

- Open Licence Software (OPS) : https://opensource.org/licenses/OSL-3.0
- GNU-GPL : http://www.gnu.org/licenses/gpl.html

d) Annexe n°3b : Plans de gestion des données de ~~dix~~ projets membres

sept

Plan de gestion de données

Antonomaz

Table des matières

---

1) Plan de gestion de données (PGD) du projet Antonomaz (« ANalyse auTOmatique et NumérisatiOn des MAZarinades »)

Présentation de la section

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

☐ ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

ABIVEN, Karine, IdHAL : karine-abiven ; ORCID : *https://orcid.org/0000-0001-9518-1040*,
Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
Rôle dans le projet : Responsable du projet

LEJEUNE, Gaël, IdHAL : gael-lejeune ; ORCID : https://orcid.org/0000-0002-4795-2362,
Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
Rôle dans le projet : Co-responsable du projet

BARTZ, Alexandre. IdHAL : alexandre-bartz ; ORCID :
https://orcid.org/0000-0003-0850-8266
Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France
Rôle dans le projet : Ingénieur du projet (en 2021).

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle,
Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD projet Antonomaz
Deux versions de ce PGD sont actuellement prévues

2) Présentation du projet et responsabilités

☐ ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")

Responsable du projet (principal researcher) et unité de rattachement

ABIVEN, Karine, IdHAL : karine-abiven; ORCID : *https://orcid.org/0000-0001-9518-1040*, Université Paris - Sorbonne,Sens-Texte-Informatique-Histoire (STIH), EA 4509, Paris IV, France

Rôle dans le projet : Responsable du projet

Financeur(s) du projet et type de financement

DIM STCN (Ile de France) en 2019-2021.
IUF (Institut Universitaire de France) en 2020-2025.

Référence de la convention de financement

ne s'applique pas.

Institution / organisme / unité porteuses du projet

Projet dans le cadre de l'Institut Universitaire de France

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

En partenariat avec la Bibliothèque Mazarine, et cofinancé par l'OBVIL (SU), le DIM STCN, le consortium CORLI.
Le projet a été labellisé par le Consortium CAHIER [https://cahier.hypotheses.org/membres] de la Tigr Huma-num (voir aussi [https://cahier.hypotheses.org/antonomaz]).

Descriptif et objectif(s) du projet

Le projet Antonomaz, "ANalyse auTOmatique et NumérisatiOn des MAZarinades" vise à exploiter une collection numérique d'environ 5000 écrits ayant pour objet les affaires politiques de la régence du cardinal Mazarin, et traditionnellement appelés "mazarinades".

Notre approche se situe dans le champ des études littéraires, de l'analyse du discours et des Humanités Numériques. Elle vise à fournir des méthodes automatiques, empruntant au Traitement Automatique des Langues, à la fouille de données et à la linguistique de corpus, pour l'analyse de ces données par les spécialistes des divers domaines concernés (littéraires, linguistes, historiens, principalement).

Objectifs :

1/ Offrir une collection numérique de ces écrits, la plus exhaustive possible, en libre accès. Dans ce but, le projet abonde une bibliothèque numérique, celle de la Bibliothèque Mazarine (Mazarinum), en favorisant les numérisations puis en automatisant le passage du mode image au mode texte. Il s'agit d'offrir aux utilisateurs une collection téléchargeable pour des traitements avancés (statistiques, par exemple), pour compléter des ressources comme le signalement en ligne de collections entières (comme celle de l'Université de Tokyo, mise en ligne par le groupe des Recherches internationales sur les Mazarinades). Le but serait aussi de permettre la sélection de corpus cohérents à l'intérieur de cette nouvelle collection numérique, en fonction des besoins de chaque usager.

2/ Améliorer les données textuelles obtenues par des transcriptions automatiques (par reconnaissance de caractères), en mettant à profit les méthodes d'apprentissage profond. Il s'agit de paramétrer finement la reconnaissance automatique des caractères originaux figurant dans l'imprimé ancien.

Rendre disponible ce modèle entraîné, qui pourrait servir à améliorer les sorties en mode

textes de fac-simile d'autres projets numériques s'intéressant aux "imprimés non-livres" (*non-book printed material*), par exemple la Newberry French Pamphlet Collection

(https://archive.org/details/newberryfrenchpamphlets?and[]=mediatype%3A%22texts%22&and[]=year%3A%221649%22)

3/ Expérimenter des applications en Traitement Automatique des Langues : datation automatique, attribution d'auteur, classification non-supervisée. Ces expériences exploitent

d'abord directement les données brutes (sorties d'OCR bruitées), dont l'analyse au grain caractère peut produire des résultats parfois meilleurs que les données lissées pour l'œil humain, bien plus coûteuses pourtant à obtenir.

S'ensuivent plusieurs pistes de travail, comme l'automatisation de la normalisation et l'annotation du mode texte, ainsi que divers types de balisage et d'extractions d'informations comme les entités nommées.

4/ Proposer une visualisation originale des liens entre ces textes polémiques : en raison de leur nature réactionnelle, ils a n'ont de sens que pris dans leur contexte fin et compris dans leur mise en réseau. Il s'agira donc de donner un accès visuel dynamique à leur enchaînement à la fois chronologique et réticulaire. Des visualisations des métadonnées seront aussi proposées (par date, par épisodes historiques, par taille de l'imprimé, par éditeurs connus, etc.)

Dates et durée

Date de début de financement et de début des travaux : 2019.
Date de fin de financement et de fin des travaux : 2025.

Mots clés du projet

- Mazarinades
- Langue et littérature françaises
- Traitement Automatique du langage naturel
- Analyse du discours
- Exploration de données
- Data visualisation
- Reconnaissance optique des caractères
- Métadonnées
- Philologie numérique
- Pamphlets

Publications (articles, pré-proposition, site web, …)

Sites web du projet :

- Site web à venir. Présentation de la collection numérique de mazarinades : accès aux PDF accessibles en ligne mis en réseau. Possibilité de recherches plein texte, de recherches d'entités nommées, et de recherches textométriques. Visualisations des métadonnées.

Sites de présentation actifs :

- http://www.dim-humanites-numeriques.fr/projets/antonomaz/
- Carnet de recherche en lien avec le projet : https://libelles.hypotheses.org/

Bases de données interrogeables en ligne :

- Mise en ligne des numérisations en format IIIF d'un ensemble de libelles conservés à la Mazarine (1ere vague de 419 libelles dont le financement a été assuré par l'OBVIL pour Antonomaz) : https://mazarinades.bibliotheque-mazarine.fr/

- Mise en ligne de la bibliographie Moreau (“Moreau En Ligne”), et de ses suppléments, structurée en format texte et interrogeable par divers champ (numéro Moreau, date, mots de la notice Moreau, nombre de pages, etc.). Une dernière version à jour de Juillet 2021 est en ligne : http://memes.sorbonne-universite.fr/visualisation/Moreau/test.html

Listes des articles publiés par le projet :

- 2021a : Karine Abiven, Gaël Lejeune, “Des données au corpus : l’exploitation numérique des mazarinades”, *Dix ans de Corpus d’auteurs*, Editions des Archives contemporaines, accepté.
- 2021b : **Karine Abiven, Jean-Baptiste Tanguy et Gaël Lejeune**, “Exploiter en corpus des données textuelles océrisées : l’écriture burlesque de la Fronde (1648-1652)”, accepté, *revue Humanités numériques*, n°4 - Humanistica.
- 28/06/20 : **Jean-Baptiste Tanguy,** “Exploiter des modèles de langue pour évaluer des sorties de logiciels d’OCR pour des documents français du XVIIe siècle”, article accepté à *RECITAL*@TALN 2020,
- 10/03/20 : **Anaëlle Baledent (GREYC, Normandie Université), Nicolas Hiebel et Gaël Lejeune**, “Dating Ancient texts: an Approach for Noisy French Documents”, article accepté à *Language Technologies for Historical and Ancient Languages (LT4HALA)*,
- 2019 : **K. Abiven** : « Le moment discursif des barricades d’août 1648 : quelle interprétation des récurrences dans le discours sur l’événement ? », *Cahiers de Narratologie* [En ligne], 35 | 2019, mis en ligne le 03 septembre 2019, URL : http://journals.openedition.org/narratologie/9264
- 29/11/19 : **K. Abiven**: « La liste de noms propres dans les libelles de la Fronde : les revendications de prestige et leur satire », *Journées d’étude Listes de noms. Ordre social et ordre du livre*, M. Roussillon et C. Schuwey, Université d’Artois, Arras.
- 05/04/19 : Séminaire à l'OBVIL : analyse stylistique de textes littéraires
- 21/03/19 : Projet Antonomaz, Séminaire LCSU
- **A. Baledent et G. Lejeune**, “Automatic Stylistic Analysis; a search for efficient and interpretable descriptors to characterize individual writing style”, in *Phraséologie et stylistique de la langue littéraire*, Ludwig Fesenmeier et Iva Novakova (eds.), Peter Lang, 2020, p. 329-342.
- 14/03/19 : Anaëlle Baledent et Gaël Lejeune, “Analyse stylistique automatique : A la recherche d'indices efficaces et pertinents pour caractériser le style de Dumas”, *Phraseorom* 2019.
- 15/01/19 : **Karine Abiven et Gaël Lejeune**, “Analyse automatique de documents anciens : tirer parti d'un corpus incomplet, hétérogène et bruité”, revue *RIDOWS* – Pdf

Autres livrables (guides, recommandations, etc.) :

Un modèle d’OCR entrainé, spécifique aux imprimés de type brochure de l’Ancien Régime : reprise du modèle entraîné par Simon Gabay et Claire Jahan (spécifique aux imprimés du XVIIe s.). Il est prévu d’entraîner ce modèle sur les futures numérisations livrées par la bibliothèque Mazarine (lot 2 des mazarinades). Cet entraînement sera effectué avec E-scriptorium ce qui permet l'utilisation du standard IIIF (en lien avec Biblissima) pour la récupération des images et des métadonnées. Une fois ce modèle entraîné, il sera mis à la

disposition de la communauté : diffusion du modèle en licence CC-BY et sur la plateforme GitHub HTR-United.

3) Présentation et description du corpus

☐ ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

Antonomaz, "ANalyse auTOmatique et NumérisatiOn des MAZarinades"

Présenter et décrivez le corpus

L'ensemble recouvre théoriquement 5063 titres connus d'écrits généralement brefs (brochures et libelles) appelés "mazarinades" (sur environ 25 000 exemplaires référencés par la bibliothèque Mazarine, car plusieurs exemplaires d'une même édition ont été conservés). Ces écrits ont été publiés pendant la Fronde (1648-1653). Il s'agit de documents imprimés, dont la teneur est soit en faveur ou en défaveur de la politique du cardinal Mazarin. Cette collection est aussi diversifiée par les genres (chansons, relations, proclamations, satires, etc.) que par la forme des publications (placard, livret, acte officiel, etc.). Le référentiel bibliographique ainsi constitué se fonde notamment sur une bibliographie établie par Célestin Moreau en 1850-1851, outil bibliographique numérisé par le projet, et sa rénovation actuelle par la bibliothèque Mazarine (en janvier 2020, 20% des notices attendues à terme).

Un des enjeux d'Antonomaz est de permettre à l'utilisateur de sélectionner, à l'intérieur de cette collection numérique, des corpus cohérents (par auteur, par genre, par événement, etc.: par exemple : écrits de Scarron, écrits burlesques, écrits relatifs aux barricades de Paris).

Période couverte par le corpus, auteur(s) concerné(s)

1648-1653

Organisation du corpus

- Nommage des documents : identifiant Moreau et ses suppléments (avec le même codage que la Base Bibliographique des Mazarinades) et indication de la source numérique. Par

exemple Moreau50_GALL (pour un imprimé référencé au numéro 50 par Moreau et trouvé dans la bibliothèque numérique de la BNF, Gallica).
- Architecture des Xml sur le Github
- Structuration en collections, sous-collections, réseaux de textes: à venir.

Mode de collecte et origine des données

- *Fac-simile* numériques issus des bibliothèques numériques Gallica, Mazarinum, Gbooks. Pour Gallica, collecte semi-automatisée grâce à l'API de cette bibliothèque.
- Pour les métadonnées, Moreau en ligne et données bibliographiques issus de la base bibliographique de la Bibliothèque Mazarine.

Etat du corpus numérique

- Corpus en cours de production : actuellement (octobre 2021) 2 970 documents en PDF recueillis en ligne, dont 2569 éditions uniques, c'est-à-dire sans compter les différents exemplaires d'une même édition. La collecte a été faite sur les bibliothèques numériques Gallica, Mazarinum, Gbooks. Choix des PDF selon la qualité (dans l'ordre de préférence Mazarinum - de très haute qualité -, Gallica - souvent numérisés sur microfilms -, GBooks - souvent la dernière page est réduplicuée de nombreuses fois).

- Une partie des fichiers ont été océrisés à ce jour via une chaîne de traitement adaptée aux textes imprimés du 17e siècle.

- Leur passage semi-automatique en xml (2000 documents à ce jour, avec structuration minimale) est en cours; et les relectures (transcription et relectures du header). Cf volumétrie ci-dessous.

Types de données:

Images et textes en PDF
Transcriptions encodées en XML-TEI (avec métadonnées)
Images IIIF

Volumétrie

Evolution de la volumétrie (en nombre) des données au 27/09/2021

Date	PDF #docs	PDF #pages	XML #docs	XML #pages	XML #tokens	XML #types)	Retranscrits #docs
02/06/2021	1.111	15.000	447	1.500	2.108.211	199.374	105
02/07/2021	2.221	*71.069*	687	*10.423*	2.647.056	242.418	105
22/07/2021	2.613	*80.524*	750	*11.368*	2.811.341	257.506	105
27/09/2021	2.613	*80.524*	*2000*	*30. 000*	*7.350.000*	257.506	105

Modifications effectuées sur les données, versions, …

Océrisation et encodage XML/TEI des imprimés. Versionnage en fonction de l'avancée des relectures.

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Liens vers les différents *githubs* créés pour les besoins du projet :

- Outils développés, antonomaz_tools : https://github.com/rundimeco/antonomaz_tools
- Encodage du corpus, Antonomaz : https://github.com/Antonomaz

Métadonnées créées et standards et formats utilisés

- Pour les données interrogeables via la base Moreau-En-Ligne :

Les métadonnées utilisées pour le moment (2019-2021) sont celles léguées par la tradition bibliographique, issues du répertoire Moreau. Cet outil en ligne permet de requêter les imprimés par numéro Moreau, par une séquence de caractères du titre, l'année, la date plus précise quand elle est accessible, le lieu de publication, le nombre de pages et la notice Moreau.

- Peu à peu, ces métadonnées sont remplacées, notamment dans l'encodage xml des documents, ces métadonnées sont corrigées et enrichies par celles issues des archives d'H. Carrier , qui préparait une bibliographie critique des Mazarinades. Voir : https://mazarinum.bibliotheque-mazarine.fr/expositions-virtuelles/item/17787-vii-enqueter-sur-les-mazarinades?oeuvre=19#page=1&viewer=picture&o=no&n=0&q=

Ces métadonnées rénovées sont le fruit du travail de la Bibliothèque mazarine, dans leur "Base Bibliographique des Mazarinades", outil évolutif lancé en 2019: lhttps://mazarinades.bibliotheque-mazarine.fr/

Les métadonnées descriptives, administratives et techniques

L'en-tête des fichiers XML/TEI contient les informations habituellement requises :

- Les informations sur le texte (titre) et quand disponibles : la date (présente à 98% sur les imprimés), le lieu (environ 80 % continent leur lieu d'impression, avec environ 40 fausses adresses), l'éditeur, l'auteur (mais 80 % d'anonymes), nombre de pages, format (in quarto pour plus de 90% des documents, très homogènes de ce point de vue).
- On renvoie quand disponible, à la notice de la Base Bibliographique de la Mazarine, qui fournit les métadonnées les plus fiables (sur 20% du corpus). Pour les autres, on signale la source de connaissance des métadonnées (bibliographie Moreau ou catalogues de bibliothèque).
- Les noms et les lieux sont liés au web sémantique par des balises contenant l'isni, le geoname et l'identifiant wikidata.
- La balise <MsDesc> inscrit la cote du document physique, le lieu de conservation. On ajoute une balise sur la présence ou non de tampons (avec réponse boléenne) qui peut renseigner sur l'origine du document. A terme on aimerait renseigner la

présence ou nom du bref imprimé dans un recueil car c'est une information très importante pour connaitre les usages qui en ont été faits.

- Les mots clés concernent la forme (vers/prose), le ou les genres et sous-genres, le sujet quand récupérable sur la base bibliographique de la Mazaraine, ou quand il a été possible de le renseigner à la main.
- La balise encodingDesc > est ainsi renseignées : « Cette édition a été réalisée dans le cadre du projet ANTONOMAZ. Son objectif principal est de fournir un texte destiné à l'exploration avec des outils électroniques. De ce fait, ce n'est ni une édition philologique, ni une édition pédagogique ou de redécouverte d'un auteur oublié.
    - Les textes encodés dans le cadre du projet ANTONOMAZ sont issus de numérisations de bibliothèques numériques publiques et de Google livres.
    - L'édition présentée ici est issue d'un processus d'OCRisation réalisé avec Kraken. .

Les métadonnées structurelles et l'annotation sémantique

Le Format Json a été utilisé pour structurer la liste des titres et la liste des épisodes historiques (qui seront ensuite encodés semi-automatiquement dans le fichier xml).
Les balises TEI servent à la structuration du texte et des métadonnées.
Elles sont contrôlées et encadrées par un schéma de validation (ODD) auquel nous renvoyons pour documenter les métadonnées et les principes d'annotation :
Projet Antonomaz, ODD, 2021, consulté le 20/10/2021, URL :
https://github.com/Antonomaz/ODD

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Dans le balisage TEI :
- isni
- geonames
- wikidata
- identifiant unique et URL pérenne pour chaque édition de texte, appelé numéro BM (quand disponible), depuis la Base bibliographique des mazarinades (https://mazarinades.bibliotheque-mazarine.fr/)

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

☐ ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

☐ ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Un serveur avec une page web : https://antonomaz.huma-num.fr
L'accès est ouvert depuis le 15 octobre 2021. Il rassemblera les divers espaces de stockage utilisé jusqu'ici :

- Un wiki destiné à documenter le projet et les données : http://stih-sorbonne-universite.fr/dokuwiki/doku.php?id=antonomaz
- Données d'encodage stockées sur un GitHub : https://github.com/orgs/Antonomaz/repositories
- Une ressource bibliographique interrogeable : http://memes.sorbonne-universite.fr/visualisation/Moreau/test.html
- Ainsi que Sharedocs, espace de partage d'Huma-num, utilisé en interne pour stocker les données.

Accès, partage et limites des données

Données primaires publiques.

Fichiers en format texte produits et accessibles sur le github du projet : Licence Creative Commons CC-by.

Métadonnées de la « Bibliographie des Mazarinades » : Licence Creative Commons CC-by-nc-nd.

Sources en format PDF (fac-simile numériques) :

Gallica : Licence ODBL.
Mazarinum: Licence CC-by-nc-nd
GBooks : ne pas utiliser les fichiers à des fins commerciales, ne pas supprimer l'attribution Google.

5) Responsabilités et ressources pour la gestion des données

☐ ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

☐ ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsables de la gestion des données :

ABIVEN, Karine, IdHAL : karine-abiven; ORCID : https://orcid.org/0000-0001-9518-1040, Université Sorbonne,Sens-Texte-Informatique-Histoire (STIH), UR 4509, France

LEJEUNE, Gaël, IdHAL : gael-lejeune ; ORCID : https://orcid.org/0000-0002-4795-2362, Sorbonne Université, Sens-Texte-Informatique-Histoire (STIH), UR 4509, France

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Equipe engagée dans la gestion des données à différentes étapes du projet :
Actuels :

- Karine Abiven MCF en Langue Française
- Gaël Lejeune MCF en Informatique
- Jean-Baptiste Tanguy Doctorant en Humanités Numériques
- Alexandre Bartz, Ingénieur

Anciens :

- 2021: Mélanie Lecha, M2 Humanités Numériques, ENS Lyon/ENSSIB
- 2021 : Camille Roblin, M2 Humanités Numériques, Lyon 2/ENSSIB
- 2021: Amélie Hip, vacataire (retranscriptions)
- 2019-2020: Sylia Kecili, stagiaire (M1 TILDE, Paris 13) sur les problématiques d'OCR
- 2018-2019 :
    - Anaëlle Baledent, stagiaire (M2) sur la datation, actuellement en thèse d'Informatique à l'Université de Caen
    - Nicolas Hiebel, stagiaire (L3) sur la datation, actuellement en M2 Langue et Informatique à Sorbonne Université
    - Jamiilah Patel, stagiaire (M1), sur la structuration des métadonnées, Masterante en Littérature à Sorbonne Université

Évaluation des coûts : 203.5 KE

- Masse salariale, via le DIM STCN et l'OBVIL (2 stagiaires, 1 vacataire non étudiante, 1 IGE) : 61 KE
- Postes de travail pour doctorant et stagiaire (CORLI, faculté des lettres Sorbonne Université) : 5 KE
- Numérisation de 800 documents avec les standards des bibliothèques publiques (IUF) : 16 KE
- Thèse (financée par la région via le DIM STCN) : 100 KE
- Contribution de la Bibliothèque Mazarine estimée à 21 KE (en Personnel : Mise en œuvre et suivi de la numérisation des sources, production et structuration des métadonnées, contrôle qualité de la numérisation ; Elaboration et maintenance de la Bibliographie des mazarinades [BM], référentiel en ligne dont l'ouverture est programmée pour juin 2019. En matériel : mise à disposition du matériel de numérisation)
- Conférences (Bordeaux 2020 : financé par Cahier, Orléans 2021 par l'IUF) : 500 E

6) Archivage des données

☐ ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

☐ ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

Zenodo pour les fichiers XML (qui applique les principes FAIR)
ou CINES : https://www.cines.fr/archivage/ (mais peut-être surdimensionné).

Durée de conservation des données

Illimité

Volume des données à conserver

Volume peu important (<5 Go si l'on ne compte que les fichiers XML-TEI : version actuelle et version lemmatisée).
Les documents (fac-similes numériques) sont hébergés sur les plateformes des bibliothèques numériques et non par le projet.

Coûts alloués à la conservation

À voir en fonction de la plateforme choisie.

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

Données en accès ouvert (pour les limitations, voir les licences des bibliothèques numériques ci-dessus)

7) Partage des données à l'issue du projet

☐ ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

☐ ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR: les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

Les données pourraient servir à alimenter des bases textuelles :

- Sur la langue pré-classique, la langue du milieu du 17e siècle étant un maillon intéressant pour l'histoire du français entre la langue renaissante et la langue dite "classique". Les mazarinades sont un laboratoire discursif qui brassent d'innombrables genres de discours.
- Sur les genres littéraires du 17e siècle.
- Sur les pamphlets ou imprimés non-livres de l'ancien régime (par exemple en réseau avec la Newberry French Pamphlet Collection, citée plus haut).

Eléments d'accompagnement qui permettent la réutilisation des données.

Eléments d'accompagnement qui permettent la réutilisation des données.
Github : https://github.com/Antonomaz
Wiki : http://stih-sorbonne-universite.fr/dokuwiki/doku.php?id=antonomaz

Publications sur les données destinées à en améliorer l'exposition

Voir les premières parties des articles: Abiven, Lejeune 2021a et Tanguy, Abiven, Lejeune 2021b. Un data paper à part entière pourra être envisagé à la fin de l'obtention du jeu de données.

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

Données publiques, accès ouvert à l'issue du projet.

Plan de gestion de données

Schola Rhetorica

Table des matières

---

1) Plan de gestion de données (PGD) du projet SCHOLA RHETORICA

Présentation de la section

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

☐ ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

NOILLE, Christine, ISNI : 0000000109073343 ; IdRef : https://www.idref.fr/032141025
professeure, Sorbonne Université, UMR 8599 CELLF, France
Rôle dans le projet : direction

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD projet SCHOLA-RHETORICA version de ce PGD est actuellement prévue

2) Présentation du projet et responsabilités

☐ ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040),
Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

SCHOLA RHETORICA

Responsable du projet (principal researcher) et unité de rattachement

NOILLE CHRISTINE, professeure, Sorbonne Université, UMR 8599 CELLF, France; direction de SCHOLA-RHETORICA

Financeur(s) du projet et type de financement

2012: ANR Hermès (porteur: Pr. F. Lavocat, Paris 7); 2014: Corpus CAHIER; 2012-2018: UMR 5316 Litt&Arts; 2018---- : UMR 8599 CELLF + UMR 5316 Litt&Arts

Référence de la convention de financement

Institution / organisme / unité porteuses du projet

Unités co-porteuses: Sorbonne Université et Université Grenoble Alpes (convention de coportage en cours de signature)

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Partenaires institutionnels (d'après : https://schola-rhetorica.org/sr/A-propos/partenaires-et-credits)

- Université Grenoble Alpes
- Sorbonne Université
- CNRS
- UMR 5316 Litt&Arts
- UMR 8599 CELLF
- Labex OBVIL
- ANR
- TGIR Huma-Num / Consortium CAHIER

Descriptif et objectif(s) du projet

Schola Rhetorica, la recherche en rhétorique

De l'Antiquité au début du XXe siècle, la rhétorique s'élabore et s'enseigne par les manuels, les commentaires, les exercices, l'imitation. Le projet Schola Rhétorica rassemble des ressources éditoriales des XVIe-XIXe siècles (voir Présentation et description du corpus) pour approfondir l'ancienne rhétorique comme art de parler, comme art de lire et comme art d'écrire.

À l'école de la rhétorique

Trois types de ressources électroniques, trois axes de consultation du corpus :

- Définir les termes, avec le **GLOSSAIRE**
- Analyser les textes, avec les **COMMENTAIRES**
- Comprendre le système rhétorique, avec les **TRAITÉS**

Dates et durée

Date de début de financement et de début des travaux : 2012
Date de fin de financement et de fin des travaux : 2026

Mots clés du projet

- Rhétorique
- Humanisme
- Commentaires / Rhétorique – Commentaires de textes
- Glossaire
- Traités / Rhétorique – Traités

Publications (articles, pré-proposition, site web, …)

Site web du projet : https://schola-rhetorica.org/

Interface de consultation des textes : http://schola-rhetorica.fr/dev/#textes//langues/fr-fr-fr

Tous les articles des membres du projet, publiés sur la revue *Exercices de rhétorique* (https://journals.openedition.org/rhetorique/)

3) Présentation et description du corpus

☐ ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

SCHOLA RHETORICA

Présenter et décrivez le corpus

Le corpus de Padoue

Le corpus de Padoue, que nous éditons ici dans la partie Commentaires est un ensemble très cohérent de quatre ouvrages publiés à Padoue de 1689 à 1729, chez le même éditeur, les Presses du Séminaire (ultérieurement imprimerie de Giovanni Manfrè), à savoir d'une part les trois commentaires rhétoriques de M. A. Ferrazzi, et d'autre part l'édition de la Rhétorique d'Aristote sur laquelle il s'appuie :

- 1689 : Aristote, De arte rhetorica, texte grec et en vis-à-vis la traduction latine de Marcantonio Majoragio (1514-1555), divisée en longs textus (paragraphes) numérotés
- 1694 : le commentaire de Ferrazzi sur 180 discours extraits de l'Histoire romaine de Tite-Live, *Exercitationes rhetoricae in orationes Titi Livii Patavini* (de nombreuses rééditions, par exemple dix rééd. de 1707 à 1710)
- 1694 : le commentaire de Ferrazzi sur 88 discours extraits de l'Énéide de Virgile, *Exercitationes rhetoricae in praecipuas P. Virgilii Maronis orationes, quae in Aeneidum libris leguntur* (de nombreuses rééditions, par exemple sept rééd. de 1720 à 1780, en Bavière)
- 1729 : le commentaire de Ferrazzi sur la totalité des 56 discours de Cicéron, *M. T. Ciceronis orationum cum argumentis, animadversionibus, et analysi M. Antonii Ferratii* (pas de rééd., mais l'éd. princeps est très répandue dans les bibliothèques)

Un corpus méthodique

Trois traits font de cet ensemble d'ouvrages édité par le Séminaire de Padoue un corpus méthodique.

- La circularité entre théorie et pratique : la Rhétorique d'Aristote est donnée avec un paragraphe propre à Padoue, et les analyses de discours que publie Ferrazzi

renvoient uniquement à cette édition de la Rhétorique, avec une insistance toute particulière sur les passions du livre II (le pathos).

- La régularité : quoique très nombreuses, les analyses suivent constamment la même procédure et emploient toujours le même type de vocabulaire critique, dont elles accentuent la monotonie de façon délibérée, pédagogique.
- La masse : 180 discours tirés de Tite-Live, 88 de l'Énéide, les 56 discours de Cicéron, dont certains particulièrement longs. Une telle masse correspond aux nécessités internes de l'enseignement de la rhétorique. Elle permet de ramener à du sériel, donc à du reconnaissable, la diversité indéfinie et déroutante des situations rhétoriques concrètes. Et pour les étudiants d'aujourd'hui, elle remplace aussi le maître de l'époque, en signalant de façon récurrente quelles étaient à ses yeux les catégories vraiment importantes.

Corpus complémentaires

1. Une bibliothèque d'une vingtaine de traités de rhétorique des 17^{e}-19^{e} siècles : une dizaine est en cours de numérisation
2. Un glossaire collaboratif (indexant pour chaque terme des définitions attestées dans les traités depuis l'antiquité).

L'ensemble répondra aux standards d'un encodage TEI et d'une annotation enrichie (voir infra). Le point fort de cette plateforme numérique est non seulement de rassembler l'ensemble des corpus qui ont constitué l'empire de la rhétorique (intérêt patrimonial) mais d'offrir, grâce à l'élaboration d'interfaces dynamiques, une multitude de parcours pour une pluralité d'approches.

Période couverte par le corpus, auteur(s) concerné(s)

Période: XVIe-XIXe siècles. Auteurs: rhétoriciens de toute l'Europe

Organisation du corpus

Trois interfaces éditoriales : l'interface des traités, des commentaires, du glossaire.

Mode de collecte et origine des données

Données libres de droit ; OCRisation sous word ; relectures / corrections, puis encodage.

Etat du corpus numérique

- Pour les commentaires: ⅓ du corpus est en ligne (les commentaires sur Virgile)
- Pour les traités: une dizaine de traités sont édités
- Pour le glossaire collaboratif (wiki), plus de trois cents entrées sont multi-renseignées

Types de données:

Données textuelles

Volumétrie

- Commentaires: en l'état, 1 million de signes, à termes 3,5 millions
- Traités: en l'état, 7 millions de signes, à terme 12 millions
- Glossaire: en l'état 2 millions de signes, à terme 4 millions

En 2021, la volumétrie de l’ensemble des données est estimée à 3 Go.

Métadonnées, créées et standards et formats utilisés

Formats : MySQL,
Standards : TEI (sortie TEI prévue pour les Traités)

Les métadonnées descriptives, administratives et techniques

Pour les textes, glossaire et commentaires : Auteur / Titre français et latin / Edition française et latine / Sommaire / Séquençage.

Les métadonnées structurelles et l’annotation sémantique

Les traités ont fait l’objet d’un balisage sémantique au format interne avec la possibilité d'une transformation XML-TEI.

Référentiels d’indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Indexation libre: références internes, y compris entre textes et termes rhétoriques.

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

☐ ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

☐ ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Accès, partage et limites des données

L'accès au site est libre sous licence Creative Commons BY-NC-SA.
Le **GLOSSAIRE** est un atelier numérique collaboratif (de type wiki), s'appuyant sur la seule exploitation de données référencées libres de droit.
Les **COMMENTAIRES** et les **TRAITÉS** : les textes sont numérisés à partir d'éditions libres de droit. Pour les **COMMENTAIRES**, les traductions sont de l'équipe éditoriale, sous licence Creative Commons BY-NC-SA.

5) Responsabilités et ressources pour la gestion des données

☐ ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

☐ ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

NOILLE, Christine, professeure, Sorbonne Université, UMR 8599 CELLF, France
Rôle dans le projet : Responsable du projet

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

Moyens humains : voir https://schola-rhetorica.org/sr/A-propos/l-equipe

6) Archivage des données

☐ ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

☐ ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

La question de l'archivage des données sera intégrée au projet de développement de Schola Rhetorica déposé en 2022.

Plateforme pour l'archivage pérenne des données

xxxxxxxxx

Durée de conservation des données

xxxxxxxxx

Volume des données à conserver

xxxxxxxxx

Coûts alloués à la conservation

xxxxxxxxx

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

xxxxxxxxx

7) Partage des données à l'issue du projet

☐ ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

☐ ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR: les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

La fairisation des données sera amorcée dans le projet 2022. L'accent sera mis sur la trouvabilité (nous disposons en général de bons systèmes de liens entre les textes) et l'accessibilité (on envisage de constituer un "usuel" en ligne, TLF, pour la rhétorique).

Potentiel de réutilisation des données

xxxxxxxxx

Eléments d'accompagnement qui permettent la réutilisation des données.

xxxxxxxxx

Publications sur les données destinées à en améliorer l'exposition

Dans une revue numérique (sur Openeditions), *Exercices de rhétorique*, créée et dirigée par la responsable de Schola-Rhetorica (Christine Noille) et son co-responsable pour la partie rhétorique (Francis Goyet)

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

xxxxxxxxx

Plan de gestion de données

E-Stampages

Table des matières

---

1) Plan de gestion de données (PGD) du projet E-Stampages

☐ ***Présentation de la section***

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

☐ ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

BRUNET, Michèle, IdHAL : michele-brunet ; ORCID : https://orcid.org/0000-0003-1818-5237, Université Lyon 2, Histoire et Sources des Mondes Antiques (HiSoMA, UMR 5189),Lyon, France
Rôle dans le projet : Responsable du projet

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD projet E-Stampages
2 versions de ce PGD est actuellement prévue

2) Présentation du projet et responsabilités

☐ ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040),
Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

E-STAMPAGES, Archivage à long terme et création d'une Bibliothèque numérique d'estampages grecs.

Responsable du projet (principal researcher) et unité de rattachement

BRUNET, Michèle, IdHAL : michele-brunet ; ORCID : https://orcid.org/0000-0003-1818-5237, Université Lyon 2, Histoire et Sources des Mondes Antiques (HiSoMA, UMR 5189),Lyon, France

Financeur(s) du projet et type de financement

Projet financé dans le cadre de l'appel à projets Bibliothèque Scientifique Numérique 5 (BSN5)

Référence de la convention de financement

xxxxxxxxx

Institution / organisme / unité porteuses du projet

- École française d'Athènes, porteur et gestionnaire dans le cadre de l'appel BSN 5, Athènes, Grèce

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Le projet s'effectue au sein d'un consortium financé dans le cadre de l'appel à projets 2014 Bibliothèque Scientifique Numérique 5 (BSN5), pour une durée de 18 mois, janvier 2015-juin 2016.
Membres du consortium :

- École française d'Athènes, porteur et gestionnaire dans le cadre de l'appel BSN 5, Athènes, Grèce
- Laboratoire Histoire et Sources des Mondes Antiques, HiSoMA, UMR 5189 du CNRS, Lyon, France
- Pôle Système d'Information et Réseau de la Maison de l'Orient et de la Méditerranée - Jean Pouilloux, Lyon, France
- The Digital Epigraphy & Archaeology Project, Université de Floride, Gainsville, Floride, USA
- (Depuis 2017) Laboratorio di Epigrafia Greca du Dipartimento di Studi Umanistici de l'Università Ca' Foscari, Venise, Italie

Descriptif et objectif(s) du projet

D'après : https://cahier.hypotheses.org/e-stampages et https://www.e-stampages.eu/s/e-stampages/page/projet

Création d'une bibliothèque numérique d'environ 6000 estampages d'inscriptions grecques sélectionnés sur des critères communs, issues des collections de l'EFA et d'HiSoMA et en provenance des sites de Thasos, Délos, Delphes et Philippes. Les images sont disponibles en version 2D et 3D, l'outil de visualisation diverses fonctionnalités destinées à faciliter le travail des épigraphistes, en particulier un variateur d'ombrage.
Objectifs :

- un archivage à long terme des originaux sous une forme dématérialisée (images .tiff), à des fins de conservation
- la diffusion en libre accès sur le Web de cette documentation scientifique essentielle pour le travail des épigraphistes, associant aux vues 2D et 3D tout un ensemble de métadonnées modélisées et structurées. L'intention est donc de créer une ressource documentaire uniquement centrée sur les estampages : il ne s'agit ni de rééditer des textes épigraphiques ni de les commenter.

Les métadonnées documentaires récupérées sur les bases de données existantes ont été modélisées, enrichies et enregistrées dans des formats et langages standardisés pour le Web sémantique, afin d'en faciliter l'interopérabilité avec d'autres catégories de ressources, complémentaires pour l'étude des inscriptions grecques — éditions électroniques des textes en TEI/EpiDoc, photographies des inscriptions, systèmes d'information géographique, etc., qui pourront être ultérieurement reliées à l'ectypothèque E-STAMPAGES.

Dates et durée

Date de début de financement et de début des travaux : janvier 2015
Date de fin de financement et de fin des travaux : fin de financement en juin 2016. Toutefois le projet se poursuit, E-Stampages est inscrit dans l'axe prioritaire de recherche Outils numériques de la recherche du programme scientifique quinquennal de l'École française d'Athènes 2017-2021 et le site E-STAMPAGES est toujours en cours d'enrichissement et d'amélioration.

Mots clés du projet

- Estampage
- Epigraphie
- Bibliothèques numériques
- Métadonnées
- Visualisation
- Ectypothèque

Publications (articles, pré-proposition, site web, …)

Site web du projet : https://www.e-stampages.eu/

Listes des articles publiés par le projet :

Antonetti, Claudia, Michèle Brunet, Eloisa Paganoni. 'Collezioni Di Calchi Epigrafici: Una Nuova Risorsa Digitale'. *Axon no. 2* (23 December 2019). http://doi.org/10.30687/Axon/2532-6848/2019/02/004

Brunet, Michèle, "E-stampages : la mise en ligne des collections d'estampages. Une nouvelle ressource pour l'étude des inscriptions grecques", *Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres vol. 2019*, 2021

Bozia, Eleni, "Assessing the role of digital libraries of squeezes in epigraphic studies", In *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference on Digital and Traditional Epigraphy in Context. 27-29 January 2016. Rome, Italy*. p373-378, Sapienza Università Editrice, 2016 https://doi.org/10.13133/978-88-9377-021-7

Bozia, Eleni, "Augmenting the workspace of epigraphists: an interaction design study", in *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference on Digital and Traditional Epigraphy in Context. 27-29 January 2016. Rome, Italy*. p171-182, Sapienza Università Editrice, 2016 https://doi.org/10.13133/978-88-9377-021-7

Adeline Levivier, Elina Leblanc et Michèle Brunet, « E-STAMPAGES : archivage et publication en ligne d'une ectypothèque d'inscriptions grecques », *Les nouvelles de l'archéologie* [En ligne], 145 | 2016. http://journals.openedition.org/nda/3801 DOI : https://doi.org/10.4000/nda.3801

Bozia, Eleni, Barmpoutis, Angelos, Wagman, Robert S., “OPEN-ACCESS EPIGRAPHY, Electronic Dissemination of 3D-digitized Archaeological Material”, in *Information Technologies for Epigraphy and Cultural Heritage: Proceedings of the first EAGLE International Conference, Paris, 2014*
https://f-origin.hypotheses.org/wp-content/blogs.dir/31/files/2014/09/Open-Access-Epigraphy.pdf

Autres livrables (guides, recommandations, etc.) : xxxxxxxxx

3) Présentation et description du corpus

☐ ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

E-STAMPAGES, Archivage à long terme et création d'une Bibliothèque numérique d'estampages grecs.

Présenter et décrivez le corpus

L'estampage est l'empreinte d'une inscription réalisée sous forme d'un moulage à l'aide d'un papier vergé sans colle. Ce matériel est indispensable aux épigraphistes pour se constituer une collection de sources consultable à distance, pratique à stocker et à transporter.
Le corpus est constitué d'environ 6000 estampages issus des collections de l'Ecole Française d'Athènes (EFA) et de l'HiSoMa, sélectionnés selon plusieurs critères pour être publiés sur https://www.e-stampages.eu/.

- **La collection de l'EFA**

Dans le cadre de la première phase du programme E-STAMPAGES, ont été retenus pour numérisation et diffusion sur le web
- les estampages les plus anciens et les estampages endommagés, dont la préservation est prioritaire
- les estampages « historiques », provenant des sites archéologiques dont l'exploration a été confiée à l'EFA depuis le XIXe siècle
- les estampages reliés aux corpus d'inscriptions dont l'EFA est l'éditeur scientifique

Collection	Dates extrêmes	Volumétrie
Delphes	1896-1987	2662
Thasos	1907-2014	1101
Délos	1903-1980	894

Béotie	1884-1891	218
Philippes	1914-1984	130
Asie Mineure	1884-1886	99
Chalcidique	1891	8
Etolie	1885	2
Crète	1889	1

- **La collection d'HiSoMa**

Dans le cadre de la première phase du programme E-STAMPAGES, ont été retenus pour numérisation et diffusion sur le web
- les estampages reliés à la collection de l'EFA par une histoire institutionnelle commune
- les estampages du fonds Jean Pouilloux, complémentaires des ensembles de Thasos et de Delphes conservés à l'École française d'Athènes

Collection	Dates extrêmes	Volumétrie
Phocide-Delphes	1975-1993	549
Thasos	1946-1956	449
Délos	1903-1925	3

Période couverte par le corpus, auteur(s) concerné(s)

Dates des estampages : 1884 - 2014

Organisation du corpus

xxxxxxxxx

Mode de collecte et origine des données

Voir https://www.e-stampages.eu/s/e-stampages/page/projet
Près de 6000 estampages d'inscriptions, provenant des quatre grands sites explorés par l'EFA Délos, Delphes, Thasos et Philippes, ont été numérisés et des vues 3D ont été créées, suivant le protocole développé par le Digital Epigraphy and Archeology project.
Les métadonnées documentaires, récupérées dans les bases de données préexistantes, ont été re-documentarisées (modélisées, enrichies et structurées) dans des formats standardisés, afin d'en faciliter l'interopérabilité à venir avec d'autres catégories de ressources, complémentaires pour l'étude des inscriptions grecques : éditions électroniques des textes en TEI/EpiDoc, photographies des inscriptions, systèmes d'information géographique, etc., qui seront reliées à l'ectypothèque E-STAMPAGES.

Etat du corpus numérique

Le corpus corpus est-il complet ? Si non pour quelles raisons ? L'état physique des sources était-il altéré, ce qui peut impacter sur la qualité de leur version numérique ?

xxxxxxxxx

Types de données

Images (estampages) numérisées / photographies des inscriptions, format jpeg et png
Jeux de données textuelles en pdf
Transcriptions XML-TEI

Volumétrie

Entre 6000 et 6200 épigraphies numérisées
xxxxx… photographies
Poids de l'ensemble des images ?
Volume autres fichiers (pdf, xml) ?

Modifications effectuées sur les données, versions, …

xxxxxxxxx

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Un thesaurus multilingue est élaboré grâce à l'outil OpenTheso, développé et maintenu au sein du réseau FRANTIQ par l'équipe de la Plateforme Tête de Réseaux Documentaires de la Maison de l'Orient et de la Méditerranée-Jean Pouilloux.

Métadonnées, créées et standards et formats utilisés

Les métadonnées sont récupérées des bases de données existantes, enrichies et redocumentarisées pour les rendre interopérables.
Standard EAD
Schémas RDF, XML
Ontologie CIDOC CRM

Les métadonnées descriptives, administratives et techniques

xxxxxxxxx

Les métadonnées structurelles et l'annotation sémantique

xxxxxxxxx

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

☐ ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

☐ ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

xxxxxxxxx

Accès, partage et limites des données

xxxxxxxxx

Droits (voir https://www.e-stampages.eu/s/e-stampages/page/credits_rights)
Mise à disposition du contenu original créé ou redocumentarisé pour l'ectypothèque E-STAMPAGES

- Les images 2D des estampages
- Les images 3D des estampages
- Les métadonnées attachées aux estampages, aux inscriptions et aux artefacts

sont publiées selon les termes de la Licence Creative Commons Attribution - Partage dans les Mêmes Conditions 4.0 International

-- Sauf indication spécifique, les photographies proviennent pour la plupart de la photothèque de l'École française d'Athènes. Toutes sont créditées à leurs auteurs respectifs. Pour toute demande de reproduction imprimée, se reporter à la page contact sur le site de l'EFA
-- Artefacts supports des inscriptions : tous droits reservés aux Musées dépositaires

- Musées archéologiques de Grèce sous la tutelle du Ministère de la Culture et du Sport de Grèce - Ελληνική Δημοκρατία, Υπουργείο Πολιτισμού & Αθλητισμού
- Musée du Louvre, Paris

Pour la consultation des artefacts originaux, prendre contact avec les musées dépositaires.

5) Responsabilités et ressources pour la gestion des données

☐ ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

☐ ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsable de la gestion des données :
BRUNET, Michèle, IdHAL : michele-brunet ; ORCID : https://orcid.org/0000-0003-1818-5237, Université Lyon 2, Histoire et Sources des Mondes Antiques (HiSoMA, UMR 5189),Lyon, France

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

4 stages de Master pro ont été effectués en renfort de l'équipe projet :

- Elina Leblanc, Master Pro « Patrimoine écrit et Edition numérique », Université de Tours, stage de 6 mois (avril-septembre 2015). Réflexion sur la structuration des métadonnées et leur modélisation en relation avec les ontologies et les formats standards, choix du CMS Omeka pour la diffusion.
- Evita Dionysopoulou, Master Pro « Archives et Images », Université Jean Jaurès Toulouse, stage de 4 mois (avril-juillet 2016). Traitement d'une partie du fonds conservé à HiSoMA dit « Fonds Homolle », inventaire des 4857 estampages, identification et saisie des métadonnées de 1282 documents, numérisation de 650. Création d'une exposition virtuelle (V.0) sur Homolle et le travail de l'épigraphiste avec le CMS Omeka.
- Hélène Vuidel, Master Pro Sibist, Enssib, stage de 4 mois (février-mai 2017) : contribution à la création d'un thesaurus sous OpenTheso.

- Rémy Ienco, Master Pro Métiers de la science des Patrimoines, CESR Université de Tours (stage 1 en convention avec CNRS/UMR 5189 Hisoma, 280 heures, mars-mai 2021 puis stage 2, 210 heures, juin-juillet 2021, en convention avec le Musée du Louvre) : préparation des métadonnées avant intégration dans le CMS pour la série Thasos, mise en œuvre du module de cartographie Mapping, finalisation du Thesaurus EpiVoc sous OpenTheso (commun aux programmes E-stampages et IGLouvre)

6) Archivage des données

☐ ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

☐ ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

xxxxxxxxx

Durée de conservation des données

xxxxxxxxx

Volume des données à conserver

xxxxxxxxx

Coûts alloués à la conservation

xxxxxxxxx

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

xxxxxxxxx

7) Partage des données à l'issue du projet

☐ ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

☐ ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR: les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

xxxxxxxxx

Eléments d'accompagnement qui permettent la réutilisation des données.

xxxxxxxxx

Publications sur les données destinées à en améliorer l'exposition

xxxxxxxxx

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

xxxxxxxxx

Plan de gestion de données

Ichtya

Table des matières

---

1) Plan de gestion de données (PGD) du projet ICHTYA

☐ ***Présentation de la section***

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

☐ ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

Gauvin, Brigitte, ISNI : 0000000061551536, Université de Caen-Normandie, CRAHAM (UMR 6273), Caen, France
Rôle dans le projet : co-responsable du projet

Buquet, Thierry, IdHal : thierry-buquet, Orcid : 0000-0003-2956-8217 ; CNRS, CRAHAM (UMR 6273), Caen, France
Rôle dans le projet : co-responsable du projet

Buard, Pierre-Yves, Pôle Document numérique, MRSH, Université de Caen Normandie – CNRS (USR 3486), Caen, France
Rôle dans le projet : Responsable technique et éditorial

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD projet Ichtya
2 versions de ce PGD sont actuellement prévues.

2) Présentation du projet et responsabilités

☐ ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040),
Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

ICHTYA Corpus de traités latins d'ichtyologie et histoire des savoirs sur la faune aquatique

Responsable du projet (principal researcher) et unité de rattachement

Gauvin, Brigitte, IdHAL : brigitte-gauvin ; Université de Caen-Normandie, CRAHAM (UMR 6273), Caen, France

Rôle dans le projet : co-responsable du projet

Buquet, Thierry, IdHal : thierry.buquet, Orcid : 0000-0003-2956-8217 ; CNRS, CRAHAM (UMR 6273), Caen, France

Rôle dans le projet : co-responsable du projet

Buard, Pierre-Yves, Pôle Document numérique, MRSH, Université de Caen Normandie – CNRS (USR 3486), Caen, France

Rôle dans le projet : Responsable technique et éditorial

Financeur(s) du projet et type de financement

CRAHAM (financement du laboratoire)

Référence de la convention de financement

Institution / organisme / unité porteuses du projet

CRAHAM - Université Caen - CNRS

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

- GDRI Zoomathia consacré à l'étude de la transmission des savoirs zoologiques, Antiquité-Moyen Âge (dir. A. Zucker, CEPAM, université de Nice).
- Sourcencyme (Sources des Encyclopédies Médiévales), piloté par Isabelle Draelants, à l'Institut de recherche et d'histoire des textes (Paris).
- Le projet ICHTYA bénéficie du soutien informatique et éditorial de la Maison de recherche en sciences humaines (Pôle document numérique, resp. Pierre-Yves Buard), et des Presses universitaires de Caen.
- Consortium CAHIER (Corpus d'Auteurs pour les Humanités, Informatisation, Édition, Recherche. Littérature, philosophie, histoire de l'art) (TGIR Huma-Num)

Descriptif et objectif(s) du projet

L'objectif du projet Ichtya est la mise en ligne d'un corpus de traités latins d'ichtyologie, permettant d'apprécier le contenu du savoir zoologique véhiculé pendant l'Antiquité et le Moyen Âge, avant la publication des grands traités d'ichtyologie du XVIe siècle. Il se fonde sur la base de la *Biblioteca Ichthyologica* de Pierre Artedi (collaborateur de Linné et fondateur de l'ichtyologie moderne). Un des objectifs du projet Ichtya serait de reconstituer la *Biblioteca Ichthyologica* en la documentant et de façon générale, proposer une étude de l'histoire de la bibliographie ichtyologique.
La Bibliothèque Ichtya s'accompagne d'un thesaurus des noms de poissons latins et vernaculaires figurant dans les textes latins et d'une bibliographie spécifique établie sur zotero.org.
L'édition critique des traités d'ichtyologie médiévaux est réalisée en XML-TEI, permettant des publications multi-modales, consultables en ligne et disponibles sous forme de livre papier.

Dans sa première phase, le projet Ichtya est dédié aux textes de la période médiévale. Il prévoit d'associer des éditions critiques ponctuelles (paru : *Hortus sanitatis*, lib. 4 *De piscibus* ; en cours Thomas de Cantimpré, *Liber de natura rerum*, 6 − 7 ; Albert le Grand, *De animalibus*, 24) et une bibliothèque documentaire, la Bibliothèque Ichtya (textes sources ou similaires : Pline, nat. 9 et 32 ; Ambroise, *Hexameron*, 5 ; Basile, *Hexameron*, 7-8, Isidore de Séville *Etymologiae* 12, 6 ; Vincent de Beauvais, *Speculum naturale*, 17) dont les outils d'exploration doivent pouvoir être mutualisés. Dans des phases ultérieures, il s'agira a) de replacer ce matériel dans la perspective de la Bibliotheca ichtyologica de Peter Artedi en tenant compte de la documentation dont il disposait (éditions consultées), b) d'envisager les traités de la Renaissance. Un dernier axe de recherche concerne l'étude des synonymies et polyonimies entre les noms de poissons dans les traités ichtyologiques, en analysant

comment les auteurs, à partir du Moyen Âge, indiquent ces équivalences de désignation et leurs méthodes d'identification zoologique.

Dates et durée

Date de début de financement et de début des travaux : 2009
Date de fin de financement et de fin des travaux :

Mots clés du projet

- poissons
- ichtyologie
- histoire de la zoologie / Zoologie – Histoire
- philologie
- édition critique
- édition électronique
- XML-TEI / Text Encoding Initiative
- EAD / Description archivistique encodée
- encyclopédies médiévales / Encyclopédies – Moyen âge
- latin

Publications (articles, pré-proposition, site web, …)

Page web descriptive du projet : https://www.craham.cnrs.fr/ichtya/

Sites et outils en ligne :
2020 - Mise en ligne de la Bibliothèque Ichtya : https://ichtya.unicaen.fr/
2020 - Mise en ligne du Thesaurus Ichtya des noms latins de poissons et de créatures aquatiques figurant dans les textes latins d'ichtyologie antique et médiévale : https://ichtya.unicaen.fr/lab/thesaurus/
2015 - Mise en ligne de la Bibliographie collaborative d'Ichtya : https://www.zotero.org/groups/356871/ichtya/library
2013 - Edition multimodale de l'*Hortus sanitatis* (livre des poissons) : http://www.unicaen.fr/puc/sources/depiscibus/accueil

Listes des articles publiés par le projet :
Voir https://www.craham.cnrs.fr/ichtya/

3) Présentation et description du corpus

☐ ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

Ichtya - Corpus de traités latins d'ichtyologie et histoire des savoirs sur la faune aquatique

Présenter et décrivez le corpus

Le corpus constitutif de la bibliothèque numérique Ichtya rassemble des textes antiques et médiévaux latins consacrés à l'ichtyologie qui furent publiés dans l'Antiquité, au Moyen Âge et à la Renaissance. Elle s'inspire de la Bibliotheca Ichtyologica de Peter Artedi (1705-1735) et a pour vocation de mettre en ligne et à disposition des lecteurs un corpus latin consacré au savoir ichtyologique.
Le corpus a été océrisé et entièrement encodé en XML-TEI. Il est actuellement composé de 21 textes (8 en ligne publique) et 5 traductions (2 en ligne publique). Ces textes, attribués ou anonymes, sont de nature hétérogène (textes sacrés, encyclopédies, dialogues, poèmes) et de longueur très variable (fragments, chapitres, livres entiers).
Les textes sont annotés (indexation des zoonymes et des citations de sources). L'indexation des zoonymes s'appuie sur la constitution d'un index ichtyologique encodé en XML TEI.

Période couverte par le corpus, auteur(s) concerné(s)

Antiquité - Moyen Âge, Renaissance. Auteurs multiples

Organisation du corpus

Les textes sont classés en fonction de leur période de publication (Antiquité, Moyen-Age, Époque moderne) et les traductions font l'objet d'une section à part.

Mode de collecte et origine des données

Les données textuelles ont fait l'objet d'une campagne de reconnaissance de caractères.

Etat du corpus numérique

Extraits et textes complets encodés, pas d'images numérisées.

Types de données:

Données textuelles transcrites et enrichies en XML-TEI

Volumétrie

2000 pages web en janvier 2020

Modifications effectuées sur les données, versions, …

Le corpus a été entièrement encodé en XML-TEI

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Le thesaurus est construit en XML-TEI. Il est composé d'autant de fichiers XML (notices) que de formes latines (au nominatif) ou vernaculaires rencontrées dans le corpus de la bibliothèque Ichtya. Chaque notice présente toujours la référence précise à la source dans laquelle le terme apparaît. Cette indication de source s'accompagne, autant que possible, d'une ou plusieurs identifications et, pour les appellations latines et grecques, de la référence scientifique qui valide ces identifications. Ces identifications peuvent être accompagnées d'une note de commentaire. Les notices peuvent aussi présenter deux sortes de renvois sous forme de liens : d'une part à la forme principale en cas de paronymie, de variante orthographique ou de forme vernaculaire, indication qui figure en tête de la fiche, à la place de l'identification ; de l'autre aux autres termes désignant le même animal sous un autre nom. L'indexation par le biais du format XML permet de faire des liens directement d'une forme à l'autre. Ce thesaurus fournit un outil de première utilité pour l'étude des synonymies et polyonymies entre les noms de poissons dans les traités ichtyologiques.
Chaque forme de nom de poisson ou créature aquatique rencontré dans le corpus de la bibliothèque Ichtya fait donc l'objet d'une notice XML-TEI et chaque occurrence est liée à une notice du thesaurus.
Génération de graphes RDF : À partir de l'encodage en arbre XML-TEI, des graphes ont pu être générés pour chaque notice, permettant de mettre en évidence les liens établis par les chercheurs : identification ; variantes graphiques ; notices en relation.
Ce système d'enrichissement sémantique permet la création et la visualisation de réseaux de notices qui se révèlent un matériel intéressant d'exploitation du corpus pour les chercheurs.

Métadonnées, créées et standards et formats utilisés

Les métadonnées descriptives, administratives et techniques

Descriptions des fonds en format XML selon les recommandations de la TEI.

Les métadonnées structurelles et l'annotation sémantique

L'ensemble du corpus (textes et thesaurus) est encodé en XML TEI avec un schéma dédié. L'ensemble de l'outillage développé et sa documentation sont accessibles à cette adresse : https://www.unicaen.fr/recherche/mrsh/document_numerique/outils/compilations

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Indexation avec l'appui d'un thésaurus spécifique créé au sein du projet.
Lien vers le thesaurus de zoologie ancienne Thezoo (http://web.cepam.cnrs.fr/opentheso/)

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

☐ ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

☐ ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Accès, partage et limites des données

La mise à disposition des fichiers XML est prévue et dans ce cas l'intégralité des données sera moissonnable et interopérable.
Le dépôt des données dans Nakala est prévu dans l'année à venir.

5) Responsabilités et ressources pour la gestion des données

☐ ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

☐ ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Buard, Pierre-Yves, Pôle Document numérique, MRSH, Université de Caen Normandie – CNRS (USR 3486), Caen, France
Rôle dans le projet : Responsable technique et éditorial

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

6) Archivage des données

☐ ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

☐ ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

xxxxxxxxx

Durée de conservation des données

xxxxxxxxx

Volume des données à conserver

xxxxxxxxx

Coûts alloués à la conservation

xxxxxxxxx

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

xxxxxxxxx

7) Partage des données à l'issue du projet

☐ ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

☐ ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR: les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

xxxxxxxxx

Eléments d'accompagnement qui permettent la réutilisation des données.

xxxxxxxxx

Publications sur les données destinées à en améliorer l'exposition

xxxxxxxxx

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

xxxxxxxxx

Plan de gestion de données

Lafabrev (La Fabrique de la Révolution)

Table des matières

---

1) Plan de gestion de données (PGD) du projet LAFABREV (La Fabrique de la Révolution)

☐ ***Présentation de la section***

Rédaction du PGD démarrée par Cécile Andrisi-Brémon le 11/10/2021
Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

☐ ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

ANDRISI-BRÉMON, Cécile, IdHAL : 1084562 ; IdRef : 193185725 ; ORCID : https://orcid.org/0000-0002-5130-339X, Autoentrepreneuse pour le compte de l'Université de Paris, en lien avec le CERILAC (UPR n°441), Paris, France ; ingénieure d'études en CDD d'octobre 2015 à juin 2019 (Centre Seebacher, CERILAC, Université Paris Diderot)

Rôle dans le projet : coordinatrice numérique (mises en ligne et relectures, en collaboration avec Paule Petitier et Olivier Ritz)

Rédaction et compilation à partir de documents rédigés par Paule Petitier

PETITIER, Paule, IdHAL : 720408 ; ISNI : 0000000117567255 ; IdRef : 033175756 , Université de Paris, CERILAC (UPR 441), Paris, France

Rôle dans le projet : porteuse du projet et responsable scientifique

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France

Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1: 19/10/2021, PGD projet LAFABREV
Trois versions de ce PGD sont actuellement prévues.

2) Présentation du projet et responsabilités

☐ ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040),
Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

La Fabrique de la Révolution (LAFABREV)

Responsable du projet (principal researcher) et unité de rattachement

PETITIER, Paule, IdHAL : 720408, Université de Paris, CERILAC (UPR 441), Paris, France
Rôle dans le projet : porteuse du projet et responsable scientifique

Autre coordinateur scientifique et numérique du projet :

RITZ, Olivier, IdHAL : olivier-ritz ; ORCID : 0000-0001-5492-9403, Université de Paris, CERILAC (UPR 441), Paris, France
Rôle dans le projet : chercheur et coordinateur scientifique et numérique

Membres du projet :

Voir la galerie de membres consultable dans la rubrique « Équipe » du site, où figurent une partie des membres : https://lafabrev-michelet.lac.univ-paris-diderot.fr/equipe

ANDRISI-BRÉMON, Cécile (voir « Auteur du plan de gestion des données » et fiche à la rubrique « Équipe »)

ARNAUD, Émilien (Masterant en histoire, Master de recherche Humanités numériques et Computationnelles, École nationale des Chartes - PSL - ENS - EPHE - EHESS, détenteur d'un précédent Master de recherche en histoire de l'Université Paris 1 Panthéon-Sorbonne - Institut d'Histoire de la Révolution française - Institut d'Histoire Moderne et Contemporaine : travail sur l'index des références bibliographiques)

BERKERY, Charlotte (Docteure en littérature française du XIXe siècle : transcriptions)

FAURE, Claudie (Transcriptrice bénévole, ancienne chargée de recherche Laboratoire Traitement et Communication de l'Information LTCI-CNRS : nombreuses transcriptions. Voir fiche à la rubrique « Équipe »)

LALLIER, Thomas (Développeur web indépendant : développement du site LAFABREV. Voir fiche à la rubrique « Équipe »)

MAGRET, Maryelle (Relectrice-correctrice indépendante : préparation des fichiers XML et complétion des métadonnées à partir des volumes conservés à la BHVP, avant transcription des papiers par les transcripteurs)

MASEDA, Oriane (Détentrice d'un Master 1 Humanités numériques de l'École des chartes : nombreuses transcriptions)

MILOT-PINSON, Cécile (Cursus d'historienne ; désormais professeure des écoles : coordination numérique du projet jusqu'à début 2016 ; mise en place, avec son frère Baptiste Milot, développeur web, de l'outil numérique «chaîne éditoriale» élaboré par Thomas Lebarbé pour le projet «Manuscrits de Stendhal» ; formation des transcripteurs et rédaction d'un manuel d'encodage synthétisé par la suite dans un protocole de transcription récapitulatif)

MASSIP, Luc (Détenteur d'un Master 2 Lexicographie, Terminographie et Traitement Automatique de Corpus LTTAC de l'Université de Lille : correction automatisée d'erreurs dans les fichiers XML en Python et développements sur le site internet (allègement des pages ; mise en place d'une navigation entre entrées d'index et papiers ; mise en place d'une recherche par expressions régulières de type regex ; ajout de la page « Téléchargements » et de la rubrique « Équipe » . Voir fiche à la rubrique « Équipe »)

OUDAI CELSO, Yamina (PhD de l'Université de Venise : transcriptions. Voir fiche à la rubrique « Équipe »)

¨PETITIER, Paule (voir « Responsable du projet » et bientôt voir fiche à la rubrique « Équipe »)

SAFA, Isabelle (Chercheuse, agrégée de lettres modernes et docteur en littérature française, autrice d'une thèse sur le roman historique d'A. Dumas, enseignante en lycée et classes préparatoires à Lille : transcriptions. Voir fiche à la rubrique « Équipe »)

WULF, Judith (Professeur à l'Université de Nantes, autrice d'une thèse sur V. Hugo : transcriptions)

Financeur(s) du projet et type de financement

- DIM STCN (Domaine d'intérêt majeur « Sciences du texte et connaissances nouvelles » de la région Île-de-France : crédits de fonctionnement pour payer les prestataires intervenant sur le projet (coordinatrice numérique et historien en charge de corriger l'index des références bibliographiques)
- UDPN (Usages des patrimoines numérisés) : crédits de personnel pour payer l'ingénieure d'études du projet (coordinatrice numérique) et des vacations de transcription (enseignants-chercheurs et doctorantes)
- Consortium CAHIER : crédits pour payer la numérisation des volumes effectuée par la société Digiscrib ; formations XLST pour l'ingénieure d'études du projet
- CERILAC, URP 441 (Centre d'Études et de Recherches Interdisciplinaires de l'UFR Lettres, Arts, Cinéma) de l'Université de Paris (ex-Paris Diderot) : crédits de personnel pour payer des vacations de transcription
- Université de Paris (ex-Paris Diderot) : crédits de personnel pour payer l'ingénieur d'études du projet (coordinatrice technique)

Référence de la convention de financement

N° DIMSTCN-2019-18

Institution / organisme / unité porteuses du projet

Université de Paris (ex-Paris Diderot)

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

DIM STCN et Consortium CAHIER

Descriptif et objectif(s) du projet

Le programme de recherche LAFABREV vise à constituer un corpus numérique à partir des notes préparatoires de l'*Histoire de la Révolution française* de Jules Michelet.

Dates et durée

Date de début de financement et de début des travaux : 2014 /2015
Date de fin de financement et de fin des travaux : 2021/2022

Mots clés du projet

Jules Michelet ; Histoire ; Révolution française ; XVIIIe siècle ; XIXe siècle

Publications (articles, pré-proposition, site web, …)

Site web du projet : https://lafabrev-michelet.lac.univ-paris-diderot.fr/

Liste des articles publiés par le projet :

Articles de Paule Petitier :

- Grâce aux ressources numériques, on sait mieux comment travaillait Michelet, 21 janvier 2020.
- La Semaine sainte de Michelet. L'émergence de l'idée des fédérations à travers les papiers préparatoires de l'*Histoire de la Révolution française*, n° 46, 2018.

Billets du carnet de recherche « Littérature et Révolution » d'Olivier Ritz

- Vers un catalogue numérique de la Révolution, 7 novembre 2018.
- « La faim passe du peuple au Roi ! », 19 juin 2018.
- Dans les petits papiers de Michelet, 21 juin 2017.
- « La Fabrique de la Révolution », 27 avril 2016.

Autres livrables (guides, recommandations, etc.) :

- Guide d'expressions régulières appliqué au projet LAFABREV, rédigé et intégré au site internet par Luc Massip (stage d'Humanités numériques) : https://lafabrev-michelet.lac.univ-paris-diderot.fr/regex
- Protocole de transcription intitulé « Tableau récapitulatif des éléments », rédigé par Cécile Milot-Pinson et actualisé par Cécile Andrisi-Brémon : bientôt téléchargeable à la page « Téléchargements » du site : https://lafabrev-michelet.lac.univ-paris-diderot.fr/telechargements

3) Présentation et description du corpus

☐ ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

La Fabrique de la Révolution (LAFABREV)

Présenter et décrivez le corpus

Ce projet d'humanités numériques se concentre sur la transcription d'un corpus manuscrit qui n'a pas encore été dépouillé : les sept volumes de notes préparatoires à l'*Histoire de la Révolution française*, conservés à la Bibliothèque historique de la Ville de Paris.

La transcription et l'indexation de quelque 1800 de ces feuillets (sur environ 2000) est susceptible de mieux faire connaître les sources documentaires sur lesquelles l'historien s'est appuyé, mais surtout sa méthode de travail et le(s) type(s) d'usage(s) qu'il fait de ces sources.

Période couverte par le corpus, auteur(s) concerné(s)

Auteur : Jules Michelet ;
Dates d'écriture : de septembre 1846 à septembre 1849

Organisation du corpus

Organisation en 7 volumes de notes (6 volumes de notes préparatoires à l'*Histoire de la Révolution française* + 1 volume spécial sur les sections parisiennes)

Mode de collecte et origine des données

Les notes préparatoires à l'*Histoire de la Révolution française* de Michelet n'avaient jamais été exploitées avant leur numérisation.

Les six volumes de notes conservés à la Bibliothèque historique de la Ville de Paris sous le nom « Histoire de la Révolution française », correspondent à un ensemble complexe de documents préparatoires, en partie réorganisés après coup en vue d'éventuelles

réutilisations. D'un point de vue matériel, leur conservation préventive au sein de la Bibliothèque par leur montage sur onglet en recueils factices, a durablement perturbé la lisibilité et l'architecture de ces fragments, organisés au départ sous forme de liasses et de chemises hiérarchisées selon un classement organique.

Il s'est donc agi en premier lieu de restituer le plus fidèlement que possible l'architecture d'origine de ces milliers de fragments et d'annotations dont le format et plus encore le statut sont fort variables.

Les liens et associations entre certains fragments dessinent en effet de véritables dossiers préparatoires, qui permettent de reconstituer les séquences narratives ou démonstratives de l'*HRF.*

Enfin, il existe un septième volume de notes, isolées de longue date, sans doute par Michelet lui-même après 1871. Le recueil factice ainsi constitué contient la seule trace d'une documentation, provenant pour l'essentiel des sections parisiennes et témoignant donc du pouvoir « sans-culotte ». En effet, à la suite de l'incendie de l'Hôtel de Ville en mai 1871, les archives de la municipalité parisienne durant la Révolution ont irrémédiablement disparu. La transcription et la mise en ligne intégrale de ce manuscrit permettent d'associer à la figure et au travail de l'écrivain celui de l'archiviste, dédoublant ainsi les formes de transmission et de fabrication des mémoires : à travers ses propres notes de travail, Michelet n'est pas seulement un historien qui écrit la Révolution française. Il est aussi le médiateur qui contribue à l'archiver.

État du corpus numérique

Le corpus numérique est composé des images des notes ou « papiers » manuscrits scannés, des fichiers XML des transcriptions et de sept index : index des personnes, des lieux, des États, des institutions, des références bibliographiques, des références littéraires et artistiques et des événements.

Tous les papiers (sauf rares exceptions) ont été transcrits, encodés, corrigés au moins une fois (pour la majeure partie des papiers), et sont en cours de nouvelle relecture à partir d'un tableau de repérage d'erreurs dans les formes normalisées (élaboré par Luc Massip et en partie complété par Olivier Ritz et Cécile Andrisi-Brémon).

Types de données:

Les métadonnées et les transcriptions sont regroupées à l'intérieur d'un même fichier XML : les métadonnées sont décrites en EAD et les transcriptions sont encodées selon un schéma spécifique au projet (DTD), qui s'inspire de celui du projet des « Manuscrits de Stendhal », copiloté par Thomas Lebarbé, et qui peut être au moins partiellement transposable en TEI.

Les images des notes (une image pour une note) affichées sur le site sont au format JPEG dans une résolution réduite et une archive contenant la totalité des images sera également bientôt téléchargeable en haute résolution (JPEG, 300 DPI), depuis la page « Téléchargements » du site :
https://lafabrev-michelet.lac.univ-paris-diderot.fr/telechargements

Chaque image présente sur le site correspond à un « papier ». Les volumes de la BHVP rassemblent des papiers de différentes longueurs : certains très longs et d'autres plus courts, voire très courts, collés ensemble sur de grandes pages. Les papiers courts ont donc été découpés dans un logiciel de traitement de l'image afin de figurer sur le site. Tous les rectos et les versos non vierges ont été numérisés. Certains versos ne semblant pas présenter d'intérêt scientifique ont été écartés de la transcription.

Toutes les images des papiers figurent sur le site « Chaîne éditoriale » du projet, de même que les fichiers XML des transcriptions, la DTD et la CSS : https://ce-michelet.app.univ-paris-diderot.fr/

Tous les scans des folios tels qu'initialement numérisés (rassemblant donc parfois plusieurs papiers) sont conservés au format TIFF (résolution 400 DPI, mode RVB) sur un serveur dédié, hébergé par l'Université de Paris.

Les index du site sont générés depuis un tableur collaboratif, exporté au format .CSV puis .JSON. Les index obtenus sont consultables sur le site par un système d'onglets accessibles à partir du pied de page et s'ouvrent à partir de l'index des noms de personnes : https://lafabrev-michelet.lac.univ-paris-diderot.fr/index-personnes

Grâce au travail de Thomas Lallier, l'utilisateur du site peut, par un système d'infobulles, naviguer des papiers vers les entrées d'index. Grâce au travail de Luc Massip, l'utilisateur peut en outre naviguer des entrées d'index vers les entités nommées correspondantes encodées à l'intérieur des transcriptions et des entités nommées encodées vers l'entrée d'index correspondante. Ce système se base sur des interrogations de la base de données réalisées à l'aide d'expressions régulières de type regex.

Volumétrie

1874 papiers (1874 images et autant de transcriptions au format XML) et 7 index consultables sur le site

Modifications effectuées sur les données, versions, …

Les modifications apportées dans les fichiers XML sont encodées à l'intérieur des métadonnées de la manière suivante :
<item_transcription date_transcription="AAAA/MM/JJ"
modifications="transcription_initiale" nom_transcripteur="Prénom + Nom"/>
Les modifications peuvent être de 3 types : "metadonnees", "transcription initiale" et "corrections".
Un bloc de validation permet en outre de préciser la date et l'auteur de la correction finale et la date et le nom de la validatrice (Paule Petitier).

Les modifications de la DTD sont inscrites en début de fichier : version 1, version 2, version 3.1 à 3.7

Les modifications du protocole de transcription (tableau récapitulatif des éléments) suivent les modifications de la DTD.

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

7 index

Métadonnées, créées et standards et formats utilisés

Métadonnées décrites en EAD principalement (sauf bloc de validation)

Les métadonnées descriptives, administratives et techniques

Balises de métadonnées utilisées :

```
<description_corpus>
        <titre_corpus>
                <titre_principal_corpus>
                <sous_titre_corpus>
        <proprietaire>
 <programme_recherche>
        <nom_programme>
        <financeur>
        <logiciel>
  <caracteristiques_document>
        <cote_fichier_numerique>
        <cote_volume>
        <cote_folio>
                <cote_folio_A>
                <cote_folio_C>
                <cote_folio_BIC>
        <description_materielle>
                <dimensions>
                        <largeur>
                        <hauteur>
                <aspect>
                <verso type_verso="" />
        <scripteur forme_normalisee="">
        <transcription>
        <item_transcription date_transcription="AAAA/MM/JJ" modifications="metadonnees"
nom_transcripteur="Maryelle Magret" />
        <item_transcription date_transcription="AAAA/MM/JJ"
modifications="transcription_initiale" nom_transcripteur="Prénom + Nom" />
        <item_transcription date_transcription="AAAA/MM/JJ" modifications="corrections"
nom_transcripteur="Prénom + Nom" />
          <validation>
        <fichier_corrige date_finalisation_correction="AAAA/MM/JJ"
nom_correcteur="Prénom + Nom" />
        <a_valider valeur="OUI" />
        <fichier_valide date_validation="AAAA/MM/JJ" nom_validateur="Prénom + Nom" />
        <a_publier valeur="OUI" />
        </validation>
```

Les métadonnées structurelles et l'annotation sémantique

Voir protocole de transcription bientôt téléchargeable à la page « Téléchargements » du site

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Indexation s'inspirant librement de la TEI et indexation complémentaire libre

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

☐ ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

☐ ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Répertoire "htdocs" en cours de téléchargement pour vérifier volume total du site LAFABREV

Accès, partage et limites des données

Partage des données selon la [licence Creative Commons Attribution - Pas d'utilisation commerciale - Partage dans les mêmes conditions 4.0 International]().

5) Responsabilités et ressources pour la gestion des données

☐ ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

☐ ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Le projet LAFABREV serait intéressé par un stockage de ses données par la TGIR HumaNum.

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

6) Archivage des données

☐ ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

☐ ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

Actuellement le site internet LAFABREV et la base de données qui lui est associée sont hébergés sur les serveurs de l'Université de Paris (ex-Paris Diderot).
La chaîne éditoriale et sa base de données ainsi que le serveur contenant les images scannées des folios des volumes de la BHVP sont également hébergés sur les serveurs de l'Université de Paris.

Durée de conservation des données

Aussi longtemps que possible

Volume des données à conserver

Calcul du volume de données à conserver en cours

Coûts alloués à la conservation

Coûts à déterminer avec HumaNum si cette solution est choisie.

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

Accès VPN au serveur (via client SFTP) et à la base de données (via phpMyAdmin)

7) Partage des données à l’issue du projet

☐ ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s’il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

☐ ***Recommandations :***

*Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR: les données doivent être trouvables (*findable*), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d’accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l’accompagnement d’une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.*

Potentiel de réutilisation des données

Fichiers XML transformables (au moins partiellement) en XML-TEI
Demande de DOI envisageable

Éléments d’accompagnement qui permettent la réutilisation des données.

Aide pour interpréter les données : protocole de transcription

Publications sur les données destinées à en améliorer l’exposition

Publications dans revues et carnet de recherche détaillées dans point “Publications” de la partie 2 de ce document

Conditions de réutilisation : licences et contrats pour l’ensemble du projet

Licence Creative Commons Attribution - Pas d’utilisation commerciale - Partage dans les mêmes conditions 4.0 International

Plan de gestion de données :

Archives savantes des Lumières. Correspondance, collections et papiers de travail d'un savant nîmois: Jean-François Séguier (1703-1784)

Table des matières

---

1) Plan de gestion de données (PGD) du projet d'édition de la correspondance de Jean-François Séguier

☐ ***Présentation de la section***

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

☐ ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

CHAPRON, Emmanuelle, IdHAL : emmanuelle-chapron ; ORCID : https://orcid.org/0000-0001-9907-7961, Aix-Marseille Université, CNRS - TELEMMe (UMR 7303), Marseille, France

Rôle dans le projet : Responsable du projet

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France

Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD projet Archives savantes des Lumières. Correspondance, collections et papiers de travail d'un savant nîmois: Jean-François Séguier (1703-1784)
1 version de ce PGD est actuellement prévue

2) Présentation du projet et responsabilités

☐ ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040),
Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

Écritures savantes au siècle des Lumières. La correspondance et les carnets de visiteurs de Jean-François Séguier

Responsable du projet (principal researcher) et unité de rattachement

CHAPRON, Emmanuelle, IdHAL : emmanuelle-chapron ; ORCID : https://orcid.org/0000-0001-9907-7961, Aix-Marseille Université, CNRS - TELEMMe (UMR 7303), Marseille, France

Rôle dans le projet : Responsable du projet

Financeur(s) du projet et type de financement

2011 : Fonds incitatif recherche, Aix Marseille université : 10 000 euros. Le financement a permis l'élaboration du site accueillant la base de données, la reproduction numérique de lettres de/à Séguier conservées hors de Nîmes, l'organisation d'une journée d'études.
2012-2017 : utilisation des crédits IUF d'Emmanuelle Chapron (à la hauteur de 25 000 euros environ)
2020 : financement CAHIER (3000 euros)

Référence de la convention de financement

Aucune convention de financement

Institution / organisme / unité porteuses du projet

Laboratoire TELEMMe (Temps, Espaces, Langages, Europe méridionale-Méditerranée, UMR 7303), CNRS - Aix-Marseille Université

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Le projet d'édition et d'étude des archives savantes de Séguier est une initiative conjointe du laboratoire Telemme et de l'Institut européen Séguier (association nîmoise fondée en 2005).

Descriptif et objectif(s) du projet

Voir : https://cahier.hypotheses.org/seguier
Centré sur le savant nîmois Jean-François Séguier (1703-1784), le projet « Écritures savantes au siècle des Lumières » est une contribution à l'histoire des mobilités et de la communication intellectuelle dans l'Europe du XVIIIe siècle. Cet antiquaire et botaniste a en effet laissé une importante correspondance (environ 3500 lettres, dont un tiers provenant de savants non régnicoles) et attiré dans son cabinet d'antiquités et d'histoire naturelle de nombreux voyageurs européens (près de 1500 pour la seule décennie 1770). Au-delà de l'éclairage que cette figure d'érudit méridional peut apporter sur le fonctionnement de la vie savante à l'écart des grandes capitales culturelles, une telle entreprise s'inscrit dans les réflexions actuelles sur ces deux modalités complémentaires de faire connaissance et travailler que sont, à l'époque moderne, la correspondance et le voyage.
Le projet consiste en l'édition numérique de la correspondance de Jean-François Séguier (1703-1784). Les relations épistolaires entretenues par l'antiquaire et botaniste nîmois avec de très nombreux savants européens, ainsi qu'avec tout un milieu d'amateurs méridionaux, ouvrent de nombreuses perspectives de recherche. Elles permettent d'interroger les formes du travail intellectuel à distance, la circulation des plantes, des graines et des livres, les cultures épistolaires (choix de la langue, civilités de l'entrée en correspondance), l'acculturation de différents milieux sociaux aux pratiques scientifiques (herborisation, collection, description des spécimens, manipulation des références bibliographiques…). Ces perspectives ont été illustrées par un colloque international **(Emmanuelle Chapron, François Pugnière (éd.)**, *Écriture épistolaire et production des savoirs au XVIIIe siècle. Les réseaux de Jean-François Séguier*, Paris, Classiques Garnier, 2019).

Pour mener à bien ce projet, un groupe de recherche, le Comité international Séguier (CIS), a été constitué et s'est réuni pour la première fois en mars 2010. Il était placé par convention sous la double tutelle de l'UMR 73030 Telemme (CNRS-Université de Provence, aujourd'hui Aix Marseille Université) et de l'Institut européen Séguier (association loi 1901, fondée en 2005), qui coordonnait depuis sa fondation les recherches autour du savant nîmois. Le CIS rassemblait des historiens, des historiens et philosophes des sciences, des littéraires de différents pays européens et était organisé en trois comités :

1) le **Comité scientifique**, sous la présidence de Brigitte Marin (Professeur d'histoire moderne, Université de Provence-MMSH), chargé du pilotage scientifique général du projet, est constitué de :

- Gabriel Audisio (Professeur émérite d'histoire moderne, Université de Provence / Institut européen Séguier)
- Jean Boutier (Directeur d'études à l'EHESS, Centre Norbert Elias, Marseille)
- Laurence Brockliss (Professeur d'histoire moderne, Université d'Oxford)
- Marina Caffiero (Professeur d'histoire moderne, Université La Sapienza, Rome)
- Michel Christol (Professeur d'histoire ancienne, Paris-I Sorbonne)
- Willem Frijhoff (Professeur d'histoire moderne, Université libre d'Amsterdam)
- Sergey Karp (Centre d'étude du XVIII^{e} siècle, Académie des sciences, Russie)
- Hans-Jürgen Lüsebrink (Professeur d'histoire moderne, Université de Sarrebrück)
- Daniel Roche (Professeur honoraire, Collège de France)

2) le **Comité de recherche**, sous la présidence d'Emmanuelle Chapron (MCF en histoire moderne, Université de Provence-Telemme), chargé des recherches sur le terrain et de l'incrémentation du site, est constitué de :

- Arnaud Bartolomei (MCF en histoire moderne, Université de Nice)
- Robert Chamboredon (Professeur agrégé d'histoire-géographie, Nîmes / Institut Séguier)
- Samuel Cordier (docteur du Muséum national d'histoire naturelle, en poste à Genève)
- Ivano Dal Prete (Visiting Scholar, Lecturer, Yale University)
- Claire Davison-Pégon (Professeur de littérature, Université de Provence)
- Véronique Krings (MCF en histoire romaine, Université Toulouse-II)
- Gilles Montègre (MCF en histoire moderne, Université Grenoble-II)
- François Pugnière (Professeur d'histoire-géographie, Nîmes / IES)
- David Rousseau (doctorant en histoire moderne sous la direction de Pierre-Yves Beaurepaire, Université de Nice)
- Anne Saada (CNRS-UMR 8547 Pays germaniques : histoire culture philosophie).

3) le **Comité d'organisation**, sous la présidence de Gabriel Audisio (Institut européen Séguier, Professeur d'histoire émérite de l'Université de Provence) coordonne les activités des groupes et assure le relais avec l'IES. Il comprend :
Evelyne Bret (Conservateur chargée des fonds anciens de la Bibliothèque municipale de Nîmes)
Hélène Deronne (MCF en histoire de l'art, Université d'Avignon)
Jean-Marie Guillon (Professeur d'histoire, Université de Provence- UMR Telemme)
Jean-Louis Meunier (président de l'Institut européen Séguier)
Jean-Michel Ott (trésorier du CIS, Institut européen Séguier)
Rüdiger Stephan (chargé des relations internationales à l'Institut européen Séguier)
Julie Thérond (chargée de communication, Institut européen Séguier)

Ce comité international Séguier n'a jamais réellement fonctionné. Il ne s'est jamais réuni en assemblée plénière après cette première rencontre et n'a été que rarement sollicité.
Après le délitement puis l'arrêt des activités de l'Institut européen Séguier (association loi 1905), l'entreprise s'est poursuivie sous la direction conjointe d'Emmanuelle Chapron (maître de conférences puis professeur d'histoire moderne, Aix Marseille Université), d'Eric

Carroll (ingénieur de recherche en informatique, CNRS, Telemme) et de François Pugnière (professeur d'histoire-géographie à Nîmes, membre associé de l'EA 4424 CRISES, Montpellier III), avec le seul soutien de l'UMR Telemme.

Dates et durée

Date de début de financement et de début des travaux : 2010
Date de fin de financement et de fin des travaux : 2025 (fin des travaux estimée)

Mots clés du projet

- Correspondance savante
- Édition numérique (Édition électronique)
- Collections
- Archives

Publications (articles, pré-proposition, site web, …)

Site web du projet : www.seguier.org (le lien n'est plus actif)

En plus du site, pour une valorisation plus rapide du projet, un blog a été créé sur la plateforme Hypothèses en septembre 2015. Il permet de mettre en évidence les dernières publications sur le site, des points d'avancement, des réflexions théoriques. Administré par Emmanuelle Chapron, il compte une soixantaine de billets.
Carnet Hypothèse dédié au projet Séguier : https://seguier.hypotheses.org/

Listes des publications liées au projet :

- 2011 : journée d'études « Conserver, archiver, éditer. Usages de la correspondance savante, xvii^{e}-xviii^{e} siècles ». Actes publiés dans la Bibliothèque de l'École des chartes, 2013 [2017] sous la direction d'Emmanuelle Chapron et Jean Boutier.
- 2016 : colloque international Savoirs à l'œuvre, savants au travail. Actes publiés dans le volume **Emmanuelle Chapron et François Pugnière (dir.)**, *Écriture épistolaire et production des savoirs au xviii^{e} siècle. Les réseaux de Jean-François Séguier*, Paris, Classiques Garnier, 2019, 315 p.

Communications individuelles, Emmanuelle Chapron :

- Journée d'étude HISTARA, Paris, 2021 ;
- Colloque Condorcet, ENS, Paris, 2018 ;
- Séminaire d'Histoire des relations internationales, Poitiers, 2018 ;
- Assemblée du consortium CAHIER, Paris, 2016.

Publications individuelles :

- **Emmanuelle Chapron**, « Monde savant et ventes de bibliothèques en France méridionale dans la seconde moitié du xviii^{e} siècle », *Annales du Midi*, 283, 2013, p. 409-429 [halshs-01487318]. 2013
- **Emmanuelle Chapron**. *L'Europe à Nîmes : les carnets de Jean-François Séguier (1732-1784)*, Editions Barthelemy, 2008

3) Présentation et description du corpus

☐ ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

Écritures savantes au siècle des Lumières. La correspondance et les carnets de visiteurs de Jean-François Séguier

Présenter et décrivez le corpus

Le cœur du corpus visé par le projet est la correspondance de Séguier, qu'on se propose d'éditer intégralement. Sa partie passive (correspondance reçue) est conservée pour l'essentiel à la Bibliothèque Carré d'Art de Nîmes et à la Bibliothèque nationale de France. Les lettres écrites par le Nîmois sont dispersées dans de nombreuses bibliothèques et fonds d'archives européens, notamment en Allemagne, en Suisse, au Royaume-Uni et en Italie. Leur recensement a supposé de longues recherches dans les catalogues et inventaires en ligne, et parfois des déplacements in situ. L'état du recensement des lettres et de leur transcription est tenu à jour sur le blog du projet : https://seguier.hypotheses.org/category/etats-des-lieux
Dans un second temps, il est prévu d'associer au projet la transcription du reste des archives de Séguier conservées à la Bibliothèque Carré d'Art de Nîmes (carnet de voyage, notes inédites, mémoires, catalogue de bibliothèque, inventaire des collections).
Environ 3710 lettres ont été repérées à ce jour :
La correspondance active représente 822 lettres et la correspondance passive, 2888 lettres. Parmi cet ensemble, 2700 lettres sont conservées à la Bibliothèque du Carré d'art de Nîmes, le reste est dispersé dans plus d'une cinquantaine d'établissements situés dans neuf pays (Allemagne, Autriche, France, Italie, Pays-Bas, Royaume-Uni, Suède, Suisse, Russie). Plus de 2200 lettres sont accessibles sur www.seguier.org.
Dans un second temps, le site accueillera l'édition des papiers de travail du savant : carnets de travail et de voyage, répertoires des visiteurs, notes de lecture, croquis, inventaires et catalogues autographes des collections.

Période couverte par le corpus, auteur(s) concerné(s)

Auteur : Séguier, Jean-François (1703-1784)

ISNI : 0000000108147857
ARK BnF : http://catalogue.bnf.fr/ark:/12148/cb10576633j
IdRef : 07713320X

Période du corpus : 1703 - 1784

Organisation du corpus

Les informations sur les données sont rassemblées et organisées dans un tableur Excel unique qui servira à leur dépôt dans Nakala.
Les colonnes rassemblent les métadonnées de la lettre : titre; expéditeur; date de la lettre; lieu et pays d'expédition; récepteur; lieu et pays de réception; langue; établissement de conservation de la lettre; cote; transcription intégrale; annotations; dessins et croquis; autres mentions descriptives; thèmes; personnes citées; ouvrages cités; contributeur [chercheur qui a transcrit la lettre]; éditeur scientifique; références bibliographiques; lien vers Gallica [si manuscrit numérisés]; droits sur l'image; licence.
Les lignes correspondent aux lettres. Chaque lettre est identifiée par un ID numérique progressif (de ID1 à ID 2263) qui permet de faire le lien avec les images de la lettre.
Les images sont nommées par cet ID et un numéro progressif (par exemple : ID263_1, ID263_2, etc.). Si on ne dispose pas des droits sur l'image, la lettre est liée à un fichier image générique "Image non disponible".

Mode de collecte et origine des données

Lettres détenues en majorité par la bibliothèque du Carré d'art de Nîmes. Elles ont été numérisées et transcrites afin d'être accessibles à la fois en mode image et en mode texte.

Etat du corpus numérique (types et natures des données, modifications effectuées sur les données, volumétrie)

Le corpus des lettres est estimé à 3500 lettres. Actuellement, 2200 ont été transcrites et environ 500 sont en cours de transcription ou de versement dans la base de données. Des recherches *in situ* devraient être entreprises pour retrouver et transcrire le reste des lettres, notamment en Russie et en Italie.

La convention liant l'Institut européen Séguier et l'UMR 7303 Telemme prévoyait la mise à disposition d'un ingénieur de recherche en informatique, Eric Carroll, pour la mise en place d'un site internet qui permette d'interroger et de consulter les données. Ce site a été conçu en deux temps :

1) Dans un premier temps (2010-2013), une interface de travail a été construite pour permettre une saisie collaborative des données (métadonnées des lettres, transcription intégrale et indexation des textes, chargement des images). Elle était fondée sur une base de données relationnelle, comprenant une vingtaine de tables, animée par une instance de SqlServer 2008 R2. La deuxième couche consistait en un développement d'une application internet riche (Ajax, JavaScript, dot Net 4.0, C#, XHTML/CSS, DublinCore/OAI, UML).

2) Dans un second temps (2013), une interface publique a été mise en place par une entreprise privée (Walter Wizman), en raison des multiples obligations professionnelles d'Eric Carroll au sein du laboratoire. A partir de l'adresse www.seguier.org, il était désormais possible d'interroger la lettre à partir de requêtes par nom (auteur de la lettre, destinataire,

individus cités dans la lettre), par date et lieu d'expédition, par institution de conservation, par thème. Il était aussi possible d'effectuer des requêtes en texte intégral ou de sélectionner les lettres accompagnées de croquis ou auxquelles sont jointes des objets, livres, graines et plantes ou petites antiquités. La liste des résultats permettait d'accéder aux éléments d'identification de la lettre et à son texte intégral, éventuellement à son image numérique, selon les conventions passées avec les établissements.

Après le départ d'Eric Carroll, qui a quitté l'UMR Telemme en février 2020, le site www.seguier.org n'a plus été entretenu, le nom de domaine n'a plus été payé et le site n'est plus accessible.

Les données textuelles ont été récupérées sous la forme d'un tableur Excell d'environ 2200 lignes et 30 colonnes. Avec les images des lettres (5000 fichiers image), l'ensemble a été stocké sur les serveurs de la MMSH et sur le Sharedocs du consortium CAHIER. Le consortium a mis à disposition du projet deux stagiaires (Andrés Echevarria Pelaes et Ala Eddine) chargés d'accompagner le dépôt des données sur Nakala.

Métadonnées, créées et standards et formats utilisés

Les métadonnées descriptives, administratives et techniques

Métadonnées de la lettre :

- descriptives : titre; expéditeur; date de la lettre; lieu et pays d'expédition; récepteur; lieu et pays de réception; langue; établissement de conservation de la lettre; cote; transcription intégrale; annotations; dessins et croquis; autres mentions descriptives; thèmes; personnes citées; ouvrages cités; références bibliographiques
- administratives: contributeur [chercheur qui a transcrit la lettre]; éditeur scientifique
- techniques: lien vers Gallica [si manuscrit numérisés]; droits sur l'image; licence.

Les métadonnées structurelles et l'annotation sémantique

Pas de données d'enrichissement sémantique ou de métadonnées structurelles

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Indexation à partir d'une liste propre au projet.

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

☐ ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

☐ ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Accès, partage et limites des données

Les données sont stockées actuellement sur les serveurs de la MMSH et sur le Sharedocs du consortium CAHIER. Elles sont en cours de dépôt sur Nakala. Un site est en cours de construction sur Nakala-web, l'accès sera libre.

La publication des images a fait l'objet de conventions particulières avec les établissements de conservation, qui précisent les mentions légales à apporter. Les conventions devraient être renouvelées avant la publication des données sur Nakala. Les images des lettres conservées par la BnF (entre autres) ne sont pas couvertes par ces conventions.
Les transcriptions sont sous licence CC_BY 4.0

5) Responsabilités et ressources pour la gestion des données

☐ ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

☐ ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsable de la gestion des données :
CHAPRON, Emmanuelle, IdHAL : emmanuelle-chapron ; ORCID : https://orcid.org/0000-0001-9907-7961, Aix-Marseille Université, CNRS - TELEMMe (UMR 7303), Marseille, France

Rôle dans le projet : Responsable du projet

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

Dans le cadre du projet Séguier, le consortium a accordé un financement à hauteur de 3 000€. Un espace de stockage et de partage de fichiers sharedocs a également été ouvert (service HumaNum).

Mise à disposition de personnel :

- La convention liant l'Institut européen Séguier et l'UMR 7303 Telemme prévoyait la mise à disposition d'**un ingénieur de recherche en informatique, Eric Carroll**, pour la mise en place d'un site internet.
- Le consortium CAHIER a mis à disposition du projet deux stagiaires (Andrés Echevarria Pelaes et Ala Eddine) chargés d'accompagner le dépôt des données sur Nakala.

Le projet a rassemblé au fil des ans de **nombreux contributeurs qui ont alimenté la base de données en corpus de lettres liés à leurs projets de master, de thèse ou de recherche** :

- Lily Servel, étudiante du master Histoire d'Aix Marseille Université, a participé au projet pendant son master et par un CDD de 2 mois en 2016
- Adeline Danerol, étudiante du master Histoire d'Aix Marseille Université, a participé au projet pendant son master et a effectué plusieurs missions ponctuelles pour le projet
- Etienne Stockland, doctorant
- Meike Knittel, doctorante
- Florence Catherine, enseignante du secondaire
- Gilles Montègre, MCF Université Grenobles Alpes
- Andrea Bruschi, docteur en histoire, rémunéré (auto-entrepreneur) pour la saisie de lettres italiennes en 2014 et 2021
- Claire Torreilles
- Véronique Chapron, retraitée bénévole

6) Archivage des données

☐ ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

☐ ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

En cours de dépôt sur Huma-Num, la question de l'archivage n'est pas encore évoquée.

Durée de conservation des données

Inconnue

Volume des données à conserver

Inconnu

Coûts alloués à la conservation

Inconnu

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

7) Partage des données à l'issue du projet

☐ ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

☐ ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR: les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

Eléments d'accompagnement qui permettent la réutilisation des données.

Publications sur les données destinées à en améliorer l'exposition

La valorisation et la documentation sur les données sont rassemblées sur le carnet Hypothèses : inventaires régulièrement mis à jour des lettres repérées et transcrites, éclairages sur certaines lettres ou ensemble de lettres, cartes, mise en relation avec d'autres archives, annonce des publications, etc. Le blog continuera à fonctionner après le dépôt sur Nakala mais nous espérons pouvoir rapatrier et diffuser une partie des informations sur le nouveau site.

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

Voir -4-

Plan de Gestion de Données : Les dossiers de *Bouvard et Pécuchet*

Table des matières

---

1) Plan de gestion de données (PGD) du projet Les dossiers de Bouvard et Pécuchet

☐ ***Présentation de la section***

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

☐ ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

Dord-Crouslé, Stéphanie, IdHAL : stephanie-dord-crousle ; ORCID : 0000-0002-6683-9509, CNRS (UMR 5317 IHRIM), France
Rôle dans le projet : responsable scientifique

L'HERMITE, Laurène, IdRef : https://www.idref.fr/236176927 ; Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le projet : co-auteure du PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021
1 version de ce PGD est actuellement prévue

2) Présentation du projet et responsabilités

☐ ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040),
Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

Les dossiers de *Bouvard et Pécuchet*

Responsable du projet (principal researcher) et unité de rattachement

Dord-Crouslé, Stéphanie, IdHAL : stephanie-dord-crousle ; ORCID : 0000-0002-6683-9509, CNRS (UMR 5317 IHRIM), France

Rôle dans le projet : responsable scientifique

Financeur(s) du projet et type de financement

Le projet a bénéficié d'un soutien financier spécifique

- du CNRS (appel d'offres « ATIP Jeunes chercheurs » 2006 du Département Sciences humaines et sociales)
- de l'ANR (appel à projets « Corpus et outils de la recherche en Sciences humaines et sociales » du programme Sciences humaines et sociales 2007 ; https://anr.fr/Projet-ANR-07-CORP-0009).
- de la Région Rhône-Alpes (allocation doctorale allouée au projet dans le cadre du Cluster de recherche n° 13 « Culture, patrimoine, création » 2007-2010)
- du Ministère des Affaires étrangères et européennes (Partenariat Hubert Curien Galilée 2009)

- de la TGIR Huma-Num par l'intermédiaire du consortium CAHIER (2012, 2013 et 2018).

Voir http://www.dossiers-flaubert.fr/projet-partenaires-soutiens

Institution / organisme / unité porteuses du projet

l'UMR 5317 IHRIM qui a pris la suite de l'UMR 5611 LIRE le 1er janvier 2016.

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Le projet – dans sa version initiale – a été réalisé entre 2006 et 2012

- dans le cadre de l'UMR 5611 LIRE (« Littérature, Idéologies, REprésentations, XVIIIe et XIXe siècles »), unité mixte de recherche qui associait le CNRS, l'Université Lumière - Lyon 2, l'Université Jean Monnet de Saint-Étienne, l'Université Stendhal - Grenoble 3 et l'ENS de Lyon
- avec le soutien technique de l'USR 3385 ISH (unité mixte de services de l'Institut des Sciences de l'Homme)
- de l'ENS-LSH (École normale supérieure - Lettres et sciences humaines) devenue l'ENS de Lyon (École normale supérieure de Lyon)
- et du TGE Adonis.

Il se poursuit aujourd'hui :

- dans le cadre de l'UMR 5317 IHRIM qui a pris la suite de l'UMR 5611 LIRE le 1er janvier 2016
- avec le soutien technique de la TGIR Huma-Num
- au sein du consortium CAHIER.

Voir http://www.dossiers-flaubert.fr/projet-partenaires-soutiens

Descriptif et objectif(s) du projet

Conservés à la bibliothèque municipale de Rouen, les dossiers de *Bouvard et Pécuchet*, le dernier roman – posthume et inachevé – de Gustave Flaubert (1821-1880), constituent un ensemble patrimonial imposant (2 400 feuillets), cohérent, d'importance scientifique et culturelle reconnue. Ils sont porteurs d'une dimension épistémologique singulière : composés pour rédiger une « encyclopédie critique en farce », ils proposent une configuration critique des savoirs au XIXe siècle, originale et révélatrice. Ils forment le socle de la présente édition. Mais d'autres dossiers existent ailleurs qui ont vocation à enrichir le site en rejoignant progressivement et virtuellement leurs semblables. Car c'est l'ensemble de ce chantier documentaire qui a servi à rédiger le premier volume de l'œuvre et aurait dû être réutilisé pour la composition d'un second volume, jamais écrit en raison de la mort soudaine du romancier.

Or, en raison de leur volume, de leur organisation complexe et indéfiniment mouvante, ainsi que de leurs contenus scientifiques extrêmement variés, les dossiers ne peuvent pas être

édités de manière satisfaisante sous une forme imprimée. C'est particulièrement vrai pour les pages préparées en vue du second volume du roman : les annotations que l'écrivain y a portées, indiquant le lieu probable du classement, sont souvent plurielles et obligent à conserver aux fragments textuels une mobilité qui est nécessairement défaite par la fixité d'une édition imprimée.

Dépassant cette limite en recourant au support électronique et à l'encodage XML-TEI intégral du corpus, la présente édition offre l'accès :

- aux images, à la transcription (formats diplomatique et textuel) et aux métadonnées des pages du corpus,
- à un moteur de recherche plein texte,
- à trois bibliothèques permettant d'identifier les références utilisées par Flaubert et de circuler dans le corpus
- et à un outil de production de « seconds volumes » possibles : l'agenceur.

Date et durée

Date de début des travaux : 2006
Date de fin des travaux : non prévue (tant qu'il y aura des dossiers à intégrer)

Mots clés du projet

- Roman -- Dossiers documentaires ;
- Text Encoding Initiative (langage de balisage) ;
- Bibliothèques numériques ;
- Flaubert, Gustave (1821-1880) Bouvard et Pécuchet ;
- Oeuvre inachevée ;
- Reconstitution conjecturale ;
- Edition posthume ;
- Manuscrits inédits ;

Publications (articles, pré-proposition, site web, …)

Site web du projet : http://www.dossiers-flaubert.fr
Le site Les dossiers de Bouvard et Pécuchet s'est vu attribuer un ISSN (International Standard Serial Number) par la Bibliothèque nationale de France : ISSN 2495-9979.
Sont ainsi soulignés et valorisés les enrichissements progressifs du site qui le constituent en ressource intégratrice.
Ressortent en particulier d'une publication en série :

- les reconstitutions conjecturales du « second volume » existant déjà sous forme papier ainsi que les agencements créés par des internautes (après validation par le comité scientifique du projet) qui seront progressivement mis en ligne sur le site
- et l'ajout à venir, sur la plateforme éditoriale, de nouveaux dossiers de notes non conservés à la bibliothèque municipale de Rouen.

Carnet de recherche du projet : https://flaubert.hypotheses.org/

Listes des articles publiés par le projet :
https://halshs.archives-ouvertes.fr/ANR-07-CORP-009

Autres livrables (guides, recommandations, etc.) :
Compte rendu de fin de projet ANR :

- **S. Dord-Crouslé**. Compte-rendu de fin de projet -Projet ANR-07-CORP-009 BOUVARD - *Les Dossiers de Bouvard et Pécuchet de Flaubert. Enrichissement, valorisation, documentation d'un corpus multi supports : Programme " Corpus et outils de la Recherche en Sciences Humaines et Sociales "* 2007. [Rapport de recherche] ANR (Agence Nationale de la Recherche - France). 2012. ⟨halshs-00760914⟩

3) Présentation et description du corpus

☐ ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

☐ ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Présentez et décrivez le corpus

Voir http://www.dossiers-flaubert.fr/edition-corpus
Au corpus originel intégralement conservé à la bibliothèque municipale de Rouen s'ajoutent maintenant, peu à peu, les dossiers conservés ailleurs. En dépassant les strictes limites qu'il s'était d'abord fixées (l'édition d'un ensemble patrimonial cohérent conservé à la bibliothèque municipale de Rouen), le site commence à réaliser pleinement le projet scientifique d'ampleur qui est le sien : donner à lire l'ensemble de la documentation réunie par Flaubert pour son entreprise encyclopédique « en farce » et permettre à l'agenceur d'y puiser des matériaux – pour certains inconnus – en vue de la création ou de l'enrichissement de « seconds volumes » possibles.
Ont été ajoutés les dossiers Rousseau, Hegel et Mirabeau.

Période couverte par le corpus, auteur(s) concerné(s)

Auteur : Flaubert, Gustave (1821-1880)
ISNI : 0000000122762442
ARK BnF : http://catalogue.bnf.fr/ark:/12148/cb11902894q
IDRef : 026866552
Période du corpus : Les documents produits par Flaubert l'ont été entre 1860 et 1880.
Mais les ouvrages pris en note ont un empan chronologique bien plus large: antiquité - 1880.

Organisation du corpus

Unités constitutives du corpus
(voir http://www.dossiers-flaubert.fr/edition-unites-constitutives) :

- **La page** (unité matérielle) : L'unité constitutive matérielle du corpus des dossiers documentaires de Bouvard et Pécuchet est *la page*. Un feuillet est formé de deux pages ou folios, un recto et un verso – dont l'un peut être vierge.
- **Les transcriptions** (associées aux pages). Elles sont de 4 types :

    - La *transcription ultra-diplomatique* se présente sous la forme d'un fichier PDF généré à partir d'un logiciel de traitement de texte. Elle reprend toutes les particularités de la graphie du scripteur.
    - La *transcription diplomatique* (format HTML et générée à partir des fichiers XML/TEI) conserve tous les traitements textuels décrits pour la version ultra-diplomatique. En revanche, elle homogénéise et rationalise une partie des dimensions topographiques et graphiques.
    - La *transcription normalisée* (format HTML et générée à partir des fichiers XML/TEI) achève d'homogénéiser le rendu topographique des pages en déterminant et en ne conservant que quelques espaces signifiants (essentiellement deux : la marge et le corps du texte). Mais surtout, elle propose un texte intelligible par tous les lecteurs, débarrassé des particularités et des graphies déviantes propres à chaque scripteur.
    - La *transcription enrichie* (format HTML et générée à partir des fichiers XML/TEI) permet de faire le lien entre les versions diplomatique et normalisée.
- **Les textes** (unités logiques à fondement matériel) : les pages regroupées selon un ordre validé scientifiquement forment des *textes* qui appartiennent à des catégories typologiques homogènes. Il s'agit d'un autre point d'entrée vers la lecture et l'exploitation des Dossiers. Techniquement, chaque texte, que ce soit en version diplomatique ou en version normalisée, présente l'agrégation – au sein d'une page HTML – du contenu balisé en XML/TEI de l'ensemble des pages concernées ; il est doté d'une URL spécifique et est accessible sur le site à partir d'une page de sommaire permettant de lister, type par type, la totalité des textes du corpus selon différents ordres (classement patrimonial, ordre alphabétique des titres, etc.)
- **Les fragments** : les pages sont composées de *fragments textuels*. Ce sont les unités logiques fondamentales de l'édition électronique du corpus : à leur niveau va être vérifiée et promue la mobilité des éléments constitutifs des dossiers documentaires de Bouvard et Pécuchet. La possibilité de créer des reconstitutions conjecturales du second volume du roman repose sur le découpage de l'intégralité du corpus en fragments textuels, opération qui le rend manipulable et infiniment réagençable. Chaque fragment textuel est accessible par l'intermédiaire d'une métadonnée (« Référence bibliographique de fragment ») attachée à la page où il apparaît, et élucidant la référence bibliographique exacte du fragment copié par Flaubert ou l'un de ses collaborateurs.
- **La citation** est le regroupement de tous les fragments présentant la réalisation textuelle de la même référence bibliographique. Chaque citation possède une page dédiée, pourvue d'une URL et présentant toutes les informations nécessaires à son identification.

Plan de nommage des fichiers :

Collection (nom)	**Collection (description)**	**Volume (nom)**	**Volume (description)**	
BnF	NAF	28825	"Littérature - esthétique"	
Montmorency	Musée JJ	495	"Notes sur	Montmorency_4

	Rousseau		rousseau”	95_f_001_r.jpg
Rouen	BM	g 225-3	Feuillets épars	
[Rouen]		Volume 1 cote g 226-1 => à corriger en :	g 226-1	
	Information requise uniquement dans le TEIHeader			
Antibes	Vente Caroline Franklin Groult, 1931	066	“Esthétique de Hegel”	Antibes_066_f_001_r

Mode de collecte et origine des données

Origine des images, manuscrits et autres pièces des dossiers :

- Bibliothèque municipale de Rouen

La numérisation du microfilm de sauvegarde des documents visés par le projet (microfilm acquis au prix public) nous a permis de constituer une base de 3500 images en noir et blanc de qualité médiocre .
Parallèlement, achat de près de 300 images HD couleur (sur 3500) au prix public grâce à une partie du financement reçu de l’ANR.
Manuscrit “définitif” Premier volume, notes, brouillons, plans, scénarios, notes de lecture. Pages préparatoires Second volume
Puis mise à disposition à titre gracieux par la bibliothèque de la numérisation des deux dossiers concernant le *Dictionnaire des idées reçues* (soit 130 images).

- Musée Jean-Jacques Rousseau et Bibliothèque d’études rousseauistes, Montmorency : mise à disposition des images à titre gracieux par convention
- Antibes, vente Caroline Franklin Groult, 1931: images uniquement, issues de collections privées, non référencées.

Etat du corpus numérique, types de données et volumétrie

Le corpus est ouvert et en cours d’enrichissement. De nouveaux dossiers sont régulièrement ajoutés. L’encodage est toujours en cours et en phase d’amélioration.

Il contient :

- Base de données SQL : Données de travail du projet, références bibliographiques (actuellement plus de 20000), etc. Voir par exemple http://www.dossiers-flaubert.fr/index.php?node=bibliotheques – 100 Mo
- Base de données XML : Transcriptions TEI – 3500 transcriptions à terme ; 2000 disponibles actuellement
- Images : Fac-similés de manuscrits – 3500 images (toutes disponibles en ligne), 5 Go
- Images : Fragments d’images découpés (partiellement disponibles – pour 300 images) – 21000 images (prévision), 3Go

Modifications effectuées sur les données, versions, …

Transcription issue d'un traitement de texte puis balisage en TEI.

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Trois bibliothèques de références bibliographiques destinées à enrichir et lier les données.
Voir : http://www.dossiers-flaubert.fr/index.php?node=bibliotheques

L'agenceur :
Il s'agit d'un outil informatique de production de "seconds volumes" possibles. Pour utiliser cet outil, il faut préalablement s'identifier (se connecter ou, lors de sa première visite, créer un compte).
Cette démarche permet à chaque demandeur de disposer d'un espace de travail personnel et privé.
Des informations utiles à la prise en main de l'agenceur sont disponibles :

- sur la page "Agencements" de présentation de l'édition
- en cliquant sur le point d'interrogation ("?") qui se trouve en haut et à droite sur certaines pages de l'espace de travail
- ou bien en consultant les pages dédiées du carnet de recherche du projet qui comportent plusieurs tutoriels (par exemple, ici).

Vous pouvez aussi regarder la vidéo expliquant le fonctionnement du site et plus particulièrement celui de l'agenceur.

Métadonnées, créées et standards et formats utilisés

Les métadonnées sont entièrement accessibles sur le site des "Dossiers…".
Exemple : http://www.dossiers-flaubert.fr/cote-Antibes_066_f_002__r-meta
Les métadonnées ne sont pas standardisées et les champs d'indexation sont libres.
Des transcriptions XML/TEI et des descriptions sont toujours en cours de réalisation.

Les métadonnées descriptives, administratives et techniques

Cote, Scripteur du manuscrit, ensemble textuel d'où provient l'extrait, nom du transcripteur.

Les métadonnées structurelles et l'annotation sémantique

Chronologie du document, provenance, classement typologique* et caractéristiques matérielles.
**Les métadonnées de classement* : pour chaque page sont proposés un classement typologique (en fonction des différents types de pages qui existent dans le corpus : notes de lecture, pages préparées pour le second volume, documentation brute imprimée, etc.) ; un classement chronologique (selon la datation plus ou moins précise qui peut être affectée à chaque page en fonction d'informations internes, comme les filiations génétiques, ou externes, la date d'emprunt d'un ouvrage consignée dans le registre d'une bibliothèque ou la mention, dans une lettre, de la période à laquelle une lecture a été faite par le romancier) ; et un classement par scripteur (Flaubert est évidemment le plus largement représenté, mais bien d'autres personnes lui ont apporté leur aide et ont laissé des traces manuscrites dans les dossiers de Rouen, au premier rang desquelles son ami Edmond Laporte, mais aussi

son « disciple » Guy de Maupassant). Ces classements permettent de proposer trois points d'accès au corpus qui s'ajoutent à celui que fournit, par défaut, le classement patrimonial, accessible par les sections du descriptif établi par l'institution de conservation ou par cotes.

Annotations ou métadonnées d'enrichissement

Annotations critiques.
Transcriptions TEI destinées à enrichir les manuscrits : transcription ultra diplomatique, diplomatique, transcription normalisée, transcription enrichie.
Références bibliographiques

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

☐ ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

☐ ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc. On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Stockage

Actuellement les données sont stockées via les services dédiés d'Huma-Num (Huma-Num Box ?)
Le transfert des données sur l'entrepôt Nakala (service Huma-Num) est prévu sous peu.

Volume des données stockées (qui sera également celui des données à sauvegarder) :

- PHP/JavaScript/CSS...: 892 Mo
- MySQL: 125 Mo
- SOLR: 38 Mo
- XML/TEI: 83 Mo
- Images (des manuscrits sous différents formats): 4,9 Go

Accès, partage et limites (d'accessibilité) des données

Une collection de 900 pages est moissonnée par Isidore

Concernant certaines images et manuscrits issus de collections de bibliothèques ou de fonds privés, il est à prévoir des règles de partage et de réutilisation :
Pour les cotes Rouen g226, g227 et g228 :

- Images consistant en des reproductions de microfilms : « Collections Bibliothèque municipale de Rouen ».
- Images consistant en des reproductions des manuscrits du Dictionnaire des idées reçues : « Collections Bibliothèque municipale de Rouen - photographie société Arkhénûm ».

- Autres images consistant en des reproductions des manuscrits de Bouvard et Pécuchet : « Collections Bibliothèque municipale de Rouen – photographie Thierry Ascencio-Parvy ».

Toute utilisation publique ou commerciale des images doit faire l'objet d'une autorisation préalable. Les demandes sont à adresser à la bibliothèque municipale de Rouen :
par courrier : Bibliothèque de Rouen, 3 rue Jacques Villon, F-76043 ROUEN CEDEX
ou par courriel : bibliotheque@rouen.fr

Pour la cote Montmorency : « Collection musée Jean-Jacques Rousseau - Ville de Montmorency - photographe Laure Querouil »
Toute utilisation publique ou commerciale des images doit faire l'objet d'une autorisation préalable. Les demandes sont à adresser au Musée Jean-Jacques Rousseau et Bibliothèque d'études rousseauistes par courrier :
Musée Jean-Jacques Rousseau et Bibliothèque d'études rousseauistes
4 rue du Mont-Louis
95160 Montmorency
ou par courriel : Rousseau-museum@ville-montmorency.fr

Pour la cote Antibes : « Collections privées »

Concernant l'utilisation des transcriptions :
L'utilisation des transcriptions à des fins privées, à des fins d'enseignement ou de recherche scientifique est autorisée, sous réserve de mentionner ainsi leur origine :

- « Transcription(s) réalisée(s) par [nom du transcripteur] pour l'édition des *Dossiers documentaires de Bouvard et Pécuchet*, sous la dir. de S. Dord-Crouslé, 2012-..., http://www.dossiers-flaubert.fr, ISSN 2495-9979. »

Pour toute publication, demander préalablement l'autorisation à la responsable de l'édition : Stéphanie Dord-Crouslé.
Le corpus complet est à citer comme suit :

- **Gustave Flaubert,** *Les dossiers documentaires de Bouvard et Pécuchet*. Édition intégrale balisée en XML-TEI accompagnée d'un outil de production de « seconds volumes » possibles, sous la dir. de Stéphanie Dord-Crouslé, 2012-..., http://www.dossiers-flaubert.fr, ISSN 2495-9979.

5) Responsabilités et ressources pour la gestion des données

☐ ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

☐ ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Responsable de la gestion des données :
Dord-Crouslé, Stéphanie, IdHAL : stephanie-dord-crousle ; ORCID : 0000-0002-6683-9509, CNRS (UMR 5317 IHRIM), France

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

Equipe engagée dans la gestion des données à différentes étapes du projet (voir http://www.dossiers-flaubert.fr/projet-equipe-technique) :
Développements informatiques

- [2016-...] Pierre-Yves Jallud (CNRS-IHRIM)
- [2014-2015] Jean-Eudes Trouslard (jet@zoulous.com) pour le module "seconds volumes à la demande" :
  Parties plans et agencements de l'espace de travail.
  - module d'extraction des données de la base TEI vers la base mySql
  - conception et développement de l'affichage des agencements et plans d'une reconstitution conjecturale
  - outils de modifications de l'arborescence des agencements et plans
  - génération en pdf du texte de la reconstitution conjecturale
- [2011-2012] Hugo Schuler (CDD)
- [2009] Stéphane Wustner (Stagiaire)
- [2008-2009] Jérémie Lagravière (CDD)

- [2007-2008] Martial Tola (CNRS-ISH)
- [2007-2012, responsable] Raphaël Tournoy (CNRS-ISH)

TEI et ingénierie documentaire
- [2016-…] Maud Ingarao (ENS Lyon-IHRIM)
- [2016-…] Paul Gaillardon (CDD puis CNRS-IHRIM)
- [2016] Christelle Cluze (Stagiaire)
- [2012-...] Nathalie Arlin (Vacataire)
- [2011] Marjorie Burghart (EHESS, CIHAM)
- [2010-2015] Laetitia Faure (CDD puis CNRS-LIRE)
- [2008-2012, responsable] Emmanuelle Morlock-Gerstenkorn (CNRS-ISH)
- [2008] Vanessa Le Rolle (Stagiaire)
- [2007-2011] Christine Berthaud (CNRS-ISH)

Aide à la transcription
- [2011] Cécile Cordier (CDD)
- [2007-2008] Claire Giguet (CDD)

Traitement des images
- [2016-…] Florence Poncet (CDD IHRIM)
- [2011-2013] Françoise Notter-Truxa (CNRS-LIRE)
- [2007-2011] Véronique Églin (INSA, LIRIS)
- [2007-2011] Vincent Malleron (Doctorant, Université Lyon 2, LIRE et LIRIS)
- [2007-2010] Christophe Lemius (CNRS-LIRE)

6) Archivage des données

☐ ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

☐ ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

L'archivage n'est pas encore mis en place mais souhaité, et ce pour la globalité des données du projet. Les informations qui suivent sont donc de l'ordre du prospectif.

Plateforme pour l'archivage pérenne des données

CINES

Durée de conservation des données

Illimitée

Volume des données à conserver

La totalité

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

voir conditions CINES

7) Partage des données à l'issue / au fil du projet

☐ ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

☐ ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR: les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Les données primaires sont accessibles (images + transcriptions, certaines encore en cours). Les métadonnées du corpus sont partiellement moissonnables via OAI-PMH (voir https://www.rechercheisidore.fr/search/?source=10670/2.q6yiyl)

Eléments d'accompagnement qui permettent la réutilisation des données.

Des informations et tutoriels sont présents sur le site des *Dossiers*. Notamment sur la page "Espace de travail" qui renvoie à l'utilisation de l'Agenceur.

Publications sur les données destinées à en améliorer l'exposition

Le carnet de recherche dédié au projet de l'édition numérique des *Dossiers de Bouvard et Pécuchet* : https://flaubert.hypotheses.org/
Ce carnet diffuse les informations et actualités liées au projet et à son évolution, de même qu'il est un lieu d'échanges et de valorisation pour les chercheurs qui souhaitent réutiliser les sources des *Dossiers*.

L'inventaire des pièces du dossier de genèse de *Bouvard et Pécuchet* :
https://flaubert.univ-rouen.fr/ressources/bp_sphere_inventaire.php

Bibliographie non exhaustive :

Stéphanie Dord-Crouslé. Vers une édition électronique des dossiers de Bouvard et Pécuchet. Stéphanie Dord-Crouslé, Stella Mangiapane et Rosa Maria Palermo Di Stefano. *Éditer le chantier documentaire de Bouvard et Pécuchet. Explorations critiques et premières réalisations numériques*, Andrea Lippolis Editore, pp.15-20, 2010. ⟨halshs-00549160⟩

Alexei Lavrentiev, Serge Heiden. Exploration textométrique du corpus des dossiers de Bouvard et Pécuchet. *Revue Flaubert,* Centre Flaubert, 2014, pp.1-12. ⟨halshs-00678874⟩

Stéphanie Dord-Crouslé. Le site et l'état d'avancement du projet Bouvard. Édition des dossiers documentaires de Bouvard et Pécuchet. *Journées d'études internationales des 11*

et 12 décembre 2008, Lyon, École Normale Supérieure - Lettres et Sciences humaines, Dec 2008, Lyon, France. ⟨halshs-00368846⟩

Pierre-Edouard Portier. Manipulations multimodales pour la construction de documents multistructurés. *Colloque: "Bouvard et Pécuchet : les " seconds volumes " possibles - Documentation, circulations, édition"*, ENS de Lyon, dir. Stéphanie Dord-Crouslé, Mar 2012, Lyon, France. ⟨halshs-00678876⟩

Caroline Angé. Édition de fragments : les enjeux de la mise en forme numérique*. colloque: "Bouvard et Pécuchet : les " seconds volumes " possibles - Documentation, circulations, édition"*, ENS de Lyon, dir. Stéphanie Dord-Crouslé, Mar 2012, Lyon, France. ⟨halshs-00678861⟩

Emmanuelle Morlock-Gerstenkorn. La pratique de l'encodage dans le projet d'édition électronique des Dossiers de Bouvard et Pécuchet : quelques exemples. Textes numériques : l'encodage, pratique savante ?, *Séminaire "Édition savante et humanités numériques"* (EHESS), Dec 2011, Paris, France. ⟨halshs-01141447⟩

Emmanuelle Morlock-Gerstenkorn. Les dossiers de Bouvard et Pécuchet de Flaubert - Fragments visuels et fragments logiques au sein du projet d'édition électronique. *Séminaire publication électronique* - IRHT Orléans, Dec 2009, Orléans, France. ⟨halshs-00438078⟩

Vincent Malleron. Outils d'analyse d'image pour les dossiers de Bouvard et Pécuchet : un panorama. *Édition des dossiers documentaires de Bouvard et Pécuchet. Journées d'études internationales des 11 et 12 décembre 2008, Lyon, École Normale Supérieure - Lettres et Sciences humaines*, Dec 2008, Lyon, France. ⟨halshs-00377381⟩

Stéphanie Dord-Crouslé. Fragments textuels et catégories de classement. Un cas d'utilisation de XML-TEI dans le dispositif éditorial du corpus BOUVARD. *Éditions critiques et génétiques en Rhône-Alpes*, Jun 2013, Grenoble, France. ⟨halshs-00838143⟩

Vincent Malleron. Le numérique et l'interdisciplinarité au service des dossiers de Bouvard et Pécuchet : Vers une mobilité retrouvée. *Séminaire de bilan et de prospective du Cluster 13 « Culture, patrimoine, création » mis en place et soutenu par la Région Rhône-Alpes*, vendredi 23 octobre 2009, Château de Montchat, Oct 2009, Lyon, France. ⟨halshs-00426391⟩

Stéphanie Dord-Crouslé, Emmanuelle Morlock-Gerstenkorn. Le "modèle abstrait" du corpus Bouvard : première approche. J*ournée d'étude " Constitution et exploitation de corpus issus de manuscrits - Lectures, écritures et nouvelles approches en recherche documentaire " organisée par Cécile Meynard et Thomas Lebarbé*, Mar 2009, Grenoble, France. ⟨halshs-00368044⟩

Vincent Malleron, Véronique Eglin, Hubert Emptoz, Stéphanie Dord-Crouslé, Philippe Régnier. *Hierarchical decomposition of handwritten manuscripts layouts. Computer Analysis of Images and Patterns*, Sep 2009, Muenster, Germany. pp.221-228, ⟨10.1007/978-3-642-03767-2⟩. ⟨halshs-00420059⟩

Stéphanie Dord-Crouslé, Emmanuelle Morlock, Raphaël Tournoy. **Nouveaux objets éditoriaux. Le site d'édition des dossiers documentaires de Bouvard et Pécuchet (Flaubert)**
Les Cahiers du numérique, Lavoisier, 2012, 7 (3-4/2011 " Empreintes de l'hypertexte. Rétrospective et évolution ", sous la dir. de Caroline Angé), pp.123-145. ⟨10.3166/LCN.7.3-4.123-145⟩

Vincent Malleron, Stéphanie Dord-Crouslé, Véronique Eglin, Hubert Emptoz, Philippe Régnier. Extraction automatisée de lignes et de fragments textuels dans les images de manuscrits d'auteur du XIXe siècle. *MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication*, Nov 2009, Avignon, France. ⟨halshs-00443548⟩

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

Conditions indiquées plus haut pour les images et manuscrits (voir : Accès, partage et limites d'accessibilité des données).
Pour les transcriptions : obligation de citation.

e) Annexe n°4 : Introduction et sommaire du livre « Dix ans de Corpus d'auteurs »

Dix ans de corpus d'auteurs

Introduction

Le 14 et 15 décembre 2011, une vingtaine d'humanistes réunis en salle J 636 de la Sorbonne[32] s'embarquait dans une drôle d'aventure au cours d'un « atelier de lancement » associant présentations de projets scientifiques, démonstrations de fonctionnement de logiciels (étaient introduits des outils comme BaseX) et discussions sur des questions d'organisation et de fonctionnement. Ce noyau allait grandir et devenir le consortium CAHIER : « Corpus d'auteurs pour les humanités : informatisation, édition, recherche ». En quelques années, il est passé de quinze à soixante projets, d'une vingtaine de personnes à quasiment deux cents, et de quelques cinq cents pages numérisées et en cours de publication, à un demi-million d'images en ligne ainsi qu'approximativement 327.000 fichiers texte.

Ce consortium ne représentait pas une institution et n'était pas non plus un partenariat entre institutions. La philosophie de l'entreprise, lancée et coordonnée par Marie-Luce Demonet (Université de Tours) de 2011 à 2015, était de créer un espace de rencontre et de discussion entre spécialistes du texte présentant un fort intérêt pour le numérique, et souvent des connaissances et des compétences dans ce dernier domaine. Il s'agissait à la fois de faire émerger une communauté, en sortant de leur isolement des entreprises similaires mais s'ignorant les unes les autres, et de transcender les clivages disciplinaires comme les différences de statut (enseignants-chercheurs, chercheurs, doctorants ou ingénieurs), tout en s'affranchissant, grâce au soutien de l'infrastructure Huma-Num, des lourdeurs administratives liées au fonctionnement en laboratoire. Au cours de ses dix années d'existence, d'abord sous la coordination de Marie-Luce Demonet, puis sous celle de Thomas Lebarbé et de Fatiha Idmhand, CAHIER a joué le rôle de point de rencontre et de contact national entre des « équipes projets » dont les questions de recherches sur les corpus d'auteurs impliquaient l'utilisation de méthodes et/ou outils informatiques.

Les quatre premières années du consortium CAHIER ont été celles de l'élargissement. Si les humanités numériques étaient, en ces années 2012 à 2015, une discipline en plein développement en Europe et dans le monde, avec de grandes conférences régulières, des revues dédiées et même des manuels devenus des références de l'état de l'art (Schreibman, Siemens et Unsworth, 2004), elles constituaient encore en France, à l'époque, une curiosité désignée par une étiquette à la fois *hype* et un peu obscure. Les présentations et les réflexions sur le domaine, sur les opportunités de l'informatique et sur les problèmes épistémologiques posés par son acclimatation dans les sciences humaines, se multipliaient (Dacos et Caverni, 2010 ; Douehi, 2011 ; Mounier, 2012 ; Gefen, 2015…), mais peut-être moins les guides pratiques, les écoles d'été et les formations concentrées sur quelques jours à destination de chercheurs confirmés cherchant une solution à un problème précis. Dès le début de son existence, CAHIER a répondu à cette demande en permettant à ses adhérents d'échanger de façon informelle à propos de leurs éditions et de leurs problèmes au cours de différents types de rencontres annuelles (groupes de travail, journées d'études et assemblées générales), en soutenant la participation à, et en publicisant, des formations organisées ailleurs, et en proposant lui-même des ateliers annuels consacrés aux méthodes de traitements des textes patrimoniaux en vue de leur publication en ligne et à l'exploitation des données textuelles.

La seconde période d'existence du consortium, de 2015 à 2019, a été caractérisée par les efforts de cristallisation d'une culture commune mise progressivement par écrit. Tout en continuant le processus d'élargissement et les actions de formation de la première période, le consortium a connu une multiplication des groupes de travail : à celui sur les « Questions juridiques », le plus ancien, se sont ajoutés ceux sur le *crowdsourcing*, la correspondance, l'évaluation des éditions numériques (EVENT) ou la typologie textuelle. Le groupe « R2CAHIER » a essayé, en même temps, d'accompagner le mouvement de l'informatisation et de l'édition vers davantage de réflexions et de propositions sur ce que signifie la recherche dans les sciences des textes avec le numérique. Autrement dit, il s'est agi d'accompagner le passage de « cahIEr » à « cahieR » pour reprendre le titre d'une communication que nous faisions à l'époque (Galleron, Idmhand et Meynard, 2016). C'est également à cette occasion que le consortium s'est doté d'un

32 Voir le programme de la rencontre sur [https://cahier.hypotheses.org/40#more-40], consulté le 28 août 2021.

comité scientifique international composé de Bertrand Jouve, Lou Burnard, Susan Schreibman et Elisabeth Burr, chargé de se prononcer sur ses bilans annuels et sur les orientations proposées par le comité de pilotage et adoptées par l'assemblée générale.

Cette évolution de CAHIER correspondait en grande partie celle des humanités numériques en France qui passaient, au cours de cette période, de la découverte à l'apprivoisement, puis à une forme de maîtrise des technologies digitales. Plus balbutiante du côté des lettres, celle-ci était déjà plus affirmée dans les projets ancrés dans la linguistique et l'histoire, où l'usage des outils numériques et « computationnels » était plus ancien. La fierté de CAHIER est d'avoir accompagné ses adhérents à passer du réflexe du « site web vitrine » à une réflexion plus poussée sur la portée et les enjeux des choix des « formats » éditoriaux, à mieux comprendre l'écosystème numérique en termes de « flux » et de « traitement » et à penser les objets d'études textuels en termes de « données » – quelles que soient les réserves qui subsistent encore envers l'utilisation de ce terme dans les sciences humaines. En somme, la nécessaire distinction entre l'édition numérique (archive éditorialisée, édition de lecture ou édition scientifique, pour reprendre la terminologie d'un des guides de CAHIER[33]) et l'interface web qui y donne accès, idéalement pas de façon exclusive, est dorénavant bien implantée dans l'ensemble des équipes de CAHIER et au-delà.

Au fur et à mesure de son agrandissement, le consortium a également vécu la structuration progressive des institutions en vue de soutenir les humanités dans leurs rapports et travaux scientifiques avec le numérique. Cette structuration a ouvert de nouvelles possibilités, mais aussi imposé de nouvelles exigences. Les membres du consortium ont pu bénéficier, en temps réel, de l'évolution HumaNum, dont la « grille de services », étoffée au fil du temps, couvre à présent l'ensemble de la chaîne de traitement d'un projet en humanités numériques, depuis l'organisation et la collecte et jusqu'à la préservation, la publication, le traitement et la réutilisation des données. Huma-Num propose ainsi un moteur de recherche pour les sciences humaines (ISIDORE), l'accès à des licences pour l'utilisation d'éditeurs XML (comme oXygen), des espaces de collaboration protégés avec une importante capacité de stockage (ShareDocs), des machines virtuelles pour tester, faire fonctionner ou partager un outil de traitement et, surtout, un espace pour le dépôt, la préservation et le partage des données à travers un entrepôt public, sûr et pérenne (Nakala). D'autres services, plus récents, ont été mis à la disposition de la communauté scientifique pour lui permettre de rédiger et publier facilement ses articles à l'aide de Markdown (Stylo) ou ses Data Papers (à l'aide de Jupyter nbviewer), d'exploiter ses données (à l'aide de R par exemple) et de mesurer l'impact de ses projets en ligne (Matomo). Ces services ont été déployés en même temps que les membres de CAHIER découvraient l'intérêt de l'*open access*, qui n'a pas fini de bouleverser le système de publication en sciences humaines. Plus récemment, c'est l'exigence de créer des données FAIR (*findable*, *accessible*, *interoperable* and *reusable*) que HumaNum a popularisée en direction de se différents consortia. Si des progrès restent à faire sur l'ensemble de ces aspects, et surtout sur le dernier point, CAHIER a joué un rôle majeur dans la découverte et la compréhension de ces principes, ainsi que dans le soutien et la formation de ses membres à leur adoption[34].

À la fin de l'année 2021, le consortium clôture « dix années de CAHIER » et n'existera plus en tant que regroupement de projets soutenus par le CNRS via l'infrastructure de recherche Huma-Num. Toutefois, à en juger par le nombre de collaborations bilatérales ou multipartenaires qui ont d'ores et déjà émergé dans le sillage du consortium, il semble certain que CAHIER ne va pas purement et simplement disparaître.

Ces dix années d'agrandissement et de « professionnalisation », si on peut dire, de la communauté scientifique CAHIER ne pouvaient se clore sans un bilan, non seulement du nombre de documents numérisés et de fichiers produits, dans de nombreux formats (image, texte, annotations, graphes…), mais également des idées et perspectives dégagées grâce aux activités d'informatisation et d'édition. C'était l'objectif visé par le colloque qui a été organisé à Bordeaux du 7 au 10 juin et qui a réuni plus d'une centaine de participants sur place et en ligne, et proposé trente communications, quatre conférences plénières, trois ateliers et une « FAIR line » ouverte à tous pendant trois jours. Seule une sélection de ces présentations figure dans ce volume, à la fois pour des raisons de cohérence et de représentativité.

Organisé en trois parties, l'ouvrage ici proposé offre d'abord, dans la partie intitulée « Les toiles de Pénélope », des contributions qui mettent l'accent sur des entreprises qui ont voulu mettre à la disposition des chercheurs, mais aussi du large public, une documentation et des sources jadis difficilement accessibles, et pour

[33] Voir « Les publications numériques de corpus d'auteur : guide de travail, grille d'analyse et recommandations », [https://cahier.hypotheses.org/guides/les-publications-numeriques-de-corpus-dauteurs], consulté le 28 août 2021.

[34] Voir en ce sens « Guide pour la FAIRisation des données des corpus d'auteur » [https://cahier.hypotheses.org/guide-pour-la-fairisation-des-donnees-des-corpus-dauteurs], consulté le 29 août 2021.

lesquelles l'édition papier est à la fois trop onéreuse et souvent peu adaptée. En préambule, les réflexions de Marie-Hélène Lay sur l'annotation manuelle mettent en perspective les travaux menés. Le lien entre les activités traditionnelles des chercheurs en SHS, comme l'établissement d'index et de concordanciers, et le travail de création de métadonnées de contenu en contexte numérique est mis en avant dans cette contribution. Par la suite, on trouvera dans cette partie des présentations de projets ayant rejoint le consortium CAHIER très tôt, comme les différents travaux sur Émile Zola (ArchiZ, Correz et ScéNa présentés dans la communication d'Olivier Lumbroso), et d'autres plus récents, comme le projet Frénaud numérique (de Marianne Froye) ou sur les journaux intimes de Bourget (de Dominique Ancelet-Netter et Guilaume Boyer). Le fil rouge qui parcourt toutes ces contributions est l'évolution des pratiques en même temps que les progrès de la technologie. On le voit à travers la fascinante reconstitution des différentes étapes de traitement du patrimoine Zola (romans, manuscrits, photographies, lettres de l'écrivain et adressées à l'écrivain), depuis la conception d'un site d'auteur, aux ambitions totalisantes mais aux parties bien cloisonnées et encore organisé par la logique de l'œuvre imprimée primant sur les parties « secondaires » (correspondance, interviews…), jusqu'aux travaux les plus récents, qui bouleversent les rapports hiérarchiques entre les parties de l'œuvre et intègrent encodage et annotation en attendant de proposer des visualisations. Autant d'entreprises dans lesquelles l'équipe Zola, une nouvelle fois, s'est déjà lancée, comme on le verra dans le chapitre final de cet ouvrage.

Toujours dans cette première partie, on trouvera aussi bien des réflexions sur l'apport du numérique à l'entreprise d'édition, que des questionnements sur les façons de faire et les nouvelles difficultés liées au changement de médium. La présentation des travaux menés sur *Le Parallèle des Anciens et des Modernes* de Charles Perrault (article de Delphine Reguig et Emmanuelle Perrin) souligne ainsi l'utilité des hyperliens et la liberté d'accumuler les notes, offerte par le détachement de la page imprimée, pour re-contextualiser une œuvre complexe, aux ambitions encyclopédiques, dont la lecture ne peut plus se faire aujourd'hui sans ces nombreuses aides permettant de l'actualiser, et de mettre en lien ses enjeux avec ceux de notre modernité. La communication de Dominique Ancelet-Netter et Guillaume Boyer montre, quant à elle, l'avantage du numérique lorsqu'il s'agit de concilier volonté de faire sortir de l'ombre un écrivain oublié (Paul Bourget) grâce à la valorisation d'un fond inédit, et l'obstacle juridique créé par le testament dudit écrivain, qui s'oppose à la publication de ses cahiers inédits. Recensement et description de l'existant, équipement avec des mots clé issus d'une analyse lexicométrique, génétique et littéraire, balisage des entités nommées sont autant d'éléments qui ne violent pas la volonté de l'auteur, et qu'internet permet de publiciser à la place d'une édition traditionnelle que l'on imagine difficilement s'intéresser à de tels objets. Travaillent, quant à elles, sur des auteurs qui ont accepté la circulation de leurs brouillons, Marianne Froye et France Marchal-Ninosque soulignent en revanche les difficultés de repenser le texte dans sa version numérisée, de maîtriser la « réorientation » qu'il subit en changeant de medium, et même d'assumer, en tant que chercheur, de nouvelles responsabilités (intellectuelles, techniques et juridiques) par rapport à l'œuvre. Tout en constituant une chance, le numérique est aussi un défi, qui impose non seulement de s'approprier des connaissances et des modes de pensée peu familiers aux humanistes, mais aussi d'inventer de nouveaux vocabulaires, et de converger vers des pratiques communes à partir de règles et de protocoles d'encodage qui, en dépit de leur rigueur, laissent une grande latitude, et dont les différentes disciplines se saisissent différemment.

La seconde partie de l'ouvrage intitulée « Un texte, des intertextes » se focalise sur une problématique courante dans la création de corpus d'auteurs, à savoir la restitution, grâce au numérique, des liens de nature diverse que l'on identifie soit entre différents témoins d'un même texte, soit entre des textes différents, d'un même auteur ou d'auteurs distincts. Texte de base et variantes, texte cité et texte citant, auteur(s) citant(s), auteur(s) source, compilateurs, scripteurs et éditeurs entretiennent des relations d'autant plus complexes que les objets dont il est question sont plus anciens et ont fait l'objet d'une longue transmission, à travers des relais plus ou moins déformants. Les problèmes auxquels se confronte dans ce cas l'éditeur scientifique en régime numérique sont à la fois liés au choix de la meilleure stratégie de représentation, dans un format numérique et avec les contraintes d'un langage structuré, de cette multitude de relations, et ceux de l'accès à des outils permettant de simplifier et de fiabiliser ce type de travail. Les deux questions s'entrecroisent dans les travaux présentés dans cette partie, à commencer par celle d'Estelle Debouy qui, après avoir discuté des différentes façons d'éditer les fragments des atellanes (en tant qu'*aparatus fontis* ou *aparatus testis*), présente un changement de perspective selon lequel le fichier XML-TEI est un sous-produit du travail numérique, et non pas son point de départ. Sa communication se focalise ainsi sur l'utilisation d'*ekdosis*, un système basé sur LaTeX, qui permet à la fois de donner à l'éditeur scientifique la possibilité de s'exprimer largement sur les choix opérés dans le cadre de son édition, et de réaliser une édition en XML-TEI à des fins d'extraction de données et d'analyse. De leur côté, Isabelle Draelants et Emmanuelle Kuhry présentent le cheminement tortueux vers l'adoption d'une interface de balisage en lien étroit avec les besoins d'un travail de recensement, identification et de mise en perspective de citations compilées dans des ouvrages encyclopédiques médiévaux. La sagacité et l'érudition de l'annotateur, engagé dans un patient démêlage de l'écheveau de renvois des compilations médiévales – parfois

explicites, mais le plus souvent allusifs, vagues, emboîtés ou même erronés – est ainsi servi par un environnement éditorial ad-hoc, élaboré au Pôle numérique de Caen. Le balisage proposé permet de garder la nécessaire lisibilité, tout en offrant une certaine souplesse de traitement intellectuel de la réalité des documents, parfois plus complexe que ce que l'emboîtement des hiérarchies permet d'exprimer. Utilisatrices du même environnement de travail et d'édition, Marie-Agnès Lucas-Avenel et de Marie Bisson se concentrent, de leur côté, sur le traitement (dans les deux sens de « création intellectuelle » et de « choix techniques ») de l'apparat de notes dans l'édition critique de l'œuvre de Geoffroi Malaterra. Les autrices soulignent l'accroissement de la lisibilité de l'édition ainsi réalisée, notamment en version dynamique grâce au site web, ainsi que les gains de temps et de cohérence que ce flux de traitement garantit, même s'il implique aussi de prendre certaines distances par rapport à la pratique ecdotique de la version papier, ou de réaliser des manipulations supplémentaires (notamment au sujet des additions et des omissions) pour s'aligner sur celle-ci.

Des conclusions similaires se dégagent du travail concernant le corpus Ichtya. L'accent de cet article tombe toutefois moins sur l'outil et les processus, que sur un autre type de valorisation des textes permis par le numérique, au-delà de la création d'éditions de lecture et/ ou scientifiques. Prenant le relais des savants qui les ont précédés depuis l'Antiquité et jusqu'au XVIII^{e} siècle, Marie Bisson, Pierre-Yves Buard, Brigitte Gauvin et Barbara Jacob décrivent les étapes de création d'un thesaurus de noms de poissons et autres créatures aquatiques. Relié finement, grâce à un jeu d'identifiants, à ses sources constituées en bibliothèque numérique, le vocabulaire ichtyologique est décliné en latin, grec, français et d'autres langues vernaculaires, puis également relié, là où cela est possible, à la terminologie et classification moderne (avec, parfois, plusieurs noms modernes correspondant à une même graphie ancienne, ou inversement). Des visualisations sous formes de graphes permettent de synthétiser dans une forme compréhensible et interprétable les notices ainsi créées, pour mieux comprendre, à terme, ce que nos prédécesseurs savaient de la vie aquatique. L'étendue impressionnante, mais aussi le caractère parcellaire et non-unifié de leur savoir, sont ainsi à la fois restitués, et compensés grâce aux technologies XML et web sémantique.

La troisième et dernière partie du volume consacrée aux « Explorations » propose des contributions dans lesquelles l'accent se déplace de la création des corpus vers leur utilisation. Sans que des problématiques comme celles évoquées plus haut soient absentes de ces chapitres, leur discussion se fait en marge d'autres questions scientifiques et portent sur les régularités identifiables dans le corpus et sur leur interprétation en lien avec les connaissances déjà acquises en lettres et langues. Les questions abordées ici ont trait à la construction de parcours alternatifs à la lecture de près, grâce à la préparation numérique des corpus – depuis l'exploitation savante, orientée par une question scientifique, et jusqu'à des approches plus ludiques et pédagogiques.

On y trouvera ainsi des travaux portant sur l'identification de textes similaires, par leurs thématiques et leur argumentation, comme ceux qui sont menés par Karine Abiven et Gaël Lejeune sur un corpus de mazarinades non corrigé après océrisation. Leur communication montre ainsi que, même si les données ne sont pas très finement annotées, une question de recherche pertinente et une bonne compréhension du fonctionnement des outils numériques peut mener au dégagement de conclusions inédites et solides. À terme, le travail d'orfèvre mené pour l'identification des sources des encyclopédies médiévales pourrait ainsi être accéléré grâce à des traitements semi-automatiques comme ceux dont ils donnent l'exemple, et qui portent à la fois sur l'acquisition du texte en format dynamique, que sur son exploitation en dépit de son bruitage. Cet appel à la conception de nouvelles approches et de nouveaux outils s'entend aussi dans la réflexion de Marie-Hélène Lay sur le besoin de développer une série de briques logicielles facilement accessibles, qui viennent accélérer et la préparation et l'exploitation du texte.

Deux autres contributions concernent la description d'un style à partir d'un corpus interrogé avec des outils numériques : celui de Maupassant, chez Ouafae Benzina, ou celui des correspondants peu lettrés pendant la Première guerre mondiale, étudiés par Agnès Steuckardt, Sybille Große, Beatrice Dal Bo et Lena Sowada. Dans la première communication, on verra ainsi un exemple d'utilisation efficace des listes de spécificités du vocabulaire, telles que fournies par un logiciel de textométrie (en l'occurrence, Hyperbase), pour observer l'évolution de l'écriture d'un auteur classique de la littérature française. La seconde offre, quant à elle, un parcours fascinant à travers les formules d'ouverture et de clôture des lettres échangées en temps de guerre. Elle montre comment certains scripteurs s'approprient des « préconstruits discursifs » en les modulant en fonction de leur propre sensibilité. L'écriture en temps de guerre, se substituant brutalement à une intimité peut-être définitivement perdue, réinvente cette dernière malgré les contraintes d'un genre (celui de la lettre) et d'un medium (celui de l'encre et du papier) peu connus, qui dissuade d'utiliser les prénoms dans les adresses et laisse peu de place pour l'expression du désir.

Un autre chapitre, issu d'une table ronde, interroge les modalités de re-création de l'œuvre de Zola, *La Bête humaine et L'Œuvre*, en l'occurrence, sur de nouveaux supports (bande dessinée, site enrichi, jeu vidéo) en vue de favoriser l'appropriation d'un classique par le public scolaire (et pas seulement). Les questions de la communication entre entreprises savantes d'édition et approches techno-créatives, de la fidélité au texte original dans le cadre d'un

« objet sémiotique secondaire » qui l'éclate, le restructure et lui ajoute des éléments étrangers, de l'efficacité intellectuelle et émotionnelle des parcours de découverte ainsi proposés sont abordées tour à tour dans cette contribution qui ouvre de belles perspectives de dissémination des travaux menés par CAHIER dans le large public.

La dernière contribution du volume adopte un point de vue surplombant, et interroge non pas un corpus ou une série de corpus, mais d'un jeu de données absorbé à partir du catalogue de la *Modern Language Association*. Ferrer présente une nouvelle méthode pour analyser la circulation et les tendances de la critique littéraire, fondée sur le calcul statistique et tirant des conclusions à partir des lois des grands nombre. Désignée par son autrice sous le nom de criticométrie, cette méthode s'inspire des approches et indicateurs de la scientométrie pour analyser et comprendre la visibilité relative et les influences des différents espaces littéraires les uns sur les autres.

Au-delà des problématiques scientifiques de chaque communication, telles qu'on a pu les découvrir plus haut, ou plus exactement à partir de celles-ci, se dessine à travers ce volume un état des lieux des humanités numériques dans les sciences du texte, en France, après deux décennies de développement soutenu. Au fil des contributions, on retrouve des questions plus anciennes mais dont la réponse ne semble pas définitivement trouvée, et d'autres qui émergent maintenant, alors que « l'apprivoisement » de la machine et des outils est (en partie) achevé. D'autres, enfin, sont des questions qui ont pu, à un certain moment, être jugées secondaires par CAHIER, comme celle des « usages », mais qui se pose aujourd'hui dans une multitude d'espaces et de communautés.

Parmi les premières, on revient sur celle de la relation entre l'ingénieur/l'ingéniérie et l'humaniste/la recherche sur les textes : quel est le rôle de chacun dans un projet d'édition ? Différents modes de collaboration sont abordés dans plusieurs communications (Froye et Marchal-Ninosque, Lucas-Avenel et Bisson, même Reguig et Perrin). Si deux modèles semblaient s'affronter à la fin des années 2000 et au début de la décennie 2010 (Dacos et Mounier, 2015), celui du « pont » semble avoir prévalu, plutôt que celui de l'humaniste qui apprend à coder, même si, indiscutablement, les humanistes qui ont écrit les chapitres de cet ouvrage et ont fait partie du consortium CAHIER sont évidemment allés plus loin qu'ils ne le pensaient dans la compréhension des possibilités, des contraintes et des modes de fonctionnement de la machine.

Une autre question est celle de la publication et du recensement des résultats d'un travail d'édition scientifique numérique. À lire les contributions publiées dans ce volume, le constat est sans appel : l'écosystème de l'édition numérique ne s'est pas étoffé depuis les premières tentatives en ce sens dont l'entreprise menée au Pôle numérique de Caen reste un exemple unique. En dépit d'efforts institutionnels pour offrir des solutions unifiées (ex. la plateforme FANUM de l'université de Franche-Comté ou la plateforme e-Man de l'Institut des textes et Manuscrits Modernes), il apparaît, pour la plupart des contributeurs et même pour une grande majorité des membres de CAHIER, que lorsqu'il s'agit d'éditer une édition numérique dans le plein sens du terme, c'est-à-dire publiciser un travail scientifique après l'avoir soumis au scrutin des pairs, en bénéficiant des capacités de distribution et de promotion d'une équipe spécialisée en la matière, la solution n'est pas évidente. La solution adoptée par les chercheurs est plutôt celle d'un contournement du problème, grâce au *do it yourself* : création d'un site web, souvent dès le début du projet, couplé à un carnet de recherche ayant à la fois la fonction de faire connaître l'objet numérique en cours de création, et de le soumettre à une forme de scrutin continu par les pairs (peu pratiqué, au demeurant). Peut-être le sentiment de perpétuel inachèvement des éditions numériques, dont il est question dans plusieurs contributions de ce volume (Ancelet-Netter et Boyer, Debouy, Froye et Marchal-Ninosque, pour n'en citer que quelques-unes) est-il aussi responsable de cet état de faits : alors que le médium papier oblige à « tirer la ligne » à un moment donné, et déclenche à partir de là l'entrée de l'édition scientifique dans un circuit éditorial, la fluidité du numérique, la possibilité et même, parfois, l'obligation de recommencer (voir les communications de Lumbroso, ou Draelants et Kuhry), ajourent *sine die* la recherche d'un éditeur professionnel et n'incitent pas ce dernier à réfléchir à un positionnement sur le « marché » (naturellement ouvert, v. supra) du texte patrimonial, enrichi grâce au numérique.

Une conséquence inattendue de cet état de faits éditorial est que la relation avec des bibliothèques reste très ténue. Alors que le moindre livre publié, aux garanties scientifiques parfois discutables, fait l'objet d'un référencement dans les catalogues de la BNF, des éditions complexes, ayant mobilisé des équipes multidisciplinaires et fortement outillées y sont totalement invisibles[35]. Il n'existe actuellement pas de solution unifiée, à la disposition des chercheurs, pour identifier les éditions en cours, achevées, ou même projetées ; le rôle joué par le portail de CAHIER en la matière

[35] Une recherche dans les catalogues de la BNF, en croisant le terme « Montaigne » avec « document électronique », ou bien du terme « MONLOE » dans tous les champs, ne mène pas à la date de rédaction de ces lignes à aucune des ressources sur l'auteur figurant dans la Bibliothèque virtuelle humaniste de Tours ([http://www.bvh.univ-tours.fr/], consulté le 27 août 2021). À noter qu'il ne s'agit pas là d'un projet éditorial récent, des mises en ligne de documents numérisés de Montaigne remontant (au moins) à 2014.

a été très modeste et touche à sa fin. Quoiqu'un « dépôt légal numérique[36] » ait été défini récemment, les moteurs de recherche sur internet continuent à rendre de plus grands services sur ce point que les instruments traditionnels pour l'identification, puis l'accès aux éditions scientifiques numériques.

Une autre conséquence de la fragilité de l'écosystème éditorial numérique – dont il est vrai que les communications ne parlent pas en tant que tel, mais que l'on entend au détour d'un chapitre (voir en ce sens certains paragraphes liminaires et conclusifs de Draelants et Kuhry, ou Nastase *et alii*, par exemple) –, est le manque de reconnaissance persistant des travaux numériques entrepris. À vrai dire, la situation est paradoxale avec, d'un côté, une forme d'exigence à prévoir un volet numérique pour les projets des humanités recherchant des financements auprès de l'Agence Nationale de la Recherche, des acteurs régionaux ou locaux, et la quasi-inexistence de systèmes et de moments d'évaluation par les pairs en aval du projet. Il n'existe pas d'équivalent français d'une revue comme *RIDE*[37], par exemple, ni un véritable jeu de critères susceptibles d'être pris en compte lors de l'évaluation de l'activité numérique d'un chercheur ou d'une équipe par des organismes nationaux comme le Conseil National des Universités (CNU) ou le Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (HCERES). Si CAHIER a essayé, au cours de son existence, de contribuer à l'émergence de ceux-ci par le biais d'articles, de grilles d'évaluation et même du présent volume, l'émergence de points de repère plus fermes reste un enjeu des années à venir.

Parmi les questions les plus récentes ouvertes par l'évolution des sciences des textes numériques, on peut mentionner celles de la création de référentiels communs et de bonnes pratiques pour l'annotation sémantique. Quand ils s'engagent dans le balisage des entités nommées, les corpus ici présentés s'appuient sur des catalogues, des ontologies ou *thesauri* disponibles sur le web, comme ISNI ou VIAF, pour n'en citer que deux, pertinents pour l'identification non ambiguë des personnes mentionnées (voir en ce sens les communications de Reguig et Perrin, ou d'Abiven et Lejeune). Mais ces besoins s'avèrent rapidement plus importants que ce qui se trouve en ligne, de sorte que les projets génèrent leurs propres bases de connaissances. L'exemple le plus évident est celui du vocabulaire de noms de poissons créé pour le projet Ichtya, mais d'autres nécessités se manifestent quand on regarde de près les contributions. Les corpus existants ou en création pourraient ainsi bénéficier de ressources négociées et partagées pour décrire plus rigoureusement les typologies textuelles (un groupe de travail de CAHIER a développé un thésaurus[38]), les « événements », les types de relations intertextuelles, etc. C'est là aussi un aspect de la réutilisabilité, dépassant la simple mise à disposition d'un fichier numérique, d'un encodage ou d'une banque d'images, ou le libre accès à un code et à des données. Quant à l'identification des bonnes pratiques, elle devient plus nécessaire que jamais au moment où les équipes n'en sont plus au déchiffrage de l'alphabet numérique, et où, face à la complexité des situations de terrain, elles sont confrontées à la nécessité de choisir (par exemple, à partir d'une multitude d'encodages possibles) et d'adapter (des éléments ou des solutions mises en pratique ailleurs). Le succès des guides (comme le guide d'annotation des correspondances[39], plusieurs fois mentionné dans les contributions) prouve l'existence d'un appétit pour des « Guidelines » plus détaillées, un besoin auquel le groupe de discussion TEI de CAHIER a également essayé de répondre au cours des dernières années d'existence du consortium.

Une autre caractéristique du moment actuel des humanités numériques que ce volume permet d'apercevoir est la relative absence des « études distantes », ou, si l'on veut, un désintérêt pour les « lectures de loin ». On peut la constater par le peu de travaux portant sur de grandes masses d'informations et par le peu de visualisations proposées dans ce volume pour illustrer les propos. Dans la plupart des cas, les visualisations proposées ont davantage pour but d'illustrer le propos que de s'inscrire dans un discours de démonstration, puisqu'il s'agit d'images ou de captures écrans, rarement de résultats issus de logiciels de traitement et de calcul. Peut-être peut-on affirmer, en paraphrasant l'article de Marianne Froye et France Marchal-Ninosque, qu'au cours des dix années d'existence de CAHIER, l'image a constitué bien plus l'alpha que l'oméga des études entreprises, point de départ d'une numérisation à partir de laquelle il s'agissait d'extraire, par des procédés manuels ou (semi)automatiques (voir en ce sens la communication d'Abiven et Lejeune), en totalité ou en partie, la substantifique moelle du texte, elle-même destinée à être ultérieurement enrichie avec des annotations diverses. Des cartes, des graphes et des arbres, pour faire référence à un ouvrage connu de Franco Moretti, on en trouvera donc très peu dans ce volume, même si l'on peut citer de suite les graphes sur les systèmes de la littérature mondiale de Carolina Ferrer, ceux générés par le projet Ichtya (chapitre Bisson – Buard – Gauvin – Jacob), ou les cartes qui figurent sur le site internet de *L'Œuvre* de Zola, voire même dans le jeu vidéo de *La Bête humaine* (contribution d'Alina Nastase, avec Laurélenn Brière, Olivier Dobremel et Jean-

36 Voir [https://www.bnf.fr/fr/le-depot-legal-numerique], consulté le 27 août 2021. On peut considérer que les ressources concernées et les modalités de dépôt semblent peu adaptées aux projets d'édition numérique ici décrits.

37 Voir [https://ride.i-d-e.de/], consulté le 10 septembre 2021.

38 Pour découvrir le thesaurus « typologie textuelle », voir [https://opentheso.huma-num.fr/opentheso/], consulté le 2 septembre 2021.

39 « L'édition numérique de correspondances – guide méthodologique », [https://cahier.hypotheses.org/guide-correspondance], consulté le 29 août 2021.

Sébastien Macke). Cette caractéristique ne doit pas être considérée comme un manque d'ambition ou un refus pour la « lecture de loin ». Elle traduit à notre sens un autre versant des sciences des textes numériques, qui n'ont pas forcément besoin de grandes masses de textes, préférant plutôt procéder avec des données « riches ».

Comme d'autres humanistes travaillant avec le numérique, ceux qui ont participé aux travaux du consortium CAHIER pendant dix ans, ceux qui ont communiqué durant le colloque « Dix ans de corpus d'auteurs » et ceux qui ont contribué à ce volume entendent « sortir de leurs routines afin de revenir fortement équipés pour forcer les autres à quitter les leurs[40] ». Comme le montre ce volume, ils le font en construisant patiemment des machines de laboratoire complexes, qui approfondissent plus qu'elles n'élargissent le domaine du visible. La voie de CAHIER, et peut-être même plus généralement la voie française des humanités numériques dans les sciences du texte, n'est pas seulement celle de la construction de grands corpus multi-langues destinés à être ultérieurement raffinés (comme dans l'action COST « Distant reading[41] »), mais plutôt celle de l'élargissement progressif de corpus raffinés, par cercles concentriques, à partir d'un premier objet ou d'une petite série d'objets traités « à fond ». Cette approche reste évidemment ancrée dans une tradition éminemment philologique, mais qui se révèle gagnante face à l'avalanche de données « captées » et mise à la disposition de tous sans contextualisation. La majorité des humanistes continue de situer avec précision ses objets d'études, même quand ils sont transformés en données. Évidemment, dans un contexte numérique, où l'automatisation peut rendre de grands services à leurs travaux longs et chronophages, se pose de plus en plus la question des outils qui leur permettent de gagner du temps pour aller de l'avant, pour que les éditions scientifiques de corpus d'auteurs ne restent pas dans un perpétuel état de projet pilote ou expérimental. Et, au-delà de la formation à des outils d'extraction d'observables, se pose également la question de l'établissement d'une science interprétative de ces derniers, car penser avec des graphiques, des cartes et des arbres ne va pas de soi dans les disciplines littéraires, quel que soit l'attrait exercé de prime abord par des visualisations plus ou moins joliment colorées.

I.Galleron, F.Idmhand

Table des matières

40 « [If you wish to] go out of your way and come back heavily equiped so as to force others to go out of their ways, [the main problem to solve is that of mobillization]. » (Latour, 1986 : p. 7). Traduction française de Ioana Galleron. Les parties entre crochets n'ont pas été traduites.

41 Voir [https://www.distant-reading.net/], consulté le 29 août 2021.

III. Explorations

- Des données au corpus : l'exploitation numérique des mazarinades
- Corpus romanesque et lexicométrie : le vocabulaire de Maupassant
- La routine et le style. Exploration outillée des formules d'ouverture et de clôture dans des correspondances peu-lettrées de la Première Guerre mondiale
- Lectures zoliennes. Des manuscrits aux adaptations, quelles données numériques en jeu ?
- De Confucius à Djebar, de Dante à Lispector : ce que la criticométrie nous apprend sur la réception des écrivains et de leurs œuvres

f) Annexe n°5 : Projet de consortium de réseau « REST CAHIER »

Projet de suite au Consortium CAHIER

Proposition élaborée par Alexey Lavrentev (Ingénieur de recherche CNRS, UMR IHRIM), Stéphanie Dord-Crouslé (Chargée de recherche CNRS, UMR IHRIM), Fatiha Idmhand (Professeur des Universités, UMR ITEM), Ioana Galleron (Professeur des Universités, UMR LATTICE)

Le réseau d'expertise scientifique et technique sur les corpus numériques d'auteurs REST CAHIER

1) Présentation du Réseau d'expertise scientifique et technique sur les corpus numériques d'auteurs CAHIER

De 2011 à 2021, CAHIER a développé, en tant que Consortium d'Huma-Num, une expertise poussée sur la numérisation des corpus d'auteurs, leur conversion en données numériques, leur édition et leur archivage. Pendant dix années, CAHIER a constitué un réseau national qui a atteint près de 250 personnes et réuni 65 projets scientifiques sur l'ensemble du territoire national. Spécialisé dans les corpus d'auteurs, le consortium a créé et partagé des connaissances et des expertises scientifiques et techniques sur le cycle de vie des données et sur plusieurs aspects couverts par l'infrastructure Huma-Num. Pour préserver cette expertise de terrain à la fois scientifique et technique, et au terme d'une décennie en tant que consortium d'Huma-Num, CAHIER propose de poursuivre ses travaux sous une nouvelle forme, celle de **"Réseau d'expertise scientifique et technique".** Le réseau d'expertise scientifique et technique REST CAHIER serait partie prenante de l'infrastructure Huma-Num; spécialisé dans les **corpus numériques d'auteurs**, il joue le rôle de pivot et de relai entre l'infrastructure technique et la communauté scientifique de terrain, et inversement.

Ce document présente les grands principes d'organisation et de fonctionnement du "Réseau d'expertise scientifique et technique CAHIER", il propose également un calendrier d'activités pour les deux premières années d'existence.

a) REST CAHIER : un réseau, quatre pôles

Inspiré par les centres "K" portés par l'infrastructure européenne CLARIN, **REST CAHIER** se propose de devenir un centre de connaissances dédié aux corpus d'auteurs. Après une décennie d'expérience en tant que Consortium d'Huma-Num, REST CAHIER a identifié quatre pôles d'expertise qui constitueront les principaux axes de travail du réseau : la **F**ormation, l'**A**nimation, l'**I**ntégration et la **R**echerche scientifique.

- **Pôle "Formation"**

En raison de leur évolution rapide, les outils et technologies des Humanités numériques nécessitent un retour constant vers la formation pratique. Pendant dix ans, CAHIER a réalisé cette veille technologique et l'a partagée à travers deux leviers principaux : d'un côté, un atelier annuel thématique d'une durée de quatre jours (de type *école de printemps*) dont la thématique était revue tous les ans de façon à répondre à l'actualité technologique et aux attentes des membres du consortium et, de l'autre, des formations brèves aux thématiques ciblées sur l'édition numérique, en particulier autour du codage avec XML-TEI (de type *école d'automne* comportant deux niveaux et organisée à Tours) et de son exploitation. Des formations à la demande, sur des techniques plus poussées au traitement automatique du texte, ont également été organisées (comme les formations à BaseX, XSLT, à Métopes ou à TXM). Le pôle "Formation" du REST CAHIER propose de poursuivre ces actions de formation sous la forme suivante :

Des formations brèves ciblées à la demande

Trois formations annuelles :

-"École d'automne à Tours" : Formation à l'encodage des textes en XML-TEI (niveaux 1 et 2)

-"École d'hiver EnExDi à Poitiers" : Formation des jeunes chercheurs et doctorants aux méthodes des

Humanités numériques
-"École de printemps" : Atelier annuel du REST CAHIER

❖ **Pôle "Animation"**

Pendant dix ans, le consortium CAHIER a assuré l'animation scientifique d'un réseau à la fois interprofessionnel, composé de chercheurs (universitaires et CNRS), de jeunes chercheurs et d'ingénieurs, et interdisciplinaire, associant non pas des institutions ou des champs disciplinaires mais des projets scientifiques soucieux de répondre à une question de recherche en croisant les méthodologies. Ce réseau a apporté un appui à la recherche et a contribué à l'élaboration de bonnes pratiques en Humanités numériques en constante interaction avec l'infrastructure Huma-Num. Le réseau REST CAHIER veut poursuivre le travail de courroie de transmission ascendante et descendante dans le cadre de son pôle "Animation". Pour cela, l'animation scientifique et technique sera assurée par des groupes de travail dont le but sera de répondre à un problème dans un temps limité ; les groupes de travail émanent de propositions faites par les membres du réseau. La solution produite par les groupes pourra prendre la forme d'un article de réflexion, de recommandations, d'un guide ou d'un outil.

Le pôle "Animation" prendra également en charge la communication du réseau, en constituant un groupe de contributeurs pour le carnet de recherches de REST CAHIER sur Hypothèses. Le groupe programmera des publications régulières et assurera la diffusion d'une lettre d'information deux fois par an. Pour le début de ses travaux, REST CAHIER propose de constituer les trois groupes de travail suivants :

Groupe de travail "TEI" : ce groupe aura pour but de poursuivre les travaux initiés dans le GT-TEI de CAHIER en vue d'acculturer et harmoniser les pratiques d'encodage en XML-TEI à l'échelle nationale ;

Groupe de travail "FAIRisation des données" : le groupe aura pour but de poursuivre, en lien avec Huma-Num, l'accompagnement de la communauté à la FAIRisation des données produites ;

Groupe de travail "Communication": le groupe aura pour but d'archiver l'ancien carnet de recherches et d'ouvrir le nouveau carnet du réseau, de l'organiser et de programmer des publications.

❖ **Pôle "Intégration"**

De 2012 à 2020, l'adhésion au consortium CAHIER était gratuite et ouverte à tous les porteurs de projets scientifiques. Une demande d'adhésion motivée était soumise à l'expertise du Comité de pilotage de CAHIER et la proposition était généralement acceptée. En contrepartie, le porteur de projet s'engageait à construire des données interopérables et à les mettre à la disposition de la communauté scientifique. Une fois membre, le projet avait accès aux formations et actions organisées par le Consortium.

Ce mode de fonctionnement sera reconduit dans le cadre de REST CAHIER, mais il sera modernisé en fonction des évolutions actuelles des Humanités numériques.

En effet, en dix ans, la notion d'interopérabilité a évolué et s'est vue progressivement supplantée par le concept de "données FAIR". Les exigences de science ouverte ont apporté de nouveaux outils (entrepôts de données) et de nouvelles exigences, puisque la FAIRisation des données fait dorénavant partie des éléments d'appréciation des projets financés par les agences. De ce point de vue, CAHIER a pu suivre l'évolution des services d'Huma-Num qui se sont transformés pour apporter aux SHS les outils qui leur permettraient de produire, traiter, stocker et préserver leurs données.

Le réseau REST CAHIER propose de poursuivre l'intégration et l'accompagnement de la communauté scientifique en revisitant son processus d'adhésion de la façon suivante : l'adhésion par projet continuera d'être la règle, de même que la motivation de l'adhésion via la remise d'un formulaire. Toutefois, une fois admis, le projet ne sera membre du réseau REST CAHIER que pour une durée limitée à trois ans, renouvelable plusieurs fois, mais soumise à conditions. En effet, en adhérant au réseau, les projets s'engageront fermement à respecter les "bonnes pratiques" et à produire des objets numériques ouverts et accessibles (FAIR) : le non-respect de cette condition constituera un motif de non renouvellement de l'adhésion. De même, pour que l'adhésion soit renouvelée, au moins un des membres du projet devra avoir participé durant les trois années écoulées aux travaux d'au moins un groupe de travail et à au moins une des formations organisées. Si ces conditions ne sont pas remplies, le renouvellement de l'adhésion ne sera pas possible. En intégrant le réseau REST CAHIER, les projets s'engagent donc à être actifs.

En contrepartie, l'intégration au réseau CAHIER permettra aux projets de bénéficier des conseils et expertises des activités du Pôle "Formation" (chaque projet bénéficiera d'une place prioritaire) qui réservera à ses membres quelques actions exclusives.

Les procédures précises et les conditions à remplir pour adhérer au REST CAHIER seront élaborées par un groupe de travail dédié durant le premier semestre de l'année civile 2022. **L'adhésion à CAHIER sera ouverte aux collègues porteurs de projets en France, en Europe et au-delà en vue d'une plus grande internationalisation du**

réseau. Les modalités financières et administratives de participation de ces projets étrangers seront définies dans les statuts du REST. La langue des travaux restera le français.

Le Pôle "Intégration" du REST-CAHIER sera constitué, en 2022 et 2023, du groupe de travail suivant :

Groupe de travail "Adhésion" : Ce groupe aura pour but d'expertiser les demandes d'adhésion au REST-CAHIER

❖ **Pôle "Recherche"**

En dix ans, le consortium CAHIER a produit de nombreux corpus ; en revanche, les livrables scientifiques comme les articles et les ouvrages ont été rares. Comme le colloque de clôture de CAHIER l'a montré, la communauté a besoin d'un espace de réflexion sur l'utilisation des corpus numériques, sur les nouvelles théories et les concepts que l'on peut fonder à partir de l'exploration des données issues d'éditions et d'archives numériques. C'est l'objectif principal du Pôle "Recherche" dont l'ouverture vers les collaborations internationales sera accentuée par le travail avec d'autres groupes tels que, par exemple, les "K centers" de Clarin et les "Working groups" de Dariah.

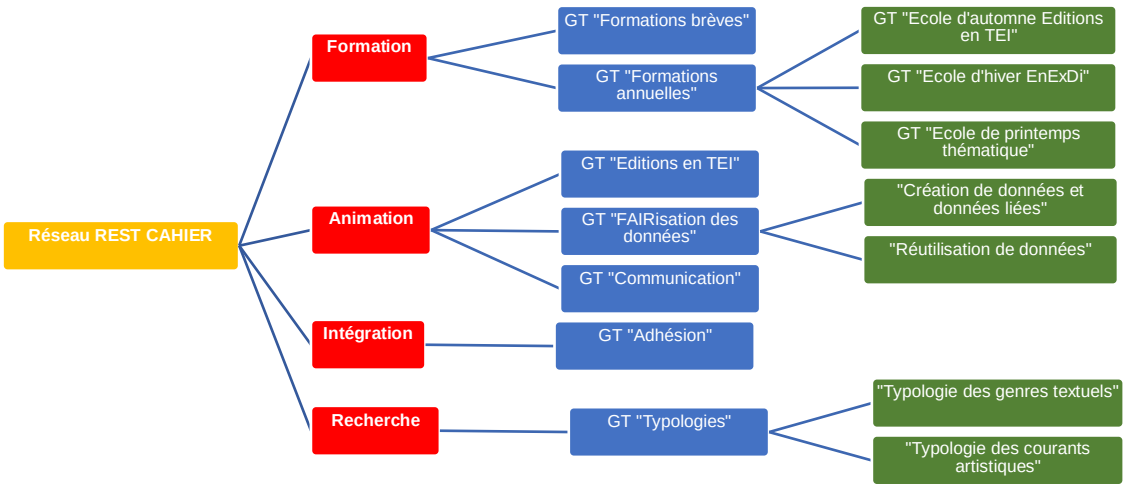
A ce stade, deux thématiques scientifiques ont été identifiées pour les groupes de recherche de 2022 :

Groupe de travail "Typologies" : Dans le sillage du GT Typologies textuelles du consortium CAHIER, ce groupe aura en charge non seulement le maintien et le développement du thésaurus déjà créé sur les "Typologies textuelles", mais également la mise en œuvre d'autres vocabulaires partagés pour les projets de corpus d'auteurs. Le groupe de travail dédié à la "Typologie des courants artistiques et littéraires" commencera ses travaux en janvier 2022.

Groupe de travail "Usages numériques" : Ce groupe aura pour but d'interroger les usages des objets numériques constitués par les communautés des sciences humaines, cette question, qui n'a pas été traitée par le consortium CAHIER, est essentielle au regard de la multiplication actuelle des objets et des livrables numériques.

Outre les deux groupes de travail susmentionnés, d'autres pourront être créés au fur et à mesure, en fonction des besoins des membres du réseau.

Pour résumer, le réseau REST CAHIER sera organisé, pour les deux premières années, de la façon suivante :



b) Alignement du réseau sur les missions et les objectifs de la TGIR Huma-Num

L'organisation proposée par le réseau REST CAHIER permettra de contribuer aux missions suivantes :

❖ **Mission de relai**

En constituant un réseau d'expertise autour des publications et éditions numériques de corpus d'auteurs avec un fort accent mis sur l'accès ouvert, l'interopérabilité et la réutilisabilité des données, le REST-CAHIER répercute des recommandations de la TGIR Huma-Num tout en jouant, à l'inverse, un rôle de relai des besoins et problèmes de la communauté en direction de l'infrastructure. En ce sens, le maintien du réseau initialement constitué par CAHIER contribue à la coordination des actions de valorisation, en augmentant la visibilité des données produites par les projets membres et en stimulant (grâce au GT "Usages", par exemple) les réflexions sur leur réutilisation dans le cadre

de travaux entrepris dans et hors du réseau.

❖ **Mission de mutualisation**

Les formations proposées par le REST-CAHIER, ainsi que le travail d'animation décrit plus haut, contribuent à mutualiser des compétences scientifiques et techniques sur la création de corpus d'auteurs entre les projets adhérents mais aussi au-delà. Des personnes ressources pourront être identifiées pour favoriser l'accès à des outils plus ou moins nouveaux (prise en main, usage expert), pour conseiller au sujet des méthodes les plus appropriées ou expertiser des propositions, etc.

❖ **Mission d'internationalisation**

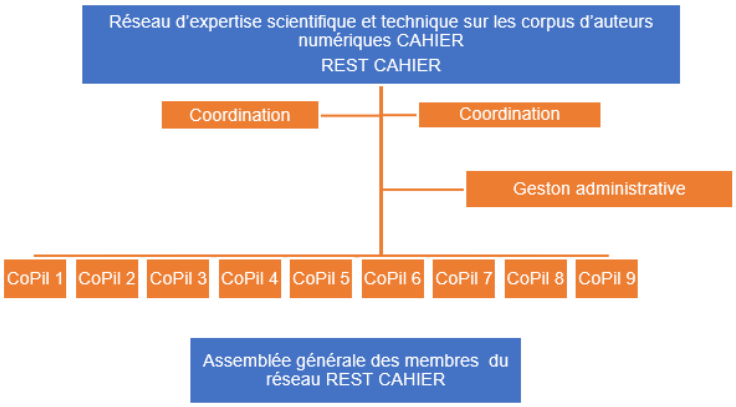
Pour mieux diffuser ses travaux à l'échelle européenne, le REST CAHIER pourra formuler, dans les prochaines années, une demande pour devenir un centre "K CLARIN" (https://www.clarin.eu/content/knowledge-centres#clarin-knowledge-centres).

2) Gouvernance du Réseau d'expertise scientifique et technique sur les corpus d'auteurs numériques CAHIER

Le réseau REST CAHIER est coordonné par un(e) coordonnateur(trice) et un(e) coordonnateur(trice) adjoint(e), élu(e)s.

L'Assemblée générale du réseau REST CAHIER est composée de tous les projets membres (sont membres tous les projets dont la demande d'adhésion a été validée), chaque projet dispose d'une voix pour élire les membres du Comité de Pilotage et les coordonnateurs.

En plus des deux coordonnateurs, le Comité de pilotage est constitué de neuf membres élus par l'Assemblée générale. Durant l'année de transition 2022 du "Consortium CAHIER d'Huma-Num" vers le "Réseau d'expertise scientifique et technique sur les corpus d'auteurs numériques CAHIER", et dans l'attente de la constitution de la nouvelle Assemblée générale des projets membres et de l'élection d'un nouveau Comité de pilotage, la gouvernance suivante est proposée :



Le tandem qui se propose de prendre en charge la coordination pour l'année de transition est formé d'Alexey Lavrentev (Ingénieur de recherche CNRS) et Stéphanie Dord-Crouslé (Chargée de recherche CNRS).

La gestion administrative du REST CAHIER continuerait d'être assurée par la MSH Val de Loire (*sous réserve de l'accord de son directeur*).

3) Calendrier de travail Réseau d'expertise scientifique et technique sur les corpus d'auteurs numériques CAHIER

❖ **Programme de travail Année 2022 (année de transition)**

Organisation des écoles thématiques saisonnières et création de groupes de travail temporaires :
Janvier à Mars 2022 GT "Statuts du REST" : Rédaction des statuts du réseau REST CAHIER
Janvier à Mars 2022 GT "Adhésion" : Rédaction du document d'adhésion et lancement du processus d'adhésion
Juin 2022 GT "Adhésion" : Réunion du GT en marge de l'atelier annuel
Juin 2022 : Atelier annuel thématique et lancement des travaux des groupes de travail "FAIRisation", "Communication", "Typologies"
Avril à Septembre 2022 : travaux du CoPil : Rédaction du projet détaillé de REST-CAHIER

❖ **Programme de travail Année 2023 (Première année du REST)**

Organisation des écoles thématiques saisonnières et lancement des travaux des groupes de travail :
GT "TEI"
GT "FAIRisation"
GT "Communication"
GT "Typologies"

❖ **Programme de travail Année 2024 (Deuxième année du REST)**

Organisation des écoles thématiques saisonnières, continuation des travaux des GT TEI, Communication, FAIR et lancement des nouveaux GT comme le GT "Usages"

4) Budget

Les dépenses et le budget annuel du réseau REST seraient répartis comme suit :

Exemple pour l'année 2023

Action	**Dépenses**	**Recettes**	**Organismes financeurs**
Formations			
Soutien "École d'automne" TEI Tours	6000	2000	**Huma-Num (via réseau REST CAHIER)**
		4000	Laboratoires et projets des participants (déplacement, logement)
Soutien "École d'hiver" EnExDi	6000	2000	**Huma-Num (via réseau REST CAHIER)**
		4000	Laboratoires et projets des participants (déplacement, logement)
Atelier annuel (École d'été)	12000	3000	**Huma-Num (via réseau REST CAHIER)**
		3000	ANF du CNRS
		6000	Laboratoires et projets des participants (déplacement, logement)
Rencontres GT			
Groupes de travail	6000	5000	**Huma-Num (via réseau REST CAHIER)**
		1000	Laboratoires et projets des participants (déplacement, logement)
Administration, gestion			
Gestion du consortium (MSH Val de Loire)	3000	3000	**Huma-Num (via réseau REST CAHIER)**
Total	**Dépenses annuelles** 33000	**Recettes annuelles** 33000	**Organismes** Huma-Num (via réseau REST CAHIER) 15000 Appel à projets ANF CNRS 3000 Laboratoires et projets des participants (déplacement, logement) 15000

Montant de la subvention qui serait demandée à Huma-Num pour 2023 = 15.000 euros

g) Annexe n°6 : Projet de consortium dédié aux outils « OLIO »

Consortium OLIO
Outils Libres Interopérables et Ouverts pour la recherche en SHS

8.9.21

Document de travail

Contexte

La pratique du numérique dans le domaine des Humanités s'est, comme dans toutes les disciplines SHS, largement développée ces dernières décennies. Grâce à des efforts soutenus institutionnellement (InSHS, TGIR Huma-Num, Collex Persée entre autres), la quantité de données numériques, structurées et utilisant des standards, explose. Le consortium Cahier, le consortium international TEI, les associations en Humanités numériques ont permis de disposer de guides de bonnes pratiques, de retours d'expériences, de tutoriels, etc. pour permettre la création de données de qualité, qui suivent les principes FAIR et ceux de la science ouverte.

Nous partons du constat que la communauté autour des corpus d'auteur s'est concentrée jusqu'à aujourd'hui sur les *données* et qu'un travail de même nature sur les outils et les pratiques est dorénavant nécessaire. Des outils pour l'édition, la description ou l'exploitation de corpus sont utilisés de longue date et des formations spécifiques se sont développées à destination de divers publics. Toutefois, il existe peu de passerelles entre ces outils et le versant "exploitation scientifique outillée" des données publiées mérite d'être développé, y compris dans ses présupposés et ses implications scientifiques.

En parallèle, la production de corpus d'auteur s'élargit avec l'apparition de nombreux acteurs dans les mondes académiques et culturels, comme en témoignent les nombreux projets de numérisation d'archives de bibliothèques ou les appels d'offres Collex. Cette masse inédite de contenus, qui répondent à des enjeux nouveaux de valorisation, de collaboration interprofessionnelle et plus largement de redéfinition de l'économie de la culture à l'heure du numérique, demande un outillage sur lequel l'expérience acquise au sein du consortium Cahier serait une vraie "plus value".

Un retour sur les usages et en amont sur les moyens, technologiques et méthodologiques, constitue une manière originale de penser les données suivant une perspective interprofessionnelle et interdisciplinaire. Comme l'ont montré les travaux relevant de la *distant reading*[1] (Moretti) , de la *machinic reading* (K. Heyles) ou des enjeux de la datavisualisation, il convient de considérer les données ou *data* plutôt que des « établis » (Latour) ou « *capta* » (Drucker).

Forte de son expérience d'une décennie de traitement de corpus, *la communauté constituée par* le consortium Cahier est désormais à même de proposer des scénarios de recherche, des parcours documentés avec des langages outils désormais bien connus, tout en offrant un espace d'innovation, du fait même de cette expérience diversifiée.

Le projet proposé par Olio en prolongation de Cahier réside dans un parti-pris : engager un questionnement critique collectif sur la production, la manipulation et l'analyse des données à la lumière des démarches individuelles menées jusqu'à présent sur des corpus singuliers. Dans quelle mesure les écarts observés dans les pratiques traduisent-ils des démarches scientifiques différentes. Comment les choix techniques contribuent à façonner l'objet scientifique étudié ?

---

1. Moretti, Franco. *Distant reading*. London/New York, Verso, 2013 (https://www.sudoc.fr/177536683).

Quels sont les impacts méthodologiques distincts entre l'exploitation de données versées dans Omeka et de données structurées en XML/TEI ? Les développements ad-hoc s'opposent-ils aux chaînes de productions génériques ? Quelles contraintes institutionnelles pèsent sur le choix des uns ou des autres ? Quelles implications entre la transcription avec outils OCR/HTR et la transcription manuelle ? Faut-il fixer dès le début du projet la nature du résultat : une édition critique spécifique, une bibliothèque numérique, un corpus sémantique, etc. ?? Le travail collectif sur ces étapes préalables à la réalisation des projets scientifiques nous semble constituer un terrain de recherche à proprement parler qui ne peut pas se réduire pas à des préliminaires techniques.

Car la question des outils, au cœur du projet Olio, n'est pas seulement celle de la technicité de la recherche ou de l'appareillage des projets : elle soulève des enjeux épistémologiques fondamentaux. En humanités numériques, les outils sont des dispositifs intellectuels autant que des instruments techniques – ils sont d'ailleurs eux-mêmes les produits d'une recherche et d'une expérimentation. On peut ainsi envisager l'outil à la fois comme le résultat et comme l'origine potentielle de questions de recherche inédites, ou de nouvelles méthodologies et approches.

Cette proposition est née au sein du groupe de travail « Réutilisabilité » de CAHIER. Les travaux de ce dernier ont mis en exergue la nécessaire mise en relation des corpus existants, au-delà des projets de recherche singuliers. Ces liens entre les corpus apportent de nouvelles problématiques scientifiques et mettent en lumière le besoin d'appropriation, voire de création, de nouveaux outils.

Olio propose d'engager un travail collectif, rassemblant chercheurs, ingénieurs, éditeurs numériques, documentalistes, et plaçant au cœur de la problématique des outils celle des usages – usages possibles, induits, contraints ou détournés, dans une tension entre normes et bonnes pratiques d'une part, inventivité et agilité de la recherche de l'autre.

La variété, tout comme les limites structurelles des outils, qui constituent des verrous autant que des occasions de retours réflexifs sur nos attentes et nos présupposés, rendent cette réflexion nécessaire. Certains outils sont propriétaires et payants d'autres sont libres et bénéficient d'une large communauté de développeurs. Il existe aussi un nombre impressionnant de "petits développements", d'utilisations confidentielles d'outils que l'on découvre au détour de colloques, journées d'étude ou autres. Assister à une conférence telle que DH peut donner le tournis tant la variété est importante. Bien sûr, il n'existe pas d'outil universel qui ferait tout, possédant toutes les options et fonctionnalités : chaque projet et chaque équipe a ses propres besoins, ses propres compétences. Il est souvent nécessaire de combiner plusieurs outils en fonction des besoins et des objectifs. Là où l'outil était très performant pour acquérir des données, il n'est plus adapté pour leur diffusion ; là où le logiciel a permis la constitution d'un corpus accompagné de métadonnées riches, il n'est plus capable d'accompagner les analyses dont le projet de recherche a besoin, etc.

Passer d'un outil à l'autre n'est pas toujours aisé. Si nos données sont stockées dans une base de données, et qu'on veuille les intégrer à un entrepôt, il faudra bien souvent passer par plusieurs étapes, parfois très techniques. Les logiciels ne disposent pas toujours d'API compatibles et l'usage de cette technique n'est pas trivial. Un corpus en XML-TEI dont on souhaiterait faire une annotation complexe n'est pas immédiatement manipulable par les outils issus d'autres disciplines car les formats diffèrent. Les exemples d'interopérabilité difficile ne manquent pas, malgré l'ouverture des corpus avec des formats standards et ouverts. Or, dans l'idéal, ne rêve-t-

on pas de créer nos données avec un logiciel parce qu'il est efficace pour nos objectifs scientifiques, mais de pouvoir les transformer sur un autre, les diffuser sur un troisième, etc. ?

Corpus d'auteur

Le périmètre de notre projet est celui qui avait été défini par le consortium Cahier : « les corpus d'auteurs et plus généralement les corpus textuels constitués en référence à l'œuvre d'un auteur, d'une tradition éditoriale, d'une forme littéraire, d'un genre ». Ce périmètre permet de prendre en compte différentes disciplines et de travailler de façon transversale (trans-séculaire, trans-disciplinaire) sur des sources de différentes natures, selon des objectifs variés : publication, édition brute ou éditorialisation, archivage, exploitation, valorisation. Il semble également nécessaire d'aménager une place spécifique aux bases de données construites lors de l'exploitation des corpus. Celles-ci permettent bien souvent d'établir des connexions entre les projets.

Objectifs d'OLIO

L'objet principal du consortium que nous proposons est ainsi l'*outil pour la recherche* : qu'il soit logiciel, site web, plateforme, entrepôt, module/plugin, service..., qu'il soit outil d'analyse, de constitution, de visualisation, de publication, d'archivage... qu'il soit spécialisé ou générique, ouvert ou propriétaire (en privilégiant tout de même les logiciels), etc. L'objectif est, en partant des pratiques actuelles, d'accompagner des démarches de signalement et d'expérimentation d'outils et de méthodologies, de réfléchir aux connexions possibles, de favoriser l'interopérabilité entre les outils et de contribuer à clarifier la nature des tournants qu'ils engagent.

Nous souhaitons intégrer autant que possible non seulement des usagers des outils, mais aussi des usagers des objets numériques qu'on fabrique grâce à ces outils. Nous imaginons cet idéal où, par exemple, un lot d'images entreposé sur Nakala, pourrait - *par le jeu des API et sans transfert des fichiers* - bénéficier d'une transcription au kilomètre par un outil HTR, passer *en un clic* par Grobid pour être structuré, puis *importé* directement sur un outil d'édition pour corriger et enrichir la transcription. Le parcours des transcriptions pourrait se poursuivre : intégration à EVT, TEI Publisher, Omeka ou autre ; dépôt sur l'entrepôt de données à côté des images initiales ; export dans des outils de TAL pour faire des analyses linguistiques, etc., etc.

L'objectif du consortium OLIO est de permettre de doter les projets d'outils adaptés aux objectifs de recherche, de travailler sur l'impact des outils sur la recherche produite et d'étudier les méthodes de recherche induites par ces outils. Cela implique la nécessité de connaître les outils (il ne s'agit pas seulement de former à leur utilisation, mais aussi d'expliciter les présupposés qui président à leur création et à leur développement), et, de là, de réfléchir collectivement aux possibilités d'utilisation et aux nouvelles formes de savoir qui peuvent en découler.

Car, à côté d'éditions (numériques ou pas) plus classiques, de plus en plus de projets créent ou utilisent des outils d'analyse, de visualisation et de fouille des données. Ces lieux d'exploration sont parfois déjà désignés sous les termes de « laboratoires numériques » ou « laboratoires de textes ». Ces objets encore mal définis, qui associent des corpus et de nouveaux outils, ouvrent de nouvelles voies/perspectives d'analyse et de recherche sur les corpus de textes.

Enfin, l'approche par l'outil et ses usages a l'avantage d'être structurellement interdisciplinaire. La plupart des outils sont employés très largement, par des communautés très diverses. L'analyse de leurs appropriations dans les différents champs disciplinaires nourrira ainsi une approche réflexive sur les Humanités (au-delà des seules Humanités numériques), en éclairant les points de rencontres, les lignes de forces et les divergences épistémologiques actuelles, sans préjuger de structurations a priori, dans la continuité des réflexions de Johanna Drucker notamment.

Un objectif complémentaire mais essentiel sera celui de la formation aux différents outils. Tout comme le consortium Cahier, dont la réussite était basée sur des ateliers réguliers et fédérant une communauté, le travail du consortium serait animé par l'organisation régulière de retours d'expériences et de formations du niveau débutant au niveau expert. Des ateliers de découverte et de travail sur corpus prototypiques permettront de rassembler des spécialistes en corpus d'auteurs mais aussi les *nouveaux entrants*, en particulier les doctorants.

Il faut en effet donner une place aux *usagers* des corpus d'auteurs : lecteurs, public, chercheurs, enseignants qui utilisent ces éditions. Ainsi on fera appel à une communauté travaillant sur tel corpus d'auteur ou tels types de corpus pour avoir des retours sur les utilisations des objets numériques ainsi produits, et sur les besoins réels des éditeurs et des usagers.

Communauté OLIO

Le consortium veut fédérer celles et ceux qui ont besoin d'utiliser des outils pour les corpus d'auteur et celles et ceux qui les créent. Il ne s'agit pas de rassembler des structures mais des projets (et l'ensemble de leurs membres) ayant comme objet un corpus d'auteur ou un outil s'adaptant à celui-ci, personnes venant des institutions de la recherche en humanités (laboratoires, MSH, écoles doctorales) mais aussi du monde des bibliothèques ou de la culture, grands pourvoyeurs ou utilisateurs de corpus d'auteur (associations, bibliothèques, musées).

Le consortium peut aussi devenir un interlocuteur expert pour les autres opérateurs de numérisation et d'édition de corpus d'auteurs (Collex, campus d'excellence, EUR, etc.).

Loin de polariser les membres du consortium en opposant pourvoyeurs et consommateurs d'outils, il s'agit d'accompagner un dialogue structurellement nécessaire et fécond. L'engagement des développeurs dans une démarche « bottom-up », qui part du besoin exprimé, est naturel et reste pertinent ; mais, bien au-delà, le consortium se propose de devenir un lieu de réflexion sur le rôle et la place de l'outil dans les Humanités (au sens large). La manière dont l'instrumentation numérique de la recherche influence les Humanités doit faire l'objet d'une réflexion collective qui est encore loin d'être épuisée aujourd'hui. Le projet de consortium OLIO souhaite s'emparer de ces questions à travers le prisme des outils, aussi bien pour envisager leurs potentialités en termes problématique scientifique que pour questionner leur apport.

Feuille de route

- Améliorer l'information sur les outils et leur offrir une visibilité à travers leur signalement ;
- Créer le cadre de réflexion et d'élaboration de chaînes de traitement à travers des retours d'expérience, des procédures spécifiques ou développées en commun, la rédaction de

vade-mecum. Par exemple : comment favoriser l'interopérabilité de ses données en cours de projet ? Comment anticiper et élaborer une chaîne de traitement sans pour autant interdire un changement d'option, en conservant de la souplesse ?

- Signaler ou développer les scripts, procédures et outils permettant le passage d'un format à l'autre, d'un environnement à l'autre ;
- Faciliter l'appropriation des outils par la communauté à travers la formation, des débats contradictoires basées sur de larges retours d'expérience et encourager les débats méthodologiques ;
- Nourrir la réflexion théorique et épistémologique : quel impact des outils sur la recherche ? quels principes, quels concepts, qu'ils soient explicites ou qu'ils relèvent de l'impensé, entrent en jeu dans la conception, le fonctionnement et l'utilisation des outils ?
- Regrouper et synthétiser l'actualité des outils pour les corpus d'auteur à travers un état de l'art annuel qui sera présenté lors de chaque AG annuelle et diffusé ensuite au sein de la communauté. Cet état de l'art annuel présentera les actualités autour des outils (nouvel outil, nouvelle version), les formations et les événements notables de l'année écoulée ; il pourra faire des focus sur des thématiques émergentes ou qui suscitent les controverses. Il n'aura pas vocation à être prescriptif mais il donnera toutes les ressources et tous les éléments de débats. Il sera conclu par un glossaire avec des définitions des notions faisant l'actualité.

Le consortium OLIO utilisera pour cela différents outils :

- Carnet de recherche avec liens vers différents réseaux sociaux : il comportera l'actualité des événements organisés par le consortium, les comptes rendus des formations, des retours d'expériences, lieu de publication du consortium et de ses instances ;
- Agenda du consortium (avec module d'inscriptions et d'évaluation des formations organisées) et agenda sur l'actualité des outils pour les corpus d'auteur ;
- Espace collaboratif interne pour l'élaboration des signalements, recommandations à diffuser ;
- Documents (vade-mecum, état de l'art annuel) publiés sur supports numérique et papier.

Des outils

Une première liste d'outils ou de références :

- Pour (chaînes de) traitement des données : Metope, Textable, Dataiku, OpenRefine, TXM...
- Pour visualisation des données : Tableau, Gephi, ElasticSearch&Kibana...
- Pour travail sur les images : API IIIF, Mirador...
- Lemmatiseur et TAL : Pyrrha ; CLTK et NLTK (Python)...
- Fouille et analyse : Textable, iramuteq, TXM, hyperbase, voyant tools, Philologic ...
- Logiciels de reconnaissance de caractères (OCR, HTR) : EScriptorium, Tesseract (ocrmypdf), Kraken, Trankribus...
- Editeur ou outil XML/TEI : Oxygen, TEIPublisher...
- Plateformes de transcription : TACT, Transcrire, Transcript/EMAN...
- Bibliothèque numérique : Omeka Classic & S, Heurist...
- Description archivistique : atom…

Fonctionnement & statuts

Le consortium sera organisé en ateliers et groupes de travail (GT). Une assemblée générale annuelle du consortium fixera la thématique et l'agenda des ateliers et des groupes de travail (GT) qui présenteront leurs travaux à chaque AG.

Cette assemblée générale annuelle élira un comité de pilotage qui sera chargé de suivre l'avancement de ces travaux et une équipe de coordination tricéphale qui sera chargée de les mettre en œuvre.

Les statuts et le mode d'adhésion seront débattus lors de la première AG du consortium.

Gouvernance

L'équipe de coordination sera tricéphale : un/une coordinateur/trice principal(e) et deux adjoint(e)s qui devront être représentatifs des différents statuts et expériences rassemblés dans le consortium : producteur / utilisateur d'outil - chercheur, enseignant chercheur / ingénieur / bibliothécaire. Une variété disciplinaire, institutionnelle et géographique sera aussi nécessaire.

Structuration

La première Assemblée générale fixera le fonctionnement du consortium, les statuts, les modalités de participation à OLIO et d'élections de ses instances. L'Assemblée générale sera ensuite l'organe central du fonctionnement du consortium.

Première assemblée générale sur une journée complète :

- présentation du consortium et de ses objectifs,
- propositions d'organisation et de statuts du consortium (structure, mandats, adhésions, etc.),
- proposition à la volée de GT suivie de séance d'échanges en sous-groupe pour la définition de ceux-ci,
- création des premiers GT dont 1 personne coordinatrice et 1 responsable communication.

Fonctionnement régulier prévu

Au cours de l'année :

- les GT organisent régulièrement des séances ;
- l'équipe de coordination veille à ce que les GT soient actifs (relance, accompagnement, aide, si nécessaire), se tient au courant des avancées et s'assure que les GT ont des périmètres bien définis (pas de doublon) ;
- une équipe de communication (une personne responsable de la communication par GT et une au sein du trio de coordination) veille à ce que le carnet de recherche informe régulièrement la communauté des travaux du consortium ;

- préparation d'un atelier annuel de formation porté par un ou plusieurs GT, la thématique est fixée en AG;
- l'équipe de coordination accompagne la production de manuel, documentation ou tout autre support utile ; elle met en place le recueil d'informations pour l'état de l'art annuel.

Milieu d'année:

- atelier annuel :
- workshop de rédaction de l'état de l'art annuel (parties thématiques). Les deux événements peuvent être reliés.

AG de fin d'année

- retour des GT : difficultés/facilités, décision de poursuivre, arrêter, se transformer ;
- discussion sur la publication des livrables produits par les GT (manuels, vade-mecum, outils, etc.) : ils sont soumis à tous *avant* diffusion et leur diffusion est votée.
- retour sur l'atelier annuel et programmation du suivant ;
- désignation dans chaque groupe d'une personne responsable FAIR et d'une personne responsable de la communication ;
- discussion et finalisation de l'état de l'art annuel pour diffusion.

En début d'année X+1

- avec les responsables FAIR de chaque groupe, fairisation de tous les livrables du consortium : ceux de l'atelier (programme, supports de formation, comptes-rendus...), ceux éventuels des GT (documents, données test, exemples type, outils...), l'état de l'art annuel.

_ _ _ _ _ _ _ _ _

Personnes ayant contribué et/ou déclaré leur intérêt pour cette proposition :

- **EC/C :** Florian Barrière, Sarah Orsini, Véronique Magri, Céline Bohnert, Geoffrey Williams, Olivier Bruneau, Bénédicte Grailles, Brigitte Ouvry-Vial, Laurent Rollet, Martine Paindorge, Bénédicte Obitz, Marie-Thérèse Cam, Magalie Moysan, Marc Douguet, Frédéric Soulu, Isabelle Cogitore, Eric Francalanza, Anne Réach-Ngô
- **BIATS/ITA :** Elisabeth Greslou, Richard Walter, Anne Garcia Fernandez, Camille Desiles, Pierre Willaime, Gwenaëlle Patat, Julie Aucagne, François Vignale, Aurélie Hess, Francine Filoche, Vincent Maillard, Arnaud Bey
- **(Post-)doctorant.e.s :** Nicolas Lasolle, Elena Prat, Clarisse Evrard, Emmylou Haffner

Pour vos questions n'hésitez pas à écrire à : olio@groupes.renater.fr

Consortium OLIO - Outils Libres Interopérables et Ouverts pour la recherche en SHS

Proposition de statuts

Le Consortium

Le Consortium OLIO est constitué de projets de recherche partenaires dont la vocation est de travailler sur des corpus d'auteurs et plus généralement sur des corpus textuels constitués en référence à l'œuvre d'un auteur, d'une tradition éditoriale, d'une forme littéraire, d'un genre. Ces textes peuvent être de statuts différents (travaux, brouillons, œuvres, correspondances, inscriptions épigraphiques... papiers officiels et plus officieux) issus ou liés à un ou plusieurs auteurs, qu'il soit écrivain, philosophe, érudit ou scientifique. Les projets peuvent donc concerner un seul auteur comme une bibliothèque numérique avec de nombreux auteurs, le tout est d'avoir des sources primaires.

Ce périmètre permet de prendre en compte différentes disciplines et de travailler de façon transversale (trans-séculaire, trans-disciplinaire) sur des sources de différentes natures et selon des objectifs variés : publication, édition brute ou éditorialisation, archivage, exploitation, valorisation.

Les corpus devront être accessible en ligne et respecter les principes FAIR.

1. Gouvernance

1.1 Coordination

1.1.1 Rôle

a) La Coordination assure la gestion courante du Consortium, en particulier elle suit les travaux des différents groupes de travail, l'organisation d'un atelier annuel et la publication d'un état de l'art annuel ;
b) Elle met en application la politique scientifique et budgétaire définie par l'Assemblée générale annuelle et le Comité de pilotage ;
c) Elle est l'intermédiaire entre les instances scientifiques de tutelle et le Comité de pilotage qu'elle instruit de toutes ses actions ;
d) Elle est l'intermédiaire entre le Conseil Scientifique et le Comité de Pilotage.

1.1.2 Élection

a) La Coordination est élue par l'assemblée générale pour la durée du contrat de labellisation ; elle est constituée d'un.e coordinateur.trice principal.e et deux adjoint(e)s qui devront être représentatifs des différents statuts et expériences rassemblés dans le consortium : producteur / utilisateur d'outil - chercheur, enseignant chercheur / ingénieur / bibliothécaire. Une variété disciplinaire, institutionnelle et géographique sera aussi nécessaire.
b) Après appel, les candidatures aux trois fonctions, accompagnées d'une profession de foi, sont publiées sur le site du Consortium et diffusées sur la liste dédiée.
c) L'élection a lieu le jour de l'Assemblée générale. Elle se déroule pour chaque fonction au scrutin uninominal à deux tours.
d) Le Comité de pilotage peut, par une motion soutenue par au moins un tiers de ses membres, demander le renouvellement de la Coordination.

1.1.3 Soutien administratif

La coordination du consortium devra s'adjoindre pour la durée de son mandat, d'un soutien administratif.
La coordination sera soutenue par un.e responsable administratif.ve à temps partiel dont la quotité de temps de travail dévolue au Consortium sera compensée annuellement à son employeur principal par le Consortium.

1.2 Comité de pilotage

1.2.1 Rôle

a) Le Comité de pilotage (CoPil) définit les orientations scientifiques et budgétaires du consortium, en respectant les décisions de l'assemblée générale.
b) Il se réunit au moins 3 fois par an en CoPil ordinaire.
c) Il peut se réunir à la demande d'au moins un tiers de ses membres en CoPil extraordinaire.
d) Le CoPil étudie les dossiers des projets candidats à l'adhésion au Consortium et statue par vote sur chaque demande.
e) Les membres du CoPil expertisent les dossiers, remplissent un rapport d'expertise et, le cas échéant, font appel à des experts extérieurs.
f) Le CoPil élabore avec la Coordination les thématiques de l'atelier annuel et de l'état de l'art annuel du Consortium.

// Un article spécifique doit être rédigé pour définir les conditions d'adhésion et de départ des projets membres

1.2.2 Constitution

Le Comité de pilotage est constitué pour la durée du contrat de labellisation :
g) de la coordination ;
h) de dix (10) membres élus au scrutin nominatif à un tour par l'Assemblée générale ;
i) d'un ou plusieurs membres honoraires désignés à la majorité, pour la durée du mandat du CoPil, en raison de leur expertise, dans la limite d'un cinquième du nombre de membres élus.

1.2.3 Élection

(a) Après appel, les candidatures sont publiées sur le site du Consortium et diffusées sur la liste dédiée.
(b) L'élection se déroule au scrutin nominal à un tour lors de l'Assemblée générale.
(c) Chaque électeur vote pour autant de candidats qu'il y a de sièges à pourvoir (10 au maximum).
(d) En cas d'égalité du nombre de voix, les candidats sont départagés par la coordination en favorisant la représentativité du Comité (sexes, métiers, régions).

1.3 Assemblée générale

1.3.1 Rôle

a) L'Assemblée générale ordinaire est convoquée annuellement afin d'être informée et de valider le bilan de l'année précédente et le projet de l'année suivante présentés par la coordination ;
b) Elle décide des thématiques et du fonctionnement des groupes de travail, de l'atelier annuel et de l'état de l'art annuel.
c) L'Assemblée générale élit la coordination
d) L'Assemblée générale élit les 10 membres élus du Comité de pilotage.

e) L'Assemblée générale vote les statuts initiaux et leurs éventuelles modifications sur proposition du COPIL.
f) L'Assemblée générale est ouverte au public.

1.3.2 Constitution

a) L'Assemblée générale est constituée par l'ensemble des projets membres du Consortium.
b) La liste des votants est établie conformément au point 2 « Éligibilité et liste électorale ».
c) Seuls les votants peuvent être porteurs de procurations, à concurrence de deux procurations par personne.

1.4 Conseil scientifique indépendant

1.4.1 Rôle

a) Le Conseil scientifique analyse chaque année le rapport remis par la coordination du Consortium à sa tutelle la TGIR Huma-Num
b) Il émet un avis scientifique sur l'activité de l'année écoulée et sur le projet de l'année à venir lors de l'Assemblée générale ordinaire. Cet avis est intégré au compte-rendu de l'Assemblée générale.

1.4.2 Constitution

Le Conseil scientifique est constitué de 2 à 4 experts internationaux, désignés par le Comité de pilotage en raison de leurs compétences et de leur expertise dans les domaines d'activité du Consortium.

2 Eligibilité et liste électorale

2.1 Eligibilité

Est éligible à la coordination et au comité de pilotage tout membre d'un projet adhérent du consortium.

2.2 Votants

a) « Chaque projet partenaire dispose de deux voix maximum.
b) Les représentants du projet, dont les noms sont indiqués sur la fiche descriptive du projet partenaire, disposent du droit de vote.
c) Les représentants du projet peuvent céder leur droit de vote à l'un des membres du projet. Dans ce cas, ils indiquent par écrit à la coordination le(s) nom(s) de la ou des personnes disposant du droit de vote.

Fonctionnement Groupe de travail dans statuts ?

Consortium OLIO - Outils Libres Interopérables et Ouverts pour la recherche en SHS

Proposition de statuts

Le Consortium

Le Consortium OLIO est constitué de projets de recherche partenaires dont la vocation est de travailler sur des corpus d'auteurs et plus généralement sur des corpus textuels constitués en référence à l'œuvre d'un auteur, d'une tradition éditoriale, d'une forme littéraire, d'un genre. Ces textes peuvent être de statuts différents (travaux, brouillons, œuvres, correspondances, inscriptions épigraphiques... papiers officiels et plus officieux) issus ou liés à un ou plusieurs auteurs, qu'il soit écrivain, philosophe, érudit ou scientifique. Les projets peuvent donc concerner un seul auteur comme une bibliothèque numérique avec de nombreux auteurs, le tout est d'avoir des sources primaires.

Ce périmètre permet de prendre en compte différentes disciplines et de travailler de façon transversale (trans-séculaire, trans-disciplinaire) sur des sources de différentes natures et selon des objectifs variés : publication, édition brute ou éditorialisation, archivage, exploitation, valorisation.

Les corpus devront être accessible en ligne et respecter les principes FAIR.

1. Gouvernance

1.1 Coordination

1.1.1 Rôle

a) La Coordination assure la gestion courante du Consortium, en particulier elle suit les travaux des différents groupes de travail, l'organisation d'un atelier annuel et la publication d'un état de l'art annuel ;
b) Elle met en application la politique scientifique et budgétaire définie par l'Assemblée générale annuelle et le Comité de pilotage ;
c) Elle est l'intermédiaire entre les instances scientifiques de tutelle et le Comité de pilotage qu'elle instruit de toutes ses actions ;
d) Elle est l'intermédiaire entre le Conseil Scientifique et le Comité de Pilotage.

1.1.2 Élection

a) La Coordination est élue par l'assemblée générale pour la durée du contrat de labellisation ; elle est constituée d'un.e coordinateur.trice principal.e et deux adjoint(e)s qui devront être représentatifs des différents statuts et expériences rassemblés dans le consortium : producteur / utilisateur d'outil - chercheur, enseignant chercheur / ingénieur / bibliothécaire. Une variété disciplinaire, institutionnelle et géographique sera aussi nécessaire.
b) Après appel, les candidatures aux trois fonctions, accompagnées d'une profession de foi, sont publiées sur le site du Consortium et diffusées sur la liste dédiée.
c) L'élection a lieu le jour de l'Assemblée générale. Elle se déroule pour chaque fonction au scrutin uninominal à deux tours.
d) Le Comité de pilotage peut, par une motion soutenue par au moins un tiers de ses membres, demander le renouvellement de la Coordination.

1.1.3 Soutien administratif

La coordination du consortium devra s'adjoindre pour la durée de son mandat, d'un soutien administratif.
La coordination sera soutenue par un.e responsable administratif.ve à temps partiel dont la quotité de temps de travail dévolue au Consortium sera compensée annuellement à son employeur principal par le Consortium.

1.2 Comité de pilotage

1.2.1 Rôle

a) Le Comité de pilotage (CoPil) définit les orientations scientifiques et budgétaires du consortium, en respectant les décisions de l'assemblée générale.
b) Il se réunit au moins 3 fois par an en CoPil ordinaire.
c) Il peut se réunir à la demande d'au moins un tiers de ses membres en CoPil extraordinaire.
d) Le CoPil étudie les dossiers des projets candidats à l'adhésion au Consortium et statue par vote sur chaque demande.
e) Les membres du CoPil expertisent les dossiers, remplissent un rapport d'expertise et, le cas échéant, font appel à des experts extérieurs.
f) Le CoPil élabore avec la Coordination les thématiques de l'atelier annuel et de l'état de l'art annuel du Consortium.

// Un article spécifique doit être rédigé pour définir les conditions d'adhésion et de départ des projets membres

1.2.2 Constitution

Le Comité de pilotage est constitué pour la durée du contrat de labellisation :
g) de la coordination ;
h) de dix (10) membres élus au scrutin nominatif à un tour par l'Assemblée générale ;
i) d'un ou plusieurs membres honoraires désignés à la majorité, pour la durée du mandat du CoPil, en raison de leur expertise, dans la limite d'un cinquième du nombre de membres élus.

1.2.3 Élection

(a) Après appel, les candidatures sont publiées sur le site du Consortium et diffusées sur la liste dédiée.
(b) L'élection se déroule au scrutin nominal à un tour lors de l'Assemblée générale.
(c) Chaque électeur vote pour autant de candidats qu'il y a de sièges à pourvoir (10 au maximum).
(d) En cas d'égalité du nombre de voix, les candidats sont départagés par la coordination en favorisant la représentativité du Comité (sexes, métiers, régions).

1.3 Assemblée générale

1.3.1 Rôle

a) L'Assemblée générale ordinaire est convoquée annuellement afin d'être informée et de valider le bilan de l'année précédente et le projet de l'année suivante présentés par la coordination ;
b) Elle décide des thématiques et du fonctionnement des groupes de travail, de l'atelier annuel et de l'état de l'art annuel.
c) L'Assemblée générale élit la coordination
d) L'Assemblée générale élit les 10 membres élus du Comité de pilotage.

e) L'Assemblée générale vote les statuts initiaux et leurs éventuelles modifications sur proposition du COPIL.
f) L'Assemblée générale est ouverte au public.

1.3.2 Constitution

a) L'Assemblée générale est constituée par l'ensemble des projets membres du Consortium.
b) La liste des votants est établie conformément au point 2 « Éligibilité et liste électorale ».
c) Seuls les votants peuvent être porteurs de procurations, à concurrence de deux procurations par personne.

1.4 Conseil scientifique indépendant

1.4.1 Rôle

a) Le Conseil scientifique analyse chaque année le rapport remis par la coordination du Consortium à sa tutelle la TGIR Huma-Num
b) Il émet un avis scientifique sur l'activité de l'année écoulée et sur le projet de l'année à venir lors de l'Assemblée générale ordinaire. Cet avis est intégré au compte-rendu de l'Assemblée générale.

1.4.2 Constitution

Le Conseil scientifique est constitué de 2 à 4 experts internationaux, désignés par le Comité de pilotage en raison de leurs compétences et de leur expertise dans les domaines d'activité du Consortium.

2 Eligibilité et liste électorale

2.1 Eligibilité

Est éligible à la coordination et au comité de pilotage tout membre d'un projet adhérent du consortium.

2.2 Votants

a) « Chaque projet partenaire dispose de deux voix maximum.
b) Les représentants du projet, dont les noms sont indiqués sur la fiche descriptive du projet partenaire, disposent du droit de vote.
c) Les représentants du projet peuvent céder leur droit de vote à l'un des membres du projet. Dans ce cas, ils indiquent par écrit à la coordination le(s) nom(s) de la ou des personnes disposant du droit de vote.

Fonctionnement Groupe de travail dans statuts ?

h) Annexe n°7 : Préfiguration du consortium « CAHIER après CAHIER »

Préfiguration du nouveau consortium

CAHIER après CAHIER

Vers un nouveau Consortium (2023-2027)

Introduction, contexte

Après dix ans d'existence, le Consortium CAHIER a construit un large réseau national, expert dans la constitution, l'exploitation et la pérennisation des données des corpus d'auteurs. Actif et pluridisciplinaire, CAHIER souhaite préserver le réseau d'échanges de pratiques, de données, d'outils et de résultats qu'il a tissé au fil des ans mais il est conscient qu'il est nécessaire de s'ouvrir à de nouvelles perspectives de recherche à travers une communauté renouvelée fédérée au sein d'un nouveau projet scientifique.

Ces dix dernières années, plusieurs consortiums d'Huma-Num se sont intéressés à la création de corpus et ont développé des expertises poussées sur la numérisation, la conversion en données numériques exploitables des écrits et sur les formats et standards qui facilitent l'interopérabilité et la réutilisabilité des informations. Structurés en communautés scientifiques expertes des corpus littéraires, linguistiques, thématiques et historiques, ils ont abordé nombre de problèmes qui traversent les Humanités numériques et ont tenté d'apporter des réponses aux questions qui concernent les types de corpus constitués pour la recherche (description typologique des œuvres), la portée de certains formats d'échange (tels que le XML-TEI) ou les droits applicables aux publications numériques des sources (droits des auteurs, des éditeurs auteurs d'annotations numériques, etc.). Ils ont également contribué à augmenter la quantité de données et de corpus disponibles, les ont FAIRisés de façon à faciliter leur intégration dans de nouvelles collections numériques. En contribuant de cette façon à la constitution de nouveaux corpus plus larges, mais également plus diversifiés et plus hétérogènes, les corpus ainsi mis en données ouvrent à de nouveaux défis à la communauté scientifique.

Cependant, tout en constituant une chance, le numérique pose encore plusieurs défis aux sciences humaines et pour la communauté CAHIER, répondre aux enjeux de demain comme les objectifs d'automatisation ou d'application de méthodes poussées de l'informatique nécessite de disposer de corpus « trouvables » et, qui plus est, de bonne qualité. La crainte des porteurs de ce projet, à l'heure où de nombreuses propositions ambitionnent d'utiliser les méthodes dites « d'intelligence artificielle », est que la création de corpus, condition *sine qua non* de tels projets, ne soit oubliée alors même que la reprise de données et le nettoyage de données collectées à la volée sont des tâches coûteuses. Le prochain projet de CAHIER souhaite poursuivre la création de corpus de qualité selon de nouvelles conditions tout en interrogeant les notions de « CORPUS NUMÉRIQUES » et leurs nouvelles modalités, ainsi que celles d'ÉCRITS et de TRACES.

CAHIER propose de traiter ces questions au sein d'une communauté scientifique renouvelée car relever ce défi nécessite l'ouverture à une communauté plus transversale : les porteurs du nouveau projet ont identifié deux domaines avec lesquels ils souhaitent collaborer : les sciences historiques, qui partagent des pratiques et des intérêts sur les écrits avec CAHIER, et les sciences juridiques.

Les "corpus" et les "écrits" en question

Dès sa fondation, le Consortium CAHIER s'est distingué par l'attention particulière qu'il a portée aux Corpus d'auteurs, aux textes et au geste éditorial qui accompagne la construction de ceux-ci et leur circulation. Si CAHIER n'a pas posé de définition claire et précise de la notion de CORPUS, il a fait sienne, en "l'adoptant" dans une certaine mesure (Lebarbé, 2009), une partie de la définition proposée par la linguistique de CORPUS (Rastier, 2004) et notamment celle d'"ensemble constitué" en vue d'atteindre un objectif scientifique (Sinclair, 1996). CAHIER a

également intégré à ses pratiques certaines des méthodes du domaine telles que l'annotation, l'analyse quantifiée et l'abstraction.

Au fil des ans, le consortium a constaté que le corpus d'auteur s'est progressivement ouvert, détaché de l'ensemble supposé fermé car lié à un auteur précis pour, de plus en plus, et notamment grâce à l'accessibilité des données, tendre à désigner un ensemble plus large et surtout plus hétérogène incluant des jeux de données dynamiques, parfois captées « à la volée » à partir de ressources disponibles et dont certaines peuvent avoir été enrichies par des annotations automatiques ou expertes (ou pourront l'être, dans le cadre d'un nouveau projet). Les corpus de CAHIER sont ainsi devenus des construits scientifiques ouverts, constitués et re-constitués en vue de répondre à une question de recherche qui évolue avec les données et qui embrasse un ou plusieurs auteur(s), une ou plusieurs thématiques, une ou plusieurs périodes historiques. La notion de "corpus" s'est étendue pour devenir plus ouverte et plus interdisciplinaire mais avec les corpus numériques se pose la question de la gestion des versions des fichiers, notamment après les interventions du chercheur (comme dans le cas des couches d'annotations) et de leurs situations, c'est-à-dire de leurs contextes et de leurs historicités.

Alors que les linguistes ont proposé de nouveaux concepts pour circonscrire les corpus en milieux numériques, tel que celui de « lieu de corpus » (Emerit, 2016), ce travail reste à faire du côté des autres sciences littéraires notamment, à travers le croisement de points de vues, la redéfinition du type et des nouvelles formes des objets scientifiques en milieu numérique. Le nouveau projet de CAHIER vise à interroger le type de modèle ou d'élaboration de corpus qui peut être proposé à l'avenir tout en s'inscrivant dans le modèle vertueux de la science ouverte et FAIR.

Intérêt de la proposition : créer une nouvelle communauté scientifique associant sciences des textes, sciences historiques et sciences juridiques.

COSME est, naturellement, le premier interlocuteur repéré par CAHIER, à la fois parce que les deux consortiums ont déjà eu l'occasion de collaborer par le passé (groupe de travail « Questions juridiques ») mais également parce qu'ils partagent plusieurs sujets : les écrits, les corpus, l'exploitation de motifs, etc. Dès octobre 2021, ils ont commencé à échanger sur le type de passerelles, de questions et d'objets qu'ils pourraient, ensemble, traiter : la « trace écrite » semble constituer un point de convergence et une nouvelle réunion de travail avec les membres de COSME sera organisée lors de l'assemblée générale de novembre 2021.

Un autre défi scientifique repéré par les porteurs du nouveau projet est celui de la redéfinition des notions d'« auteur » et d'« auctorialité » ; avec le numérique, les corpus sont devenus plus vastes mais également plus hétérogènes et la notion d'auteur se voit elle aussi diluée dans un environnement aux auctorialités et « contributions » multiples. C'est avec les juristes et les historiens du droit qu'il semble important, intéressant et pertinent de discuter sur ces questions et notions. Les historiens et chercheurs en Droit portent une attention poussée à leurs sources écrites or leurs travaux sont actuellement absents des réseaux scientifiques soutenus par Huma-Num. CAHIER a pu compter sur la collaboration "technique" des sciences juridiques lors de l'élaboration de son *Guide juridique* (Dord-Crouslé, et. al., 2017), mais les apports scientifiques des juristes pourraient nourrir, beaucoup plus largement, les travaux théoriques envisagés par le consortium sur les notions de "corpus", "d'écrits", de traces et sur l'évolution des notions "d'auteur" et "d'auctorialités". L'équipe qui porte le projet est actuellement en train de prendre des contacts afin d'intégrer dans la nouvelle communauté de « CAHIER après CAHIER », le domaine des sciences juridiques.

La majorité des travaux réalisés à ce jour par CAHIER et par d'autres consortiums d'Huma-Num, comme COSME ou CORLI par exemple, ont mis l'accent sur les "bonnes pratiques" en matière de création de fichiers, d'usage de formats et de standards qui assurent le stockage à long terme et la réutilisation des données, et en matière de formation aux outils d'accompagnement et de construction de corpus. Ces consortiums n'ont toutefois pas partagé leurs questionnements théoriques alors qu'ils portent des intérêts similaires pour certains des problèmes posés par les écrits, les traces écrites et leurs circulations numériques. COSME a par exemple initié une réflexion sur les problèmes posés par le champ DATE dans les fichiers numériques tandis que CAHIER a porté différents travaux sur les problèmes juridiques posés par le numérique. De même, CAHIER a développé des réflexions allant dans le sens d'une séquentialisation des étapes de constitution des CORPUS (passant par la mise à disposition de plusieurs états d'un même fichier, du plus "pauvre" au plus "riche", *in* Galleron et al., 2018) à des fins de "réutilisabilité" technique et scientifique (Galleron, Idmhand, 2020), et élaboré une typologie destinée à faciliter la mise en lien des objets dans le web de données (par exemple pour la "Typologie des textes", un article est en cours de rédaction). Enfin, et comme cela a été constaté au sein de CAHIER, les pratiques scientifiques des spécialistes des sciences du texte sont souvent hétérogènes or il est essentiel de parvenir à mettre à plat les divergences pour trouver des formats pivot, et cette tâche ne peut être effectuée que par des humanistes qui possèdent déjà une bonne pratique du numérique. Il apparaît donc qu'au regard des propositions et résultats des différents consortiums (comme CAHIER et COSME), celles-ci

pourraient être croisées et approfondies dans le cadre d'un nouveau réseau scientifique en vue de partager des modèles, mais également des théories et des méthodologies pour la création de grands (et riches) corpus numériques pour la recherche.

Alignement du nouveau consortium sur les missions et les objectifs de la TGIR Huma-Num

Le nouveau consortium constitué aura pour but de produire des connaissances sur ces concepts, mais également sur d'autres notions problématiques au regard de l'évolution numérique. En lien avec l'infrastructure, le consortium contribuera aux missions suivantes :

- ✓ Mission de relais

Ce consortium mettra fortement l'accent sur l'accès ouvert, l'interopérabilité et la réutilisabilité des données. De cette façon, il répercutera les recommandations de la TGIR Huma-Num vers le terrain tout en jouant un rôle de relais vers l'infrastructure, en communiquant les besoins et problèmes de la communauté. En ce sens, le maintien des réseaux initialement constitués par le(s) consortium(s) contribueront à la coordination des actions de valorisation et augmenteront la visibilité des données produites par les membres. Ils stimuleront les réflexions sur leur réutilisation dans le cadre de travaux entrepris dans et hors du réseau.

- ✓ Mission de mutualisation

La création de corpus dynamiques constitue un terreau fertile pour les réflexions interdisciplinaires induites par les Humanités numériques. Le croisement des points de vue techniques, technologiques, ontologiques et théoriques restera la clé de voûte du nouveau consortium. Celui-ci identifiera des personnes ou équipes ressources qui faciliteront l'accès, le développement et la transformation des documents et des données de façon à ce qu'ils puissent être utilisés et exploités selon les besoins de la communauté, et la mutualisation des outils. Il sera important d'assurer une veille et de conseiller les communautés scientifiques au sujet des techniques et méthodes les plus appropriées.

Les formations proposées par le consortium ainsi que le travail d'animation mené par celui-ci contribueront au partage et à la dissémination des connaissances scientifiques, des compétences techniques et des technologies.

Programme de l'année 2022

Les travaux menés par durant l'année 2022 auront pour objectif d'associer les chercheurs de ces consortiums en vue d'élaborer un nouveau projet. Le premier semestre de l'année 2022 sera consacré à la prise de contact, puis à l'organisation de réunions de travail mensuelles en vue de rédiger un nouveau projet de consortium associant littéraires, historiens et juristes autour des concepts de "CORPUS", d'"ECRITS", un concept qui évolue rapidement en raison des "intermédialités" (Méchoulan, 2017), et de "TRACES". Ces réunions auront pour but d'établir le programme de travail qui permettra d'élaborer de nouvelles connaissances sur ces concepts et de mettre en place de nouvelles recommandations pour l'échange de données (formalisation de la présentation des dates, choix d'une date en l'absence de datation, typologie de périodes historiques, etc.) ou sur la mise en place d'un «cercle vertueux » d'annotations.

Cette proposition initie un nouveau cycle de réunions de travail destinées à étayer la proposition de consortium qui sera soumise à Huma-Num à l'automne 2022. Suite à la réduction et à l'annulation (ou au report) d'un nombre important des actions budgétées en 2020 et 2021 (réunions des groupes de travail ou formations en présentiel) en raison de la pandémie de Covid-19, CAHIER dispose d'un reliquat budgétaire qui lui permet de travailler, durant l'année 2022, à la concrétisation de cette proposition.

Le calendrier proposé a pour objectif de rédiger la proposition détaillée et les réunions de travail seront largement ouvertes à de nouveaux membres et à de nouvelles équipes.

Parallèlement à l'élaboration du projet, CAHIER poursuivra l'accompagnement initié en 2020 et 2021 en vue de FAIRiser les données du consortium. Le groupe de travail "FAIRisation des données" sera lui aussi ouvert à d'autres consortiums.

- ✓ Novembre 2021 à Novembre 2022

Les différents groupes de travail cités ci-dessous sont constitués pour une année. Un appel à participation sera susceptible de compléter cette liste :

GT : "Projet scientifique nouveau consortium" : Ce groupe réunira les membres de CAHIER et invitera des membres de consortiums en fin de labellisation (COSME par exemple), des membres de laboratoires en sciences juridiques ainsi que des personnes extérieures intéressées par le projet. Ce groupe organisera des séminaires de travail (de type *workshops*) et commencera ses travaux dès l'Assemblée générale de CAHIER fin novembre 2021

GT : "Ateliers thématiques du consortium CAHIER" : ce GT organisera, pour une année encore, les écoles

thématiques et formations annuelles de CAHIER. Elles contribueront à renforcer et consolider la communauté en vue du nouveau consortium

Groupe de travail "FAIRisation des données" : le GT poursuivra, en lien avec Huma-Num, l'accompagnement de la communauté à la FAIRisation des données produites. Il sera ouvert aux autres consortiums et à toutes celles et ceux qui souhaitent FAIRiser leurs données

Groupe de travail "Communication" : ce GT aura pour but d'archiver les travaux des anciens consortiums (ancien carnet de recherches de CAHIER par exemple, etc.)

Groupe de travail "Typologies" : dans le sillage du GT Typologies textuelles du consortium CAHIER, il aura en charge non seulement le maintien et le développement du thésaurus déjà créé sur les "Typologies textuelles", mais également la mise en oeuvre d'autres vocabulaires partagés pour les projets de corpus.

- ✓ Mai 2022 - Juin 2022 : séminaire de travail "Nouveau projet"

Organisé comme workshop sur deux ou trois jours par le GT "Projet scientifique" il s'agira d'un "atelier annuel" transformé dont le but sera :

réunir en présentiel les groupes de travail
organiser des séances de formation si besoin (à des outils d'exploitation interopérables et FAIR, etc.)
présenter des corpus en vue du nouveau projet
éventuellement, organiser des réunions ou petits ateliers thématiques

Bibliographie

Dord-Crouslé, Stéphanie, Greslou, Elisabeth, Hue-Gay, Elysabeth et Pierrot, Denise, *Un guide juridique pour l'édition numérique de corpus d'auteurs*, Consortium CAHIER, 2017 URL: https://cahier.hypotheses.org/3074 ; https://halshs.archives-ouvertes.fr/halshs-03400177

Emerit, Laetitia, « La notion de lieu de corpus : un nouvel outil pour l'étude des terrains numériques en linguistique », Corela [En ligne], 14-1 | 2016, URL : http://journals.openedition.org/corela/4594 ; DOI : 10.4000/corela.4594

Galleron, Ioana et Idmhand, Fatiha. (2019) "'Réutilisabilité' : L'utilisateur dans l'édition électronique", revue Humanistica, numéro 1, 2019, https://revues.univ-lyon3.fr/humanites-numeriques/

Galleron, Ioana et Idmhand, Fatiha. (2020) « Why Go from Texts to Data, or the Digital Humanities as A Critique of the Humanities », Word and Text, no. X, p. 53-69, http://jlsl.upg-ploiesti.ro/site_engleza/No_1_2020.html

Galleron, Ioana et Idmhand, Fatiha (eds). (2021) Dix ans avec CAHIER: des corpus d'auteur pour les humanités à leur exploitation numérique. Paris, Éditions des archives contemporaines (sous presse).

Lebarbé, Thomas, « Du corpus littéraire au corpus linguistique : dématérialisation, restructuration, lectures rhizomatiques et analyses linguistiques des manuscrits », *Corpus* [Online], 8 | 2009. URL : http://journals.openedition.org/corpus/1694 ; DOI : https://doi.org/10.4000/corpus.1694

Méchoulan, Éric, « Intermédialité, ou comment penser les transmissions », *Fabula / Les colloques*, "Création, intermédialité, dispositif", 2017. URL : http://www.fabula.org/colloques/document4278.php.

Rastier, François, , Enjeux épistémologiques de la linguistique de corpus. Texto ! [en ligne], juin 2004. Rubrique Dits et inédits. Disponible sur : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html. Sinclair, John, *Preliminary Recommendations on Corpus Typology*, EAGLES (Expert Advisory Group on Language Engineering Standards) Report, 1996. URL : http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html.

Sinclair, John, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press Esp, 1991.

Version du projet qui a été présentée aux membres du consortium le 30-09-2021 et soumise au vote

CAHIER après CAHIER
Vers un nouveau Consortium (2023-2027)

Introduction, contexte

Après dix ans d'existence, le Consortium CAHIER a construit un large réseau scientifique national, expert dans la constitution, l'exploitation et la pérennisation des données des corpus d'auteurs. La communauté, active et pluridisciplinaire, souhaite préserver le réseau d'échanges de pratiques, de données, d'outils et de résultats qu'elle a tissé au fil des ans mais elle a conscience qu'il est nécessaire de l'ouvrir à de nouvelles perspectives de recherche, à travers une communauté renouvelée fédérée au sein d'un nouveau projet scientifique.

Ces dix dernières années, plusieurs consortiums d'Huma-Num se sont intéressés à la création de corpus et ont développé des expertises poussées sur la numérisation, la conversion en données numériques exploitables des écrits et sur les formats et standards qui facilitent l'interopérabilité et la réutilisabilité des informations. Structurés en communautés scientifiques expertes des corpus littéraires, linguistiques, thématiques et historiques, ils ont abordé nombre de problèmes qui traversent les Humanités numériques et ont tenté d'apporter des réponses en ce qui concerne les types de corpus constitués pour la recherche (description typologique des œuvres), la portée de certains formats d'échange (tels que le XML-TEI) ou les droits applicables aux publications numériques des sources (droits des auteurs, des éditeurs auteurs d'annotations numériques, etc.). Ils ont également contribué à augmenter la quantité de données et de corpus disponibles, les ont FAIRisées de façon à faciliter l'intégration de ceux-ci dans de nouvelles collections numériques. En contribuant de cette façon à constituer de nouveaux corpus plus larges, mais également plus diversifiés et plus hétérogènes, les corpus ainsi mis en données posent de nouveaux défis à la communauté scientifique que CAHIER propose de traiter au sein d'une nouvelle communauté scientifique interdisciplinaire dont le but sera d'interroger les enjeux et nouvelles modalités des CORPUS et des ECRITS.

Les "corpus" et les "écrits" en question

Dès sa fondation, le Consortium CAHIER s'est distingué par l'attention particulière qu'il a portée aux Corpus d'auteurs, aux textes et au geste éditorial qui accompagne la construction de ceux-ci et leur circulation. Si CAHIER n'a pas posé de définition claire et précise de la notion de CORPUS, il a fait sienne, en "l'adoptant" dans une certaine mesure (Lebarbé, 2009), une partie de la définition proposée par la linguistique de CORPUS (Rastier, 2004) et notamment celle d'"ensemble constitué" en vue d'atteindre un objectif scientifique (Sinclair, 1996). CAHIER a également intégré à ses pratiques certaines des méthodes du domaine telles que l'annotation, l'analyse quantifiée et l'abstraction.

Au fil des ans, le corpus d'auteur s'est progressivement ouvert et détaché de l'ensemble supposé fermé car lié à un auteur précis; de plus en plus, et notamment grâce à l'accessibilité des données, il tend à désigner un ensemble plus large et surtout plus hétérogène incluant des jeux de données dynamiques, parfois captées « à la volée » à partir de ressources disponibles et dont certaines peuvent avoir été enrichies par des annotations automatiques ou expertes (ou pourront l'être, dans le cadre d'un nouveau projet). Les corpus de CAHIER sont ainsi devenus des construits scientifiques ouverts, constitués et re-constitués en vue de répondre à une question de recherche qui évolue avec les données, qui embrasse un ou plusieurs auteur(s), une ou plusieurs thématiques, une ou plusieurs périodes historiques. La notion de "corpus" s'est elle aussi étendue pour devenir plus ouverte et plus interdisciplinaire car avec les corpus numériques, se pose la question de la gestion des versions des fichiers, notamment après les interventions du chercheur (comme dans le cas des couches d'annotations) et de leurs situations, c'est-à-dire de leurs contextes et de leurs historicités.

Les linguistes ont proposé de nouveaux concepts pour circonscrire les corpus en milieux numériques, tel que celui de « lieu de corpus » (Emerit, 2016), mais un travail reste à faire du côté des autres sciences humaines pour croiser les points de vues et définir le type et les nouvelles formes de leurs objets scientifiques en milieu numérique. Quel modèle d'élaboration des corpus intégrant leurs dimensions contextuelles et s'inscrivant dans le modèle vertueux de la science ouverte et FAIR peut être proposé?

Intérêt de la proposition : créer une nouvelle communauté scientifique associant sciences des textes, sciences historiques et sciences juridiques.

La majorité des travaux réalisés à ce jour par CAHIER et par d'autres consortiums d'Huma-Num, comme

COSME par exemple, ont mis l'accent sur les "bonnes pratiques" en matière de création de fichiers, d'usage de formats et de standards qui assurent le stockage à long terme et la réutilisation des données et en matière de formation aux outils d'accompagnement et de construction de corpus. Ces consortiums n'ont toutefois pas partagé leurs questionnements théoriques alors qu'ils portent des intérêts similaires pour certains des problèmes posés par les écrits et leurs circulations numériques. COSME a par exemple initié une réflexion sur les problèmes posés par le champ DATE dans les fichiers numériques tandis que CAHIER a porté différents travaux sur les problèmes juridiques posés par le numérique. De même, CAHIER a développé des réflexions allant dans le sens d'une séquentialisation des étapes de constitution des CORPUS (passant par la mise à disposition de plusieurs états d'un même fichier, du plus "pauvre" au plus "riche", *in* Galleron et al., 2018) à des fins de "reutilisabilité" technique et scientifique (Galleron, Idmhand, 2020), et élaboré une typologie destinée à faciliter la mise en lien des objets dans le web de données (par exemple pour la "Typologie des textes", un article est en cours de rédaction).

Il apparaît donc qu'au regard des propositions et résultats des différents consortiums (comme CAHIER et COSME), celles-ci pourraient être croisées et approfondies dans le cadre d'un nouveau réseau scientifique en vue de partager des théories et des méthodologies pour la création de grands (et riches) corpus numériques pour la recherche. La proposition ici formulée vise à tenter d'associer les chercheurs de ces consortiums pour mettre en place de nouvelles recommandations pour l'échange de données (formalisation de la présentation des dates, choix d'une date en l'absence de datation, typologie de périodes historiques, etc.) ou la mise en place d'un « cercle vertueux » d'annotations. Ils pourraient également publier des travaux théoriques sur l'évolution de la notion de CORPUS dans les sciences historiques et littéraires et sur celle d'"ECRITS", un concept qui évolue rapidement en raison des "intermédialités" (Méchoulan, 2017).

CAHIER veut également intégrer dans cette nouvelle communauté les spécialistes des sciences juridiques. Alors que les historiens et chercheurs en Droit portent une attention poussée aux sources écrites du Droit, leurs travaux sont actuellement absents des réseaux scientifiques soutenus par Huma-Num. CAHIER a pu compter sur la collaboration "technique" des sciences juridiques lors de l'élaboration de son *Guide juridique* (Dord-Crouslé, et. al., 2017), mais les apports scientifiques des juristes pourraient nourrir, de façon plus riche, les travaux théoriques envisagés sur les notions de "corpus", "d'écrits" mais également sur l'évolution des notions "d'auteur" et "d'auctorialités" envisagés par le consortium.

Alignement sur les missions et les objectifs de la TGIR Huma-Num

En conséquence, CAHIER propose de consacrer l'année 2021-2022 à la prise de contact, à l'organisation de réunions de travail et à la rédaction du projet d'un nouveau projet de consortium associant littéraires, historiens et juristes en vue d'interroger les concepts de "CORPUS" et d' "ECRITS". Le nouveau consortium constitué aura pour but de produire de nouvelles connaissances sur ces concepts, mais également sur d'autres notions problématiques au regard de l'évolution numérique et, en lien avec l'infrastructure, il contribuera aux missions suivantes :

- ✓ Mission de relais

Ce nouveau consortium mettra fortement l'accent sur l'accès ouvert, l'interopérabilité et la réutilisabilité des données. De cette façon, il répercutera les recommandations de la TGIR Huma-Num vers le terrain tout en jouant un rôle de relais vers l'infrastructure, en communiquant les besoins et problèmes de la communauté. En ce sens, le maintien des réseaux initialement constitués par le(s) consortium(s) contribueront à la coordination des actions de valorisation et augmenteront la visibilité des données produites par les membres. Ils stimuleront les réflexions sur leur réutilisation dans le cadre de travaux entrepris dans et hors du réseau.

- ✓ Mission de mutualisation

La création de corpus dynamiques constitue un terreau fertile aux réflexions interdisciplinaires induites par les Humanités numériques. Le croisement des points de vue techniques, technologiques, ontologiques et théoriques restera la clé de voûte du nouveau consortium. Celui-ci identifiera des personnes ou équipes ressources qui faciliteront l'accès, le développement et la transformation des documents et des données de façon à ce qu'ils puissent être utilisés et exploités selon les besoins de la communauté, et la mutualisation des outils. Il sera important d'assurer une veille et de conseiller les communautés scientifiques au sujet des techniques et méthodes les plus appropriées.

Les formations proposées par le consortium ainsi que le travail d'animation mené par celui-ci contribueront au partage et à la dissémination des connaissances scientifiques, des compétences techniques et des technologies.

Programme de l'année 2021-2022

Cette proposition initie un nouveau cycle de réunions de travail destinées à étayer la proposition de consortium qui sera soumise à Huma-Num à l'automne 2022. Suite à la réduction et à l'annulation (ou au report) d'un nombre important des actions budgétées en 2020 et 2021 (réunions des groupes de travail ou formations en présentiel) en raison de la pandémie de Covid-19, CAHIER dispose d'un reliquat budgétaire qui lui permet de travailler, durant

l'année 2022, à la concrétisation de cette proposition.

Le calendrier proposé a pour objectif de rédiger la proposition détaillée: les réunions de travail seront largement ouvertes à de nouveaux projets et membres, et à de nouvelles équipes.

Parallèlement à l'élaboration du projet, CAHIER poursuivra l'accompagnement initié en 2020 et 2021 en vue de FAIRiser les données du consortium. Le groupe de travail "FAIRisation des données" sera ouvert à d'autres consortiums.

- ✓ Janvier-Décembre 2022

Les différents groupes de travail cités ci-dessous sont constitués pour une année. Un appel à participation sera susceptible de compléter cette liste :

GT : "Projet scientifique nouveau consortium" : Ce groupe réunira les membres de CAHIER et invitera des membres de consortiums en fin de labellisation (COSME par exemple), des membres de laboratoires en sciences juridiques ainsi que des personnes extérieures intéressées par le projet. Ce groupe organisera des séminaires de travail (de type *workshops*) qui seront organisés.

GT : "Ateliers thématiques du consortium CAHIER" : ce GT organisera, pour une année encore, les écoles thématiques annuelles de CAHIER. Elles contribueront à renforcer et consolider la communauté en vue du nouveau projet

Groupe de travail "FAIRisation des données" : le GT poursuivra, en lien avec Huma-Num, l'accompagnement de la communauté à la FAIRisation des données produites. Il sera ouvert aux autres consortiums et à toutes celles et ceux qui sont intéressés

Groupe de travail "Communication" : ce GT aura pour but d'archiver les travaux des anciens consortiums (ancien carnet de recherches de CAHIER par exemple, etc.)

Groupe de travail "Typologies" : dans le sillage du GT Typologies textuelles du consortium CAHIER, il aura en charge non seulement le maintien et le développement du thésaurus déjà créé sur les "Typologies textuelles", mais également la mise en oeuvre d'autres vocabulaires partagés pour les projets de corpus.

- ✓ Mai / juin 2022 : séminaire de travail "Nouveau projet"

Organisé comme workshop sur deux ou trois jours par le GT "Projet scientifique" il s'agira d'un "atelier annuel" transformé dont le but sera :

- réunir en présentiel les groupes de travail
- organiser des séances de formation si besoin (à des outils d'exploitation interopérables et FAIR, etc.)
- présenter des corpus en vue du nouveau projet
- éventuellement, organiser des réunions ou petits ateliers thématiques

Bibliographie

Dord-Crouslé, Stéphanie, Greslou, Elisabeth, Hue-Gay, Elysabeth et Pierrot, Denise, *Un guide juridique pour l'édition numérique de corpus d'auteurs*, Consortium CAHIER, 2017 URL: https://cahier.hypotheses.org/3074

Emerit, Laetitia, « La notion de lieu de corpus : un nouvel outil pour l'étude des terrains numériques en linguistique », Corela [En ligne], 14-1 | 2016, URL : http://journals.openedition.org/corela/4594 ; DOI : 10.4000/corela.4594

Galleron, Ioana et Idmhand, Fatiha. (2019) "'Réutilisabilité' : L'utilisateur dans l'édition électronique", revue Humanistica, numéro 1, 2019, https://revues.univ-lyon3.fr/humanites-numeriques/

Galleron, Ioana et Idmhand, Fatiha. (2020) « Why Go from Texts to Data, or the Digital Humanities as A Critique of the Humanities », Word and Text, no. X, p. 53-69, http://jlsl.upg-ploiesti.ro/site_engleza/No_1_2020.html

Galleron, Ioana et Idmhand, Fatiha (eds). (2021) Dix ans avec CAHIER: des corpus d'auteur pour les humanités à leur exploitation numérique. Paris, Éditions des archives contemporaines (sous presse).

Lebarbé, Thomas, « Du corpus littéraire au corpus linguistique : dématérialisation, restructuration, lectures rhizomatiques et analyses linguistiques des manuscrits », *Corpus* [Online], 8 | 2009. URL : http://journals.openedition.org/corpus/1694 ; DOI : https://doi.org/10.4000/corpus.1694

Méchoulan, Éric, « Intermédialité, ou comment penser les transmissions », *Fabula / Les colloques*, "Création, intermédialité, dispositif", 2017. URL : http://www.fabula.org/colloques/document4278.php.

Rastier, François, , Enjeux épistémologiques de la linguistique de corpus. Texto ! [en ligne], juin 2004. Rubrique Dits et inédits. Disponible sur : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html.

Sinclair, John, *Preliminary Recommendations on Corpus Typology*, EAGLES (Expert Advisory Group on Language Engineering Standards) Report, 1996. URL : http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html.

Sinclair, John, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press Esp, 1991.