“ Consortium-HN ARIANE ”. Synthèse du projet scientifique

Fatiha Idmhand, Ioana Galleron, Sabine Loudcher

« Consortium-HN ARIANE »

Analyses, **R**echerches, **I**ntelligence **A**rtificielle et **N**ouvelles **É**ditions de textes

(2023-2027)

1. **Présentation**

Le consortium ARIANE (Analyses, Recherches, Intelligence Artificielle et Nouvelles Editions numériques) a été labellisé par Huma-Num le 18/01/2023 pour une période de 4 ans. ARIANE réunit des spécialistes du texte (littéraires, linguistes, historiens…) et de l'informatique en vue de créer un espace de dialogue véritablement interdisciplinaire entre ces deux communautés. L'objectif du consortium ARIANE est de progresser dans la connaissance et le raffinement des méthodes informatiques appliquées aux objets et données des sciences humaines et plus particulièrement, des sciences du texte ; il veut constituer un espace de discussion portant sur l'interprétation des résultats obtenus à l'aide de méthodes (semi)automatiques d'analyse des textes, car le sens et la validité des extractions effectuées ne sont pas immédiats. Pour cela, le consortium vise à faciliter l'accès de la communauté aux algorithmes, scripts et chaînes de traitement qui simplifient la manipulation et l'enrichissement des textes et à stimuler la création de nouvelles connaissances à leur sujet et sur leurs contextes. Au sein d'Huma-Num, le consortium contribuera aux réflexions et même au développement de nouveaux outils, plateformes et interfaces.

2. **Disciplines concernées par les travaux du consortium**

   Disciplines principales : littérature, génétique des œuvres, linguistique, informatique

   Disciplines secondaires : histoire, sciences de la documentation, droit

3. **Coordinatrices du consortium**

   Fatiha IDMHAND, Université de Poitiers, Institut des Textes et Manuscrits Modernes (UMR 8132), fatihaidmhand@yahoo.es

   Ioana GALLERON, Université Sorbonne-Nouvelle, LATTICE (UMR 8094), ioana.galleron@sorbonne-nouvelle.fr

   Sabine LOUDCHER, Université Lyon 2, ERIC (UR 3083), sabine.loudcher@univ-lyon2.fr

4. **Email de contact** : coordination-cst-hn-ariane@groupes.renater.fr

5. **Gestion financière du consortium (Unité CNRS)** : MSHS Poitiers
6. **Listes de diffusion**

Liste de discussion publique ouverte à tous et toutes : https://groupes.renater.fr/sympa/info/infos-cst-ariane (infos-cst-ariane@groupes.renater.fr)

Liste réservée aux membres du Consortium : membres-cst-ariane@groupes.renater.fr

7. **Objectifs du consortium**

Alors que la création de données en SHS va croissant, les nouvelles épistémologies et les connaissances fondées sur celles-ci restent encore à constituer. La raison de cet état de fait est, d'une part, que les données produites massivement sont souvent "pauvres" et que, dès lors, elles ne permettent pas de répondre aux questions complexes des SHS et d'aller véritablement au-delà de ce qui est déjà connu. D'autre part, le temps nécessaire à la création de ces données (constituées le plus souvent de façon manuelle), ainsi que leur hétérogénéité, freinent à la fois les réinvestissements méthodologiques et la réutilisabilité des corpus. Ainsi, nombre des corpus produits dans le cadre de consortiums tels que CAHIER par exemple, sont restés au stade de la numérisation en mode image, accompagnée d'un jeu minimal de métadonnées. Et quoiqu'elles suivent les recommandations de standards (comme le Dublin Core), l'exploitation de ces données et métadonnées reste encore limitée, voire impossible, en raison, notamment, de pratiques fort disparates et de l'absence de vocabulaires communs pour la saisie de mots-clés, pour n'évoquer qu'un aspect du problème. Par ailleurs, dans les disciplines du champ littéraire (langue et littérature), le fait de disposer de textes « bruts » ou minimalement structurés tels que ceux qui peuvent être issus de campagnes massives d'océrisation, n'est pas toujours suffisant pour répondre aux questions de recherche. En effet, les unités sémantiques sur lesquelles les chercheurs du domaine ont besoin de s'appuyer se situent à d'autres niveaux : celui de la structure non manifeste du texte (les mouvements rhétoriques ou les maillons narratifs), ou celui des segments supra-lexicaux complexes (tels que les métaphores, les chaînes de co-référence, les citations d'autres textes, signalées ou non - pour ne donner, là aussi, que quelques exemples). Pour d'autres travaux, c'est l'accès à des corpus alignés qui se révèle nécessaire, alors que leur constitution est extrêmement lente et complexe, car presque exclusivement manuelle. Enfin, beaucoup reste à faire dans le domaine même de l'acquisition des textes, car les sources les moins explorées, et donc les plus susceptibles de mener à la création de nouvelles connaissances et lectures du passé, sont souvent les plus difficiles à numériser, en raison de leur état de conservation, de la qualité de leur support et écriture, etc. – ce qui les exclut souvent des campagnes menées en bibliothèque, pose des problèmes d'océrisation et suscite des réticences compréhensibles du côté des chercheurs. Il devient dès lors essentiel de trouver de nouvelles façons de traiter et d'explorer ces ressources textuelles numériques, plus ou moins connues et présentant des degrés de raffinement divers : c'est l'objectif du Consortium HN ARIANE.

Le but d'ARIANE est de créer l'espace d'un dialogue véritablement interdisciplinaire entre les communautés scientifiques des spécialistes des sciences du texte et de l'informatique, en vue de

répondre aux enjeux décrits plus haut et de progresser dans la connaissance et le raffinement des méthodes informatiques pour la manipulation des objets et données des sciences humaines et plus particulièrement, des sciences du texte. Pour cela, le consortium facilitera l'accès de la communauté aux algorithmes, scripts et chaînes de traitement qui simplifient la manipulation et l'enrichissement des textes, il stimulera la création de nouvelles connaissances à leur sujet et sur leurs contextes et, au sein d'Huma-Num, il contribuera aux réflexions, voire au développement de nouveaux outils, plateformes et interfaces.

8. Valeur ajoutée du Consortium

Les travaux proposés par le consortium ARIANE s'inscrivent dans le domaine du traitement automatique des textes. Ils s'appuient sur une série d'algorithmes et de méthodes issus du traitement automatique des données et de la langue et ont pour but de développer une approche spécifique pour le traitement automatique des textes. Ces textes nécessitent une approche spécifique en raison des objets concernés (documents patrimoniaux le plus souvent) et de la finalité visée, celle de l'aide à l'interprétation. Le domaine se trouve confronté à de nombreux défis en raison de la dimension réticulaire des objets étudiés (texte = tissage, texte = contexte) laquelle pose de multiples problèmes lorsqu'il s'agit d'opérer leur discrétisation en vue de leur numérisation, de la constitution d'une base de données, d'un encodage, ou de leur mise en lien avec d'autres textes et contextes (de production, d'édition, de lecture, etc.). En se proposant de travailler à une meilleure compréhension et à la généralisation des méthodes de la science des données pour l'acquisition, la structuration et l'exploitation des textes, ARIANE contribuera à favoriser l'appropriation critique d'une nouvelle génération d'outils numériques et de méthodes innovantes par/pour les chercheurs des sciences humaines. Le but du consortium est la montée en compétences des chercheurs et la montée en gamme des outils, deux conditions nécessaires pour redonner aux premiers la possibilité de se recentrer sur le cœur de leur réflexion.

La valeur ajoutée du consortium proposé se situe également au niveau épistémologique. En effet, les travaux du consortium, tels que préfigurés plus haut, s'accompagnent de réflexions sur les tensions et les passages entre textes et données, entre lecture et extraction d'information, autrement dit, entre « humanités » et « numérique ». Se situant dans le prolongement de contributions plus anciennes ou plus récentes (v. Manovitch, 2003, McCarty, 2020, 2021, Citton, 2015, etc.), le consortium s'interrogera sur l'alternative à laquelle est confronté le chercheur en régime numérique : partir d'une théorie *a priori* suivie de la recherche de données permettant de la valider, ou le cheminement inverse, partir des données et élaborer ultérieurement une théorie.

Enfin offrant un espace d'expérimentation et de discussion interdisciplinaire sur les résultats des expériences, le consortium entend stimuler la réutilisabilité, qui reste actuellement, comme on l'a déjà dit, très en-deçà de ce que l'on pourrait attendre compte tenu du nombre de projets de numérisation réalisés depuis plus de dix ans. De même, en s'interrogeant sur ce que l'on voit et ce qui est occulté par le processus de « datification » des textes, en promouvant des parcours de recherche qui vont des hypothèses à la réflexion sur les données nécessaires pour les

tester, en insistant sur le droit à l'échec dans les humanités numériques, le consortium espère contribuer à la transformation des protocoles d'apport de preuve dans les sciences humaines et sociales.

9. Axe de recherches et groupes de travail

Pour sa première labellisation, le consortium travaillera selon les trois axes de travail suivants dans le cadre de groupes de travail du consortium (GT) et de groupes de travail partagés (GTP)

a) AXE 1. Éditions numériques de qualité

Les travaux entrepris dans le cadre de cet axe prolongeront en partie les travaux d'acquisition des textes en formats numériques dynamiques, entrepris par exemple (mais pas exclusivement) dans le cadre de CAHIER. Cependant, un accent plus prononcé sera mis sur de nouveaux objectifs tels que les méthodes les plus avancées d'océrisation et d'enrichissement (semi)automatique des textes avec des métadonnées et des balises sémantiques ainsi que la création de référentiels partagés, en appui des opérations de numérisation et annotation. La formation des membres à ces méthodes fera partie des travaux.

Trois groupes de travail seront immédiatement proposés pour cet axe :

- GT1 : Labellisation de projets en Humanités numériques et charte.
- GT2 : Acquisition de données et transcription assistée par ordinateur (OCR, HTR).
- GT3 : Éditions numériques de qualité.

b) AXE 2. Deep reading

Cet axe se concentrera sur l'épistémologie, les outils et les méthodes de la fouille de données et de métadonnées. L'objectif est de contribuer à la création de nouvelles connaissances en histoire et histoire de la littérature, théorie littéraire, stylistique et poétique, que ce soit à partir de (grands) corpus, de collections constituées selon une norme explicite, ou de textes individuels. Les objectifs de cet axe seront, notamment, d'actualiser la veille sur les outils et les méthodes d'analyse des textes assistée par ordinateur (textométrie, sentiment analysis, analyse de réseaux, fouille de métadonnées) et de former les membres du consortium à l'utilisation de ces nouvelles méthodes, de recommander la création de nouveaux outils/nouvelles fonctionnalités pour répondre à des questions spécifiques et, éventuellement, de s'engager dans la création de ces outils/ fonctionnalités en partenariat avec d'autres consortiums, institutions ou projets.

Les travaux de cet axe apporteront des valeurs ajoutées évidentes ; deux groupes de travail seront immédiatement proposés pour cet axe :

- GT4. Analyse automatique des textes.
- GT5. Métadonnées et modélisation de données.
- GT6. Open French Corpus - Thésaurus (groupes de travail partagé avec CORLI).

c) AXE 3. Problématiques transversales

Des problématiques et questions transversales aux deux axes seront également traitées et donneront lieu à des travaux de l'ensemble des membres du consortium. Deux sujets ont déjà été identifiés par les membres, ils concernent les questions juridiques et les questions d'éthique de la numérisation et de la mise en ligne. Alors que l'un des objectifs des humanités numériques est de rendre un maximum de textes accessibles à un maximum de personnes, la question qui se pose est celle de savoir si l'on peut vraiment mettre en ligne tous les textes, donner accès à tous les textes, et comment. En effet, certains des objectifs de la science ouverte entrent parfois en contradiction avec le droit de la propriété intellectuelle, le droit des données personnelles, voire avec le droit au respect de la vie privée, ceci s'explique par le fait que les règles ne sont pas toujours bien comprises mais également parce que la doctrine n'a pas encore apporté toutes les réponses aux problèmes identifiés. Si la rédaction de documents juridiques contribue à lever quelques verrous juridiques, une réflexion transversale se révèle nécessaire, de même que le suivi de la jurisprudence dans ce domaine.

10. Adhésion au consortium

L'adhésion au Consortium-HN ARIANE est ouverte à tous les chercheurs, jeunes chercheurs et ingénieurs désireux de contribuer à ses travaux. L'adhésion au consortium est individuelle, la participation effective aux activités du Consortium est appréciée individuellement. Rejoint le consortium qui adhère aux principes de la science ouverte (données FAIR, logiciels partagés, etc.). En rejoignant le consortium, les chercheurs, jeunes chercheurs et ingénieurs s'engagent à participer activement à au moins un groupe de travail : https://framaforms.org/demande-dadhesion-au-consortium-hn-ariane-1679997583

11. Liste des membres du Consortium à la date du 09/02/2023

Équipes porteuses du projet. Coordinatrices					
	N° de la structure	Laboratoire, Acronyme et Nom	Contact (nom, prénom, fonction, email)	Site Web	Domaine
1.	UMR 8132	ITEM Institut des Textes et Manuscrits Modernes	Idmhand, Fatiha (PU) fatihaidmhand@yahoo.es	https://fatihaidmhand.ovh/	Littérature, Langues Archives Humanités numériques
2.	UMR 8094	LATTICE Laboratoire Langues, Textes, Traitements informatiques, Cognition	Galleron, Ioana (PU) ioana.galleron@sorbonne-nouvelle.fr	http://www.univ-paris3.fr/mme-galleron-ioana-468922.kjsp	Littérature Théâtre Humanités numériques
3.	UR 3083	ERIC	Loudcher, Sabine (PU) sabine.loudcher@univ-lyon2.fr	https://eric.univ-lyon2.fr/~sabine/	Informatique Métadonnées Big data
Membres du consortium					
1.		ObTIC - Sorbonne Université	Alrahabi, Motassem (IR) motasem.alrahabi@gmail.com	http://lalic.paris-sorbonne.fr/PAGESPERSO/alrahabi/index.html	Linguistique TAL Humanités numériques
2.	UR 4335	3LAM Laboratoire Langues, Littératures, Linguistique des universités d'Angers et du Mans	Baillot, Anne (PU) anne.baillot@univ-lemans.fr	https://cv.archives-ouvertes.fr/annebaillot	Littérature, Langues Archives Humanités numériques
3.	UMR 8558	CRH EHESS Centre de Recherches Historiques	Brando, Carmen (IR) carmen.brando@ehess.fr	http://crh.ehess.fr/index.php?5088	Informatique
4.	UR 3083	ERIC	Jérôme Darmont (PU) jerome.darmont@univ-lyon2.fr	https://eric.univ-lyon2.fr/jdarmont/	Informatique Gestion de données

5.	UPR 3251	LIMSI Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur	Devillers, Laurence (PU) laurence.devillers@limsi.fr	https://laurence-devillers.com/	Informatique IA TAL
6.	UMR 5317	IHRIM Institut d'histoire des représentations et des idées dans les modernités	Dord-Crouslé, Stéphanie (CR) Stephanie.DordCrousle@ens-lyon.fr	https://cv.archives-ouvertes.fr/stephanie-dord-crousle	Littérature Archives Humanités numériques
7.	UMR 8643	LSV Laboratoire Spécification et Vérification	Dowek, Gilles (PU) gilles.dowek@ens-paris-saclay.fr	http://www.lsv.fr/~dowek/	Informatique Mathématiques
8.	EA 4661	ELLIADD Edition, Littératures, Langages, Informatique, Arts, Didactique, Discours	Froye, Marianne (MCF) marianne.froye@univ-fcomte.fr	https://cv.archives-ouvertes.fr/marianne-froye?langchosen=fr	Littérature Archives Humanités numériques
9.	UR4573	Transversales Droit, Contrat, Territoires	Kahn, Anne-Emmanuelle (MCF HDR) anne-emmanuelle.kahn@univ-lyon2.fr	https://dct.msh-lse.fr/node/47	Droit Droit de la propriété intellectuelle (propriété industrielle et propriété littéraire et artistique).
10.	UMR 7323	CESR, BVH Centre d'Études Supérieures de la Renaissance	Lastraioli, Chiara (PR) chiara.lastraioli@univ-tours.fr	https://www.cesr.cnrs.fr/chercheurs/chiara-lastraioli	Littérature, Langues Archives Humanités numériques
11.	UMR 5317	IHRIM	Lavrentev, Alexei (IR) alexei.lavrentev@ens-lyon.fr	http://ihrim.ens-lyon.fr/auteur/lavrentev-alexei	Linguistique TAL Humanités numériques
12.	UMR 6004	LSN Laboratoire des Sciences du Numérique de Nantes	Marinica, Claudia (MCF) claudia.marinica@univ-nantes.fr	https://perso-etis.ensea.fr/marinica/	Informatique Données hétérogènes Big data
13.	MNS HS	MNSHS Méthodes Numériques pour les Sciences de l'Humain et de la Société (Epitech Paris)	Puren, Marie (MCF) marie.puren@epitech.eu	https://cv.archives-ouvertes.fr/marie-puren	Littérature, Histoire Archives Humanités numériques
14.	UMR 8094	LATTICE Laboratoire Langues, Textes, Traitements informatiques, Cognition	Poibeau, Thierry (DR) thierry.poibeau@ens.psl.eu	https://www.lattice.cnrs.fr/membres/direction/thierry-poibeau/	Linguistique Informatique TAL
15.	UR4573	Transversales Droit, Contrat, Territoires	Quiquerez, Alexandre (MCF HDR) alexandre.quiquerez@univ-lyon2.fr	https://dct.msh-lse.fr/node/50	Droit Droit de la propriété intellectuelle
16.	UMR5317	IHRIM Institut d'Histoire des Représentations et des Idées dans les Modernités	Reboul, Marianne (MCF) marianne.reboul@ens-lyon.fr	http://ihrim.ens-lyon.fr/auteur/reboul-marianne	Littérature, Langues anciennes Humanités numériques
17.	RNSR : 20172248N	INRIA Institut national de recherche en sciences et technologies du numérique	Romary, Laurent (DR) laurent.romary@inria.fr	https://cv.archives-ouvertes.fr/laurentromary	Linguistique Informatique TAL
18.	UMR 8546	AOrOc Archéologie et philologie d'Orient et d'Occident	Stokes, Peter (DR) peter.stokes@ephe.psl.eu	https://www.ephe.psl.eu/annuaire/peter-stokes	Langues anciennes Informatique Humanités numériques
19.	UMR 5190	LARHRA/Lyon 2	Vernus, Pierre (MCF) pierre.vernus@msh-lse.fr	http://larhra.ish-lyon.cnrs.fr/membre/348	Histoire Archives Humanités numériques
20.	EA2449	DYPAC	Vitali, Giovanni (MCF) giovannipietrovitali@gmail.com	www.dypac.uvsq.fr	Linguistique, Histoire Archives Humanités numériques
21.	UMR 7323	CESR, BVH Centre d'Études Supérieures de la Renaissance	Marie-Luce Demonet (PR émérite) marie-luce.demonet@univ-tours.fr	https://www.cesr.cnrs.fr/chercheurs/marie-luce-demonet	Littérature Edition Humanités numériques