a.c.c.n.

architecture.conception.culture.numérique

2025-2026

Cartographie de la recherche doctorale en architecture sur les thématiques du numérique

RAPPORT DE MISSION DU RESEAU ACCN

ADRIEN SAINT-SARDOS, KEVIN JACQUOT, ISABELLE FASSE,
CAMILO CIFUENTES, PHILIPPE MARIN

LES RESULTATS DE L'ETUDE SONT CONSULTABLE A L'ADRESSE : HTTPS://ACCN.ARCHI.FR/CARTOGRAPHIE-DOCTORALE

Sommaire

Partie 1

Partie 2

Partie 1. La place des thématiques du Numérique dans la recherche doctorale des ENSA-P

Introduction

- Contexte et objectifs

Le présent rapport a été rédigé entre octobre et décembre 2025. Il présente les résultats de la mission de vacation de recherche commanditée par le Bureau Enseignement et Recherche (BER) du Ministère de la Culture, et confiée au réseau Architecture, Conception et Culture Numérique (ACCN), l'un des cinq Réseaux Scientifiques et Pédagogiques en Architecture (RSPA). Cette mission de vacation, d'une durée de trois mois, tend à **évaluer la place des thématiques et études liées au Numérique dans la recherche doctorale en Architecture au sein des ENSA-P** depuis 2010.

Cette volonté de suivre les travaux doctoraux en architecture n'est pas une exception, et la littérature en ligne présente des bilans comparables dans d'autres pays [(1),(2),(3),(4)]. Elle tient au caractère récent du cursus doctoral en architecture : en France, ce cursus apparaît dans les années 90, et se standardise sous l'impulsion des politiques d'uniformisation de l'éducation supérieure dans les années 2000 [voir la page dédiée sur le site du Ministère de la culture].

Le terme 'Numérique' est employé à de nombreuses reprises dans ce rapport, notamment pour qualifier les thématiques étudiées par la recherche. Sans pouvoir donner ici une définition univoque, nous y associons toute recherche utilisant ou étudiant précisément des méthodes et des outils du monde informatique, pour mener un projet d'architecture, concevoir, simuler, gérer des données massives, représenter ou transmettre de l'information.

- Usage du numérique

Le présent rapport est le fruit d'une conception intégralement humaine, aucun outil génératif n'a été utilisé pour sa rédaction ou sa relecture. Les outils d'IA ont été utilisés ponctuellement en soutien à la conception de codes d'analyses, lors d'analyses sémantiques poussées, et dans le cadre des travaux d'interface utilisateur décrits en partie IV. Aucune donnée sensible au sens de la RGPD n'a été fournie a des interfaces en ligne pendant la durée du projet.

Par soucis de traçabilité, les termes techniques liés au domaine du numérique seront en général exprimés en anglais, avec une définition et traduction proposées lors de leur première évocation.

En vue d'une une possible réutilisation, et par soucis de transparence, les codes utilisés pendant l'étude sont placés en annexe du présent rapport, dans une version annotée. Les datasets utilisés ont été partagés au BER, les auteurs laissent leur diffusion à leur discrétion. Toute démarche pouvant contribuer à l'interopérabilité/réutilisabilité des codes et des données (*FAIRness*) pourra être entreprise avec l'assentiment des auteurs.

- Remerciements

Les auteurs remercient en premier lieu Valérie Wathier pour son soutien à cette mission d'exploration et l'animation de la communauté des Réseaux. Nous exprimons notre gratitude à Armelle Le Mouëllic

pour la qualité de son travail de suivi des projets doctoraux au sein des ENSA-P, qui a servi de base à tous les bases de données et toutes analyses construites dans ce rapport.

Nous exprimons également notre gratitude aux auteurs du rapport *La place des thématiques écologiques dans les structures de recherche des ENSA-P* issus du réseau ENSA-éco, dont nous recommandons la lecture pour obtenir un angle de vue complémentaire de la recherche doctorale en architecture.

Description du Corpus

Résultat : Présentation du dataset

Dans le cadre de cet étude, nous considérons un dataset (ensemble de données) dérivant du document 250916_Liste docteurs diplômés des ENSA 2010-2025_diffusion RSPA.xlsx fourni BER à l'été 2025. Cette donnée est une liste de 629 doctorats soutenus, structurée comme présenté sur le Tableau 1. Dans la mesure du possible, nous avons conservé la structure originale du fichier à des fins d'interopérabilité des données.

Un rajout précieux tient dans les colonnes '*these_id*' et '*Abstract*'. Elles relient les doctorats à leur identifiant officiel sur theses.fr, et au résumé, en français, qui leur est associé sur internet. Avec l'identifiant, il est possible de reconstruire immédiatement une url valide : '*https://theses.fr/these_id*'. Il est à noter qu'un grand nombre de ces résumés ont dû être complétés *à la main* du fait de pages incomplètes ou d'échecs du moteur de requêtes de theses.fr.

La colonne la plus importante dans un premier temps est intitulée 'isNum'. Elle traduit l'appartenance ou non du doctorat aux recherches liées au Numérique, et donc au sous-dataset ARCHI-NUM. Ce sous-dataset contient exactement 100 thèses liées aux thématiques du numérique et soutenues entre 2010 et 2025.

Tableau 1 : Structure des dataset ARCHI et ARCHI-NUM, version 2025-11.

Nom de l'entrée	Vigilance RGPD	Remplissage dans ARCHI	Commentaire
Soutenance		5/5	Date de soutenance, format dd/mm/yyyy
Inscription		1/5	Date de l'inscription, non standardisée.
year		5/5	Année de soutenance, directement obtenue de la colonne 'Soutenance', format yyyy
Genre		5/5	Genre de la personne diplômée, complété manuellement, format 'H/F'
Nom	X	5/5	Nom de la personne diplômée.
Prénom	X	5/5	Prénom de la personne diplômée.
Mail	X	4/5	Adresse mail de la personne diplômée.
ENSA		5/5	Ecole Nationale Supérieure d'Architecture de rattachement pour le doctorat, format 'nom de ville'
Nature inscription		NA	Précisions en cas de rattachement exotique.
Co-tutelle ?		5/5	Booléen NON ou OUI+acteur hors ENSA.
Unités de recherche		5/5	Unité de recherche associée au projet de thèse.

UMR-EA		5/5	Unité de recherche associée au projet de thèse, nom standardisé sur la base du document BER *Liste septembre 2025 des unités de recherche des ENSA + responsables*
Ecole doctorale			ED de rattachement.
Direction de thèse	X	5/5	Liste des personnes ayant encadré le doctorat, séparées d'un saut de ligne
Financement		1/5	Précision sur le type de financement, très souvent non-communiqué « n.c. »
Titre de la thèse		5/5	Titre complet, traduit en français si nécessaire.
Discipline		5/5	Issu de la catégorie 'Discipline' de theses.fr
Mots clé		5/5	Issus des 'mots clés acceptés' de theses.fr
these_id		4,5/5	Identifiant de la thèse. Par défaut, celui de these.fr, format ex. '*2017LIL30011' ou 's54986*'. En cas de non-référencement, proposition du numéro TEL ou HAL.
isNum		5/5	Thèse en lien avec les thématiques du numérique, appartenance au dataset ARCHI-NUM, format Booléen TRUE/FALSE
Abstract		4,5/5	Résumé de la thèse obtenu via these.fr, HAL, ou sur les sites d'ENSA-P.

La Figure 1 résume la démarche d'isolation des datasets évoqués. Elle évoque l'existence de thèses de doctorats reliés à l'Architecture mais hébergés en dehors des ENSA-P. Quoiqu'en dehors du scope de cette étude, cet ensemble étendu compte sans aucun doute une part de doctorats tournés vers le Numérique.

A ce dataset ARCHI-NUM est associé un dossier contenant les manuscrits de 77 thèses en format PDF, obtenus sur la plateforme HAL ou sur demande à leurs auteurs et autrices. Cette bibliothèque de manuscrits pourra servir de support à de l'analyse sémantique automatisée.

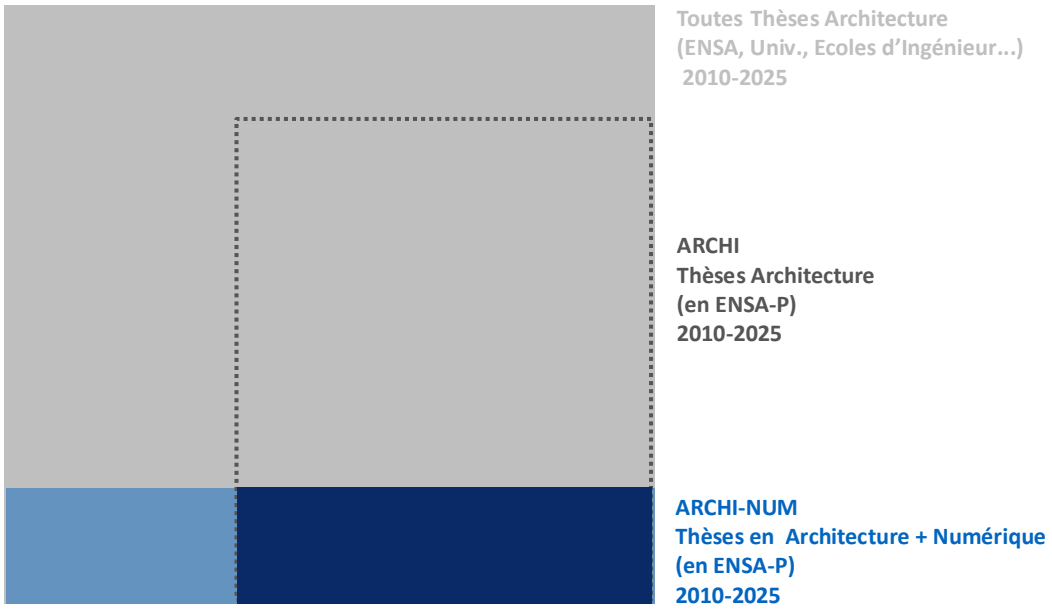


Figure 1 : Thèses de doctorat en architecture entre 2010 et 2025, périmètres et affiliations. Les proportions entre les ensembles thèses ENSA-P et thèses hors ENSA-P ne sont pas basées sur des chiffres connus.

Méthodes : obtention et enrichissement de la donnée

Identifiants et Résumés des thèses.

Pour obtenir un lien direct entre la liste initiale du BER et la donnée déposée sur theses.fr, nous avons utilisé l'Interface de Programmation d'Application, permettant d'envoyer des requêtes automatisées et structurées à la base de données des thèses. Dans cette requête, sont recherchés uniquement les travaux de thèses, pour les auteurs et autrices dont nous donnons le nom et le prénom. L'expérience montre que l'ajout d'autres conditions dans la requête (Titre, date, …) diminue le taux de réponse de l'API sans améliorer significativement la qualité des réponses.

Ce faisant est récupéré l'immatriculation de la thèse dans la base de données, qui est stockée dans l'entrée 'these_id'. A ce jour, l'API ne permet pas de récupérer automatiquement le résumé de la thèse, pourtant disponible sur une page standard (voir Figure 2). Aussi sommes-nous forcés d'utiliser une étape complémentaire de *scrapping*, i.e. une récupération de texte par lecture d'une page HTML en ligne, basée sur l'utilisation de la librairie python populaire *beautifulSoup*.

Figure 2 : Exemple de page « Thèse » sur theses.fr pour le doctorat de Claire DUCLOS soutenue en 2024. Lien consulté le 20/11/2025 : https://theses.fr/2024HESAC001

S'ensuit une étape de nettoyage de la donnée brute, *data cleaning*, pour vérifier la correspondance des auteurs avec leurs thèses, et la récupération des résumés. Pour une partie du corpus, de l'ordre de 5%, les abstracts ont dû être recherché sur des sites tiers, ou directement dans les manuscrits. 17 sur les 648 thèses listées sont encore sans abstracts à ce jour.

IsNum.

L'appartenance d'une thèse au dataset ARCHI-NUM, et donc sa proximité avec les thématiques du Numérique, a été décidée par un opérateur humain, avec relecture de l'ensemble des attribution par un autre opérateur humain. Sur l'ensemble des quelques centaines de lignes du dataset ARCHI, seules les attributions d'une dizaine de thèses ont donné lieu à une discussion.

Genre.

La donnée du genre de l'autrice ou de l'auteur était lacunaire dans le dataset original fourni par le BER. Cette donnée est effectivement difficile à obtenir, en particulier pour des prénoms qui nous sont étrangers en France. Deux stratégies successive pour compléter cette entrée : 1) effectuer une recherche internet, consulter les sites des laboratoires, et identifier les anciens doctorants et doctorantes sur des réseaux comme Linkedin ; 2) utiliser un outil numérique connaissant l'usage genré des prénoms genderize.io. 3) envoyer un mail aux encadrants en cas de doute persistant.

Les manuscrits ont été récupérés sur la plateforme HAL, ou par réponse à une demande écrite aux diplomé.es et à leurs encadrants.

Discussion : limites, pistes

Exhaustivité et fraîcheur de la donnée

Le dataset fourni s'arrête au 1er juillet 2025, et n'a pas été complété par les thèses soutenues depuis. Une discussion devra se tenir sur la façon de **maintenir la donnée**, voire sur une éventuelle stratégie automatisée de suivi des soutenances. Des solutions programmatiques existent pour 'surveiller' les modifications de pages web d'intérêt, que l'on pourrait appliquer sur les pages événement des sites ENSA-P.

Si le dataset présente un grand nombre des thèses de doctorat tenues entre 2010 et 2024, nous avons pu identifier quelques lacunes, en général dans des cas de doctorats co-hébergés par plusieurs UMR. Plutôt que de compléter ces lacunes, à la main, au fur et à mesure de leur découverte, nous avons choisi de garder intact le scope du dataset, et de commencer la construction d'un code de recherche systématique des doctorats en architecture en France. Ce travail dépasse largement le cadre de la commande du Ministère, cette recherche systématique permettant en effet d'**identifier les thèses en Architecture soutenues en dehors des ENSA-P**.

La consolidation du dataset

Dans la mesure du possible, la manipulation du dataset se fait par des lignes de code, c'est le cas pour l'homogénéisation des UMR de rattachement, la récupération de identifiants, etc. La curation de la donnée pratiquée dans le présent rapport reste pourtant partiellement humaine : à quelques reprises, il a été nécessaire de parcourir intégralement le dataset, ligne à ligne, pour détecter et corriger les entrées présentant un défaut : titre exprimé en italien, coquille orthographique dans le titre, espace surnuméraire dans une date ou un nom. Ce mode de curation hybride, qui reste adapté pour la taille de notre dataset (quelques centaines d'entrées) permet 1) de détecter les problèmes ponctuels et de mettre en place des corrections programmatiques si besoin 2) de bien connaître la donnée.

Analyse de la donnée

Résultats : Acteurs et évolutions de la recherche doctorale

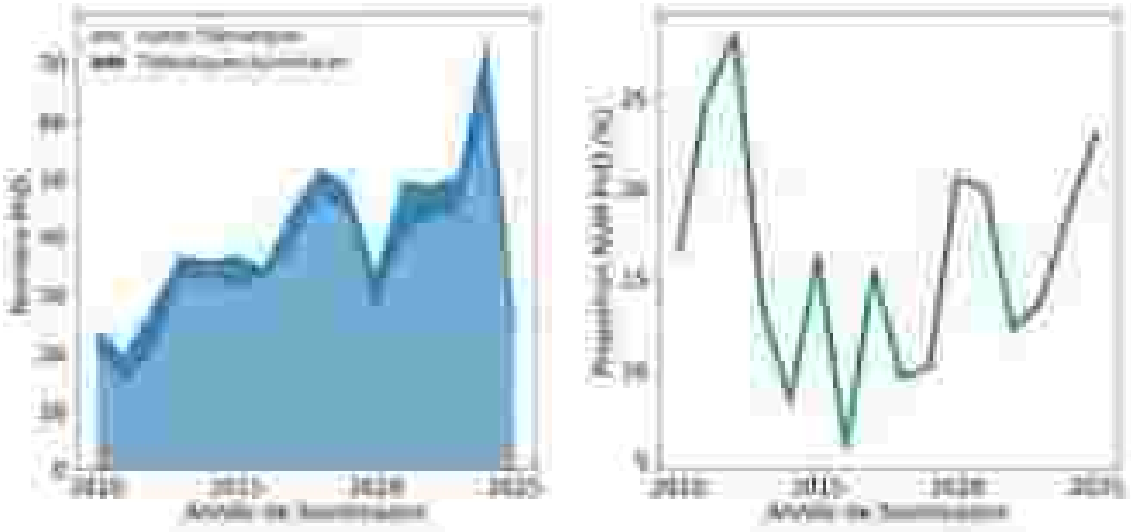


Figure 3 : (A) Evolution du nombre de thèses soutenues en architecture au sein des ENSA-P ; mise en surbrillance des thèses associées aux thématiques numériques ; (B) Proportion de thèses liées aux thématiques du Numérique dans le dataset ARCHI.

Ce premier paragraphe présente quelques analyses de tendance basées sur les datasets ARCHI et ARCHI-NUM. La Figure 3.A montre ainsi l'évolution du nombre de doctorats soutenus dans les quinze dernières années, qui semble avoir triplé sur cette période, abstraction faite du pic négatif probablement lié à la pandémie. La proportion de thèses en lien avec le numérique, présentée Figure 3.B, reste quasi-constante, à 16%±6% sur la période considérée.

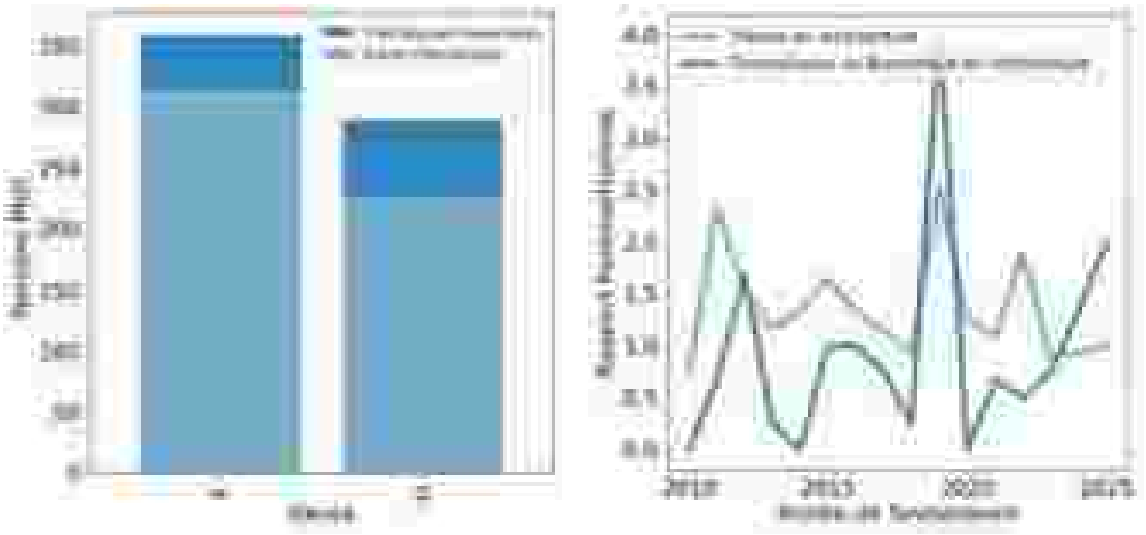


Figure 4 : (A) Genre des docteurs en architecture depuis 2010 et (B) proportion Femme-Homme chez les doctorant·e·s dans les datasets ARCHI et ARCHI-NUM.

La Figure 4 présente les tendances en termes de genre des doctorantes et doctorants. Sur l'ensemble des thèses du dataset ARCHI, on compte 358 docteures (dont 45 sur des thématiques Numériques),

pour 289 docteurs (dont 63 sur des thématiques Numériques). Sur la période considérée, le rapport Femme-Homme est de 1.35± 0.49 pour ensemble des thèses ARCHI, contre 0.93± 0.97 pour les thèses ARCHI-NUM. Cette répartition est cohérente avec celle du public étudiant en ENSA-P, dont 61% sont des femmes (chiffre ministère de la Culture, 2022).

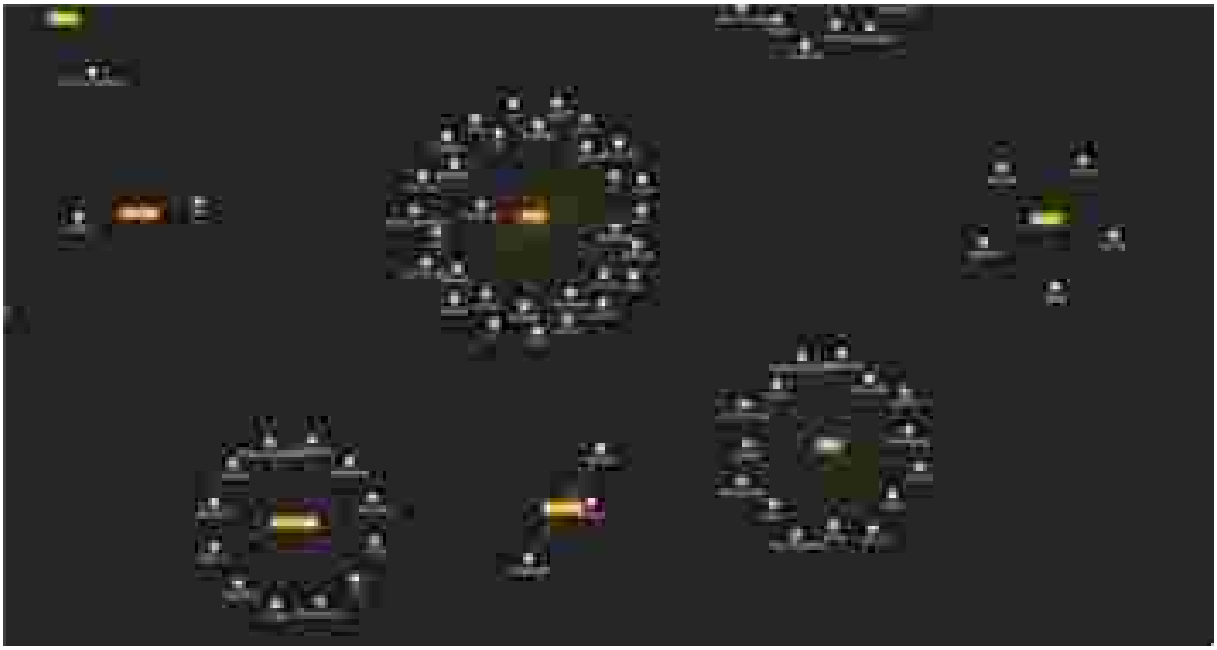


Figure 5 : (A) Acteurs ENSA-P des thèses en lien avec les thématiques du Numérique, associés aux titre des doctorats qui leur sont rattachés. (B) Instantané du Graphe Dynamique des thèses ARCHI-NUM et de leurs ENSA-P associées. Disponible en ligne sur le lien : https://accn.archi.fr/cartographie-doctorale

Les Figure 6 et Figure 6 présentent les ENSA-P de rattachement des doctorats du dataset ARCHI-NUM, présentée sous forme de liste ou de graphe dynamique. L'ENSA de Nantes et son laboratoire AAU-CRENEAU (UMR 1563) sont les premiers producteurs de thèses en lien avec le Numérique (% des

thèses), suivis par les ENSA-P de Paris-Malaquais et de Nancy (respectivement XX et YY% des thèses soutenues depuis 2010). Cette analyse peut descendre à la granularité des laboratoires. A cette fin, nous nous sommes basés sur le document du BER 'Liste 2025 Unités de recherche et responsables' pour homogénéiser la liste des UMR de rattachement des doctorats. Dans le corpus ARCHI-NUM, 10 des 14 ENSA-P représentées ne le sont que par une seule UMR, aussi ne proposons-nous pas ici de tableau des UMR, jugé redondant.

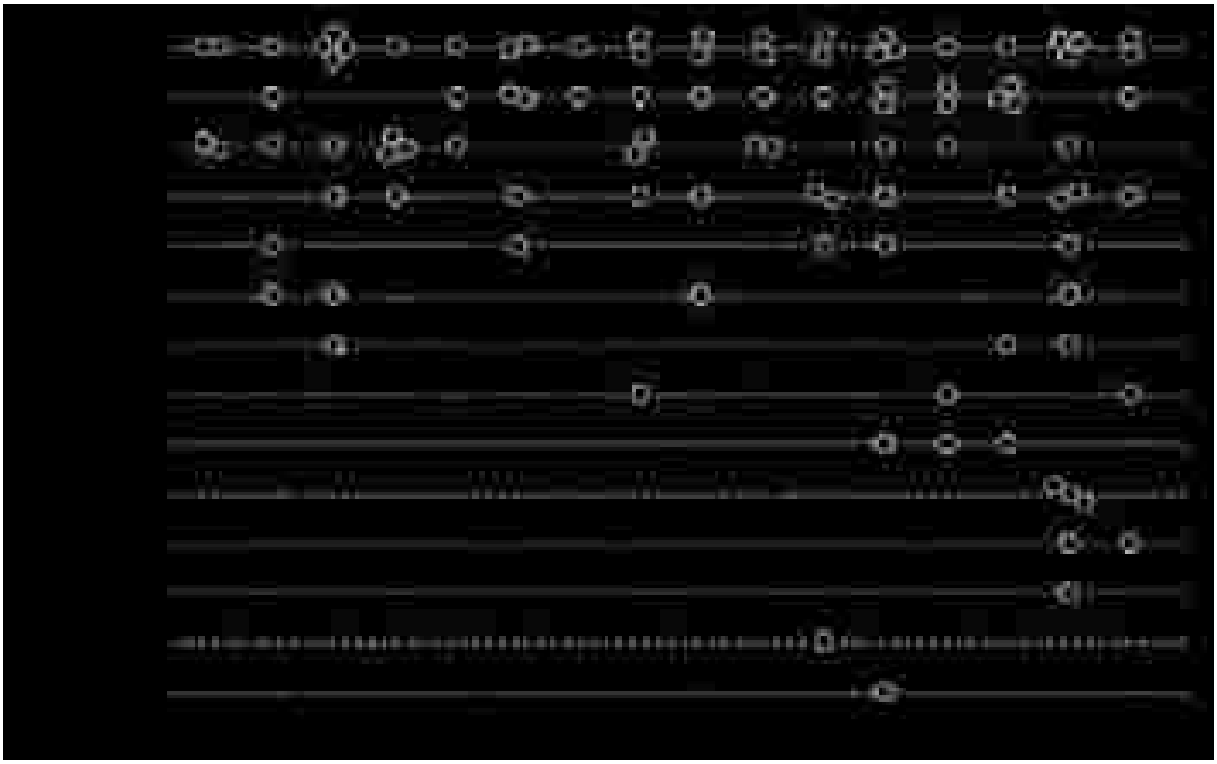


Figure 6 : Production de thèses en lien avec le Numérique Acteurs ENSA-P en fonction du temps.

Matériel et Méthode

L'ensemble des analyses statistiques est réalisé en langage python. Les visuels sont produits par la librairie python *Matplotlib* associée à la librairie *pandas,* ou par la plateforme en ligne *RawGraph*.

Discussion

Le choix a été fait de ne pas utiliser les histogrammes pour représenter les acteurs académiques de la recherche (ED, ENSA-P, laboratoires), afin d'éviter une « podium-isation » tout à fait factice : chaque acteur a ses domaines de recherches, hérités de leurs orientations, politiques, prérogatives, et de l'intérêt des personnes pour telle ou telle thématique liée à l'architecture. Reste que nos visualisations invisibilisent les ENSA-P où ne sont menées aucunes recherches liées au numérique, comme celle de Clermont-Ferrand, pourtant orientée vers des thématiques primordiales de rénovation énergétique ou de gestion des matériaux.

- De l'intérêt d'aller voir hors ENSA

Sur le sujet des acteurs de la recherche doctorale, une analyse similaire à celle menée à l'échelle des ENSA et des UMR devrait pouvoir être conduite à l'échelle des Ecoles Doctorales. Un obstacle majeur réside dans le caractère impermanent de leur dénomination et de leur accréditation. Un exemple, l'ED 498 « Sciences pour l'Ingénieur, Géosciences, Architecture », auxquelles 13% des thèses du dataset

sont rattachées, a cessé son activité, et n'apparaît même pas sur la base de donnée gouvernementale Liste des écoles doctorales accréditées.

Notre étude se restreignant aux ENSA-P, la courbe de tendance des thématiques numériques décrit un panorama probablement incomplet de la recherche doctorale française en *Digital Architecture*. Des travaux préliminaires ont commencé dans le réseau ACCN pour tenter de capter l'ensemble des thèses produites en architecture et y identifier les thèses en lien avec le Numérique. Ces travaux dépassent largement le cadre de la commande, à la fois par son périmètre et son volume, mais des études préliminaires sont en cours pour automatiser la récupération de thèses d'architecture pertinentes.

Analyses textuelles et sémantiques

Mots clés et leurs occurrences

Une approche frugale de l'analyse de texte liée à la recherche doctorale : observer l'occurrence de différents mots-clés, en particulier de ceux connotés 'Numérique'. Cette analyse, pratiquée à l'échelle d'un résumé de thèse permet d'en embrasser assez directement l'orientation thématique. Il devient presque possible, avant même d'aborder la partie suivante, d'étiqueter d'un seul regard une thèse avec sa thématique principale : dans le cas de la Figure 7, les thèses de Nicolas BIORET et de Hana REZGUI seraient peut-être labellisées « Données SIG » et « projet BIM ».

Figure 7 : Nuages de mot associés aux résumés de deux thèses du corpus, celle de Nicolas BIORET (2010) et celle de Hana REZGUI (2024).

Plus intéressant peut-être : pratiquer cette analyse à l'échelle d'un sous-ensemble des données. On propose ici (Figure 8) une comparaison sur plusieurs UMR de rattachement, mais un exercice similaire pourrait être mené sur d'autres types de communautés : quels mots clés utilisés dans les résumés thèse en 2020 ? à l'ENSA-P de Montpellier ?

Figure 8 : Nuages de mot associés à l'ensemble des résumés produits entre 2010 et 2025 à (gauche) et Bordeaux (droite).

Cette approche reste purement qualitative et visuelle. A notre connaissance, la littérature ne présente pas de méthode quantitative basée sur la visualisation en nuages de mots. Pour cause : ces analyses peuvent être fortement biaisées par la volubilité des doctorants, un long résumé avec des redondances pesant fortement sur la mesure donnée par un nuage de mots. Reste que de telles visualisations sont une piste à étudier pour permettre la consultation rapide des documents, chaque mot-clé pouvant être rendu interactif modulo un travail de design d'interface.

Outre la visualisation discutable, l'approche par mot-clé peut être rendue quantitative, et un suivi peut être mené à l'échelle du dataset. Sur la Figure 9, on représente ainsi l'apparition de quelques mots-thématiques dans les résumés de thèse. Pour s'affranchir du biais de volubilité évoqué plus haut, une thèse qui mentionnerait n fois un mot compte pour 1 thèse mentionnant ce mot. Pour tenir compte du caractère non-ponctuel du doctorat, qui porte une thématique sur toute sa durée, nous proposons aussi une visualisation lissée, qui montre par exemple la montée des thématiques du BIM et de la réalité virtuelle ou augmentée sur les dix dernières années.

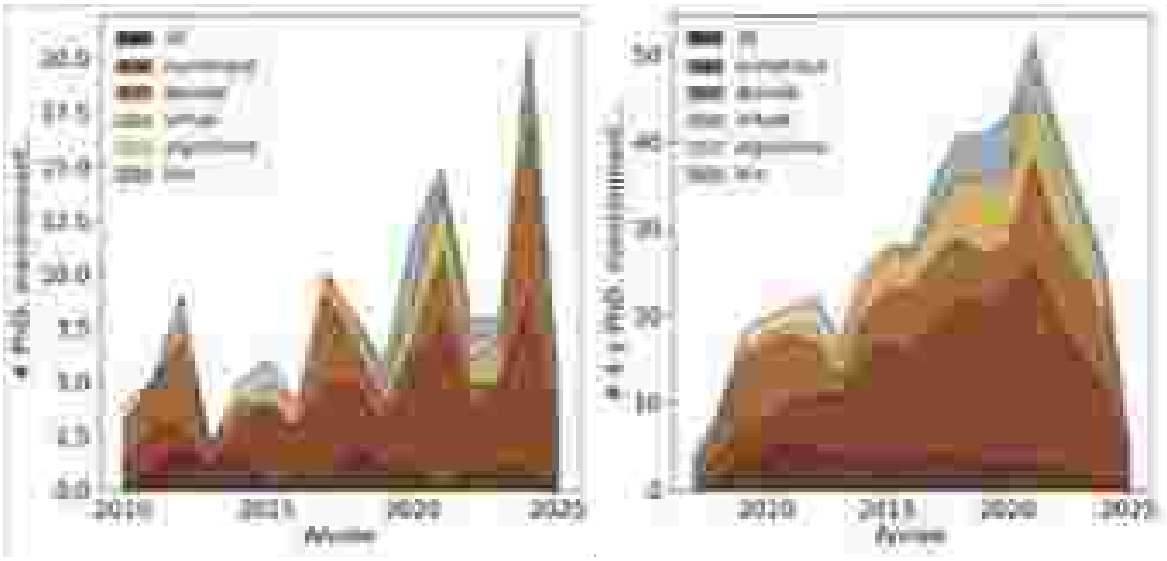


Figure 9 : Occurrences de mots clés dans les résumés de thèse. (Gauche) découpage par année, un résumé qui contient le mot compte comme +1 sur son année de soutenance. (Droite) découpage par période de 4 ans, un résumé qui contient le mot compte comme +1 sur sa période, la période étant définie comme [(Année de Soutenance – 4 ans) – (Année de Soutenance)].

Labellisation thématique supervisée

Il apparaît intéressant de pouvoir étiqueter, labéliser les thèses du corpus, pour pouvoir effectuer des rapprochements thématiques, et dépasser l'échelle d'analyse de la thèse unique. Si cette labélisation peut se faire ligne à ligne, par l'humain, elle pose deux questions : 1) comment s'affranchir des biais de perception des thèses par les opérateurs humains qui les labélisent ? 2) comment choisir les labels pour couvrir l'ensemble des thématiques de façon égale ? 3) comment généraliser l'approche à des datasets de taille plus conséquente ?

Dans un premier temps, on se propose d'attacher à chaque ensemble {Titre + Résumé} l'un des labels listés ci-après, supposés couvrir et baliser l'ensemble des travaux en lien avec le Numérique pour l'Architecture. Cette liste est une proposition de structuration, basée sur la lecture des grands axes de conférences spécialisées (SCAN, EduBIM), de documents de cadrage [CMA ARCHI (2023)], ou de livres sur la thématique du Numérique pour l'architecture [Atlas of Digital Architecture, Ed. Birkhauser, 2020].

Tableau 2 : liste de labels pour classifier les thèses en Numérique pour l'Architecture.

Label pour la classification supervisée	Commentaires
Conception paramétrique, conception générative (IA)	
Fabrication, matériaux et construction robotisée	
Réalité virtuelle et augmentée	
Cartographie, système d'information géographique (SIG)	
Maquette numérique (BIM) et simulation	
Patrimoine et maquette numérique (HBIM)	
Théorie des média et des études numériques	

Sur la Figure 10, on reprend ainsi l'affichage des ENSA-P de rattachement des doctorats vu plus haut, après attribution d'une couleur thématique à chaque doctorat. Au-delà des acteurs historiques de Nantes, Paris-Malaquais et Nancy, il est intéressant de remarquer, depuis 2021, une diversification des acteurs ENSA-P impliqués dans les doctorats ARCHI-NUM, avec l'apparition dans la liste des écoles de Montpellier, Strasbourg, Lille, Grenoble ou Toulouse.

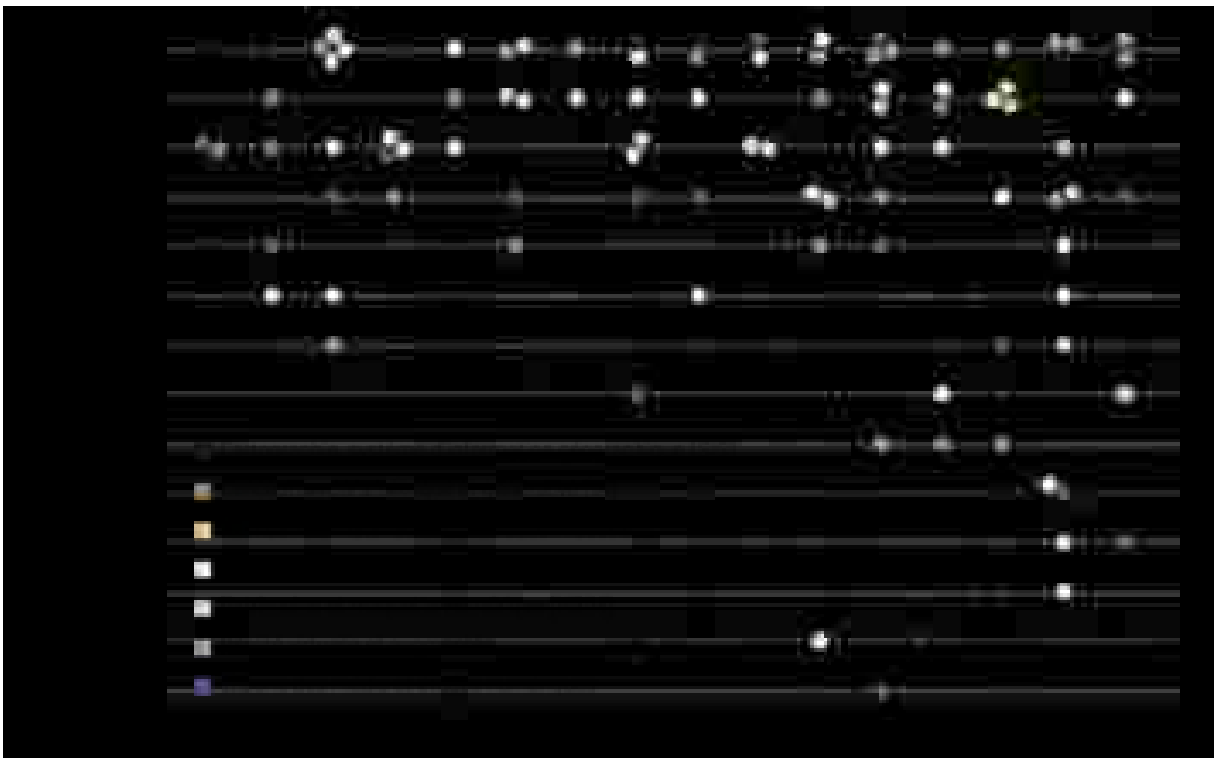
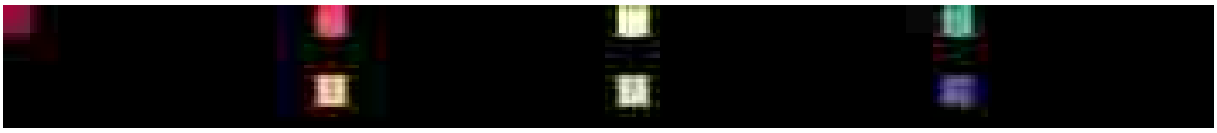


Figure 10 : Représentation des ENSA-P avec coloration des thèses selon leur appartenance thématique, déterminée par une labellisation « zero-shot ».



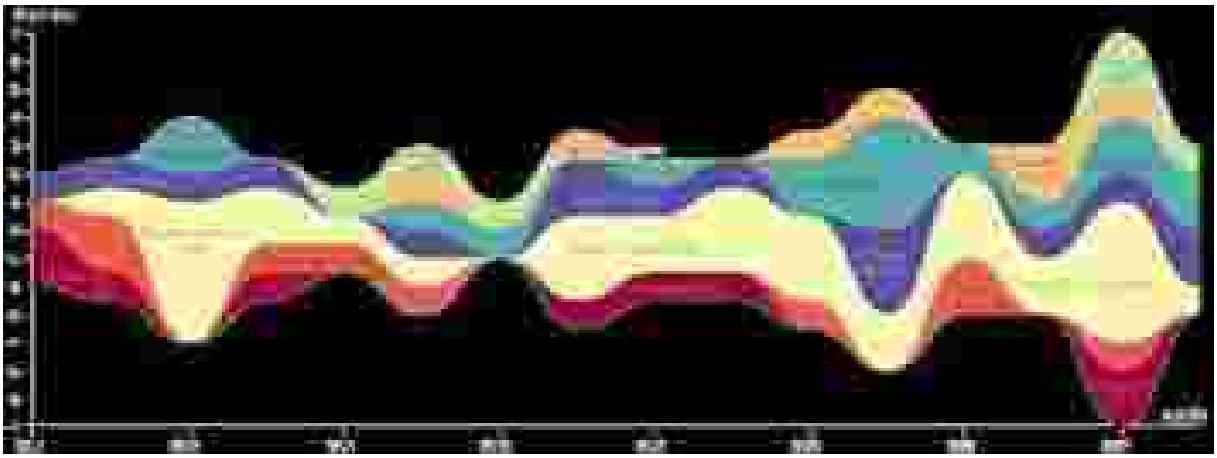


Figure 112 : Part des différentes appartenances thématiques, déterminées par une labellisation « zero-shot » des abstracts.

L'expérimentation de labellisation supervisée a aussi été l'occasion de tester la qualité de la variable « *isNum ? » :* Les cent thèses du corpus ont ainsi été passées à un LLM, en demandant « is this PhD related to architecture ? (true or false)", sur la base des titres, ou des titres rallongés par la liste des mots clés de la thèse. Les résultats sont particulièrement satisfaisant, la machine détecte peu de faux positifs, et atteint une Précision de 92.3% (titres uniquement) et de 88.9% (titres et mots-clés). En termes de Rappel (*recall*), les scores sont plus mitigés (54% titres ; 67% titres et mots-clés) : la machine échoue à identifier certaines thèses en lien avec le numérique, notamment pour des raisons de titres vagues, ou faute de connaître l'abstract pour vérifier son assomption.

Labellisation thématique non-supervisée

La méthode précédente proposait un étiquetage des thèses selon une grille prédéfinie par le comité de pilotage ACCN. En ce sens, cette méthode contient une part d'arbitraire, et peut être biaisée par des connaissances incomplètes, des préférences thématiques, des visions différentes du domaine considéré. Plus important, elle repose sur la capacité d'un groupe à atteindre un consensus sur la structure de l'ontologie à suivre.

On propose ici une autre démarche de classification un corpus de texte, schématisée par la Figure 12, qui **rapproche les résumés selon leur proximité sémantique, puis qui fait émerger des thématiques à partir des mots retrouvés dans un groupe de résumés proches.** Une étape de réduction de dimension permet de faciliter l'identification et la visualisation de ces regroupements. Il est à noter que cette approche, en particulier lors de la projection en dimension 2, demande de trouver un compromis de paramètres, pour parvenir à séparer les vecteurs sans pour autant les faire s'effondrer les uns sur les autres.

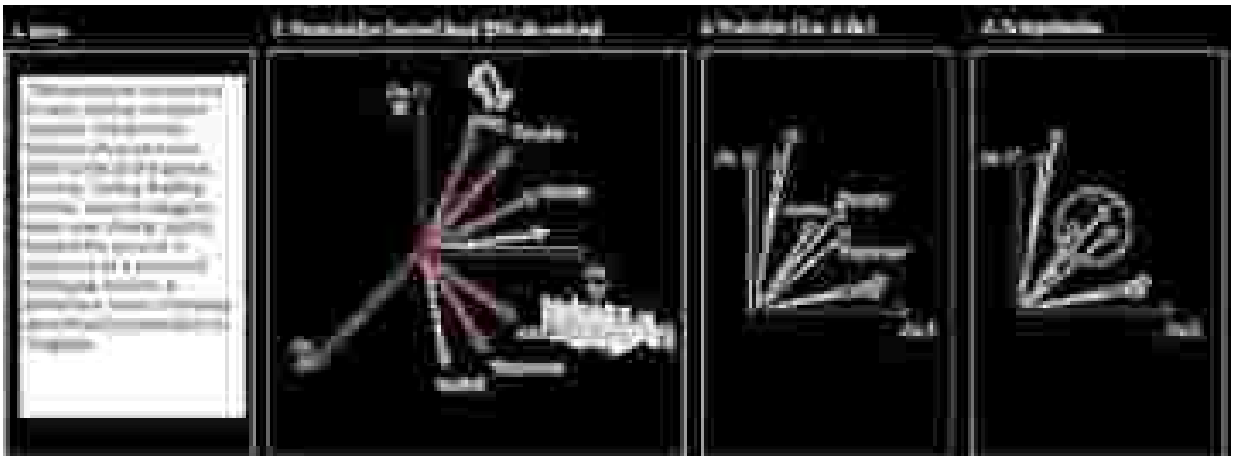
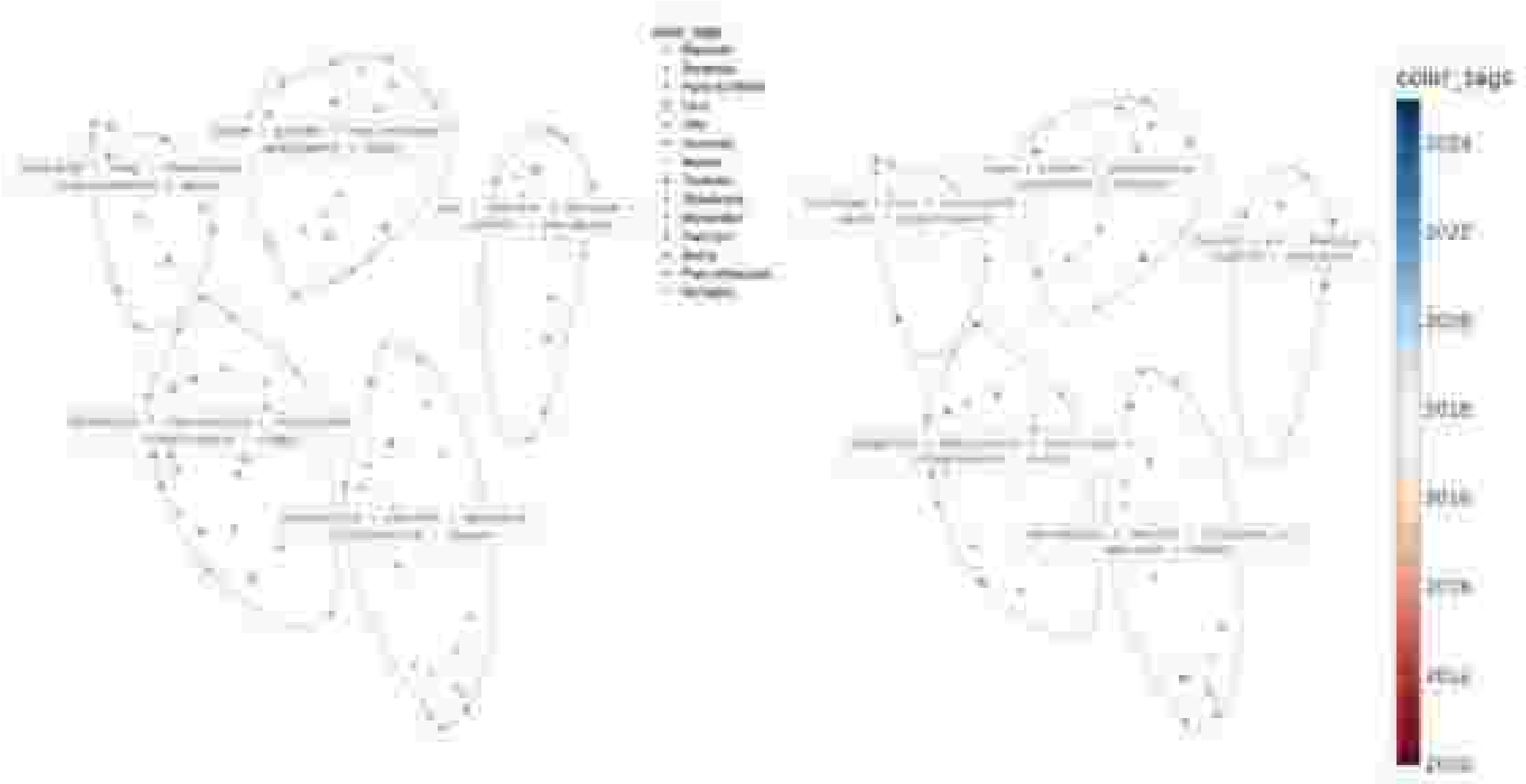


Figure 12 : Processus de labellisation automatisée, pour passer d'un texte (1), à des vecteurs dans un modèle de langue (2), que l'on peut projeter (3), pour identifier un cluster (4). Ici, l'exemple est fourni sur un texte issu de la biologie, où sont repérables des mots décrivant des animaux, des écosystèmes et des actions liés au mouvement.

Un exemple de résultat de cette démarche est montré en Figure 13, sur laquelle sont représentée les différentes thèses, projetées avec l'algorithme UMAP (min_dist = 0.2 ; n_neighbors = 5), puis regroupées par clustering agglomératif (n_cluster = 5). Se dégagent cinq clusters, étiquetés par des mots trouvés dans les abstracts des thèses de chaque cluster. Les résumés de thèse étant écrits en franaçais, il y a fort à parier que les occurrences de «'building » soient associables à *smart building* ou à *Building Information Modelling.* Dans le sens de lecture, nous pouvons inférer que ces clusters regroupent respectivement des thèses sur le BIM, sur la notion de projet, sur les performance environnementales, sur les structures géométriques, et sur les ambiances numériques. Une colorisation en fonction de la date de soutenance permet de discerner des travaux relativement récents dans le cluster Géométrie-Structure comparé au cluster Projet. Est également détectable la présence quasi exclusive des ENSA-P de Nantes et de Versailles dans le cluster Ambiances Numériques. L'ENSA-P de Nancy est sur-représentée dans le cluster Géométrie-Structure, celles de Paris-La Villette et Paris Est ressortent particulièrement dans le cluster Projet.

Figure 13 : Représentations projetées et clusterisées des abstracts de thèse du corpus ARCHI-NUM, colorisation selon la date de soutenance. Ce graphe est disponible dans une version dynamique au format HTML, dans laquelle le survol d'un point affiche le titre et l'année de soutenance de la thèse.



Matériel et Méthodes

L'analyse par mots clés et la production de nuages de mots utilisent la librairie *Wordcloud* de python. Conformément à l'usage, ont été exclus des analyses les *stopwords* suivants : ['d', 'du', 'de', 'la', 'des', 'le', 'et', 'est', 'elle', 'une', 'en', 'que', 'aux', 'qui', 'ces', 'les', 'dans', 'sur', 'l', 'un', 'avec', 'pour', 'par', 'il', 'nous', 'ou', 'à', 'ce', 'a', 'ont', 'sont', 'cas', 'plus', 'leur', 'se', 's', 'vous', 'au', 'c', 'aussi', 'toutes', 'autre', 'comme', 'cette', "d'un", "d'une", 'mon', 'ton', 'son', 'nos', 'notre', 'vos', 'leurs', 'deux'].

L'analyse sémantique supervisée utilise le modèle open source « LLM GPT-OSS 20b-cloud » disponible via la plateforme Ollama (https://ollama.com/library/gpt-oss:20b-cloud). Le prompt utilisé pour l'attribution des étiquettes est donné en Annexe.

L'analyse sémantique non-supervisée suit les procédures de la librairie BunkaTopics développée sur python, et utilise le modèle de langage de taille intermédiaire basé sur BERT « *all-MiniLM-L6-v2* » disponible sur la plateforme Huggingface (https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2). Les algorithmes utilisés pour la projection (UMAP, tSNE) et la Clusterisation (K-Means, HDBSCAN, Agglomerative Clustering) sont tirés des librairies python du même nom et de la librairie *sklearn*. Le graphe obtenu sur la Figure 13 correspond au choix de paramètres suivants :

Il a peut-être été remarqué que les modèles de langue choisis sont modestes en taille (BERT, GPT-20b). Il serait pourtant très simple d'insérer un appel à un modèle GPT dans le processus de Bunka Topics, ou de demander une labellisation supervisée à un LLM classique. Nous suivons en ce sens les recommandations de frugalité émises par de nombreux acteurs du numérique responsable, comme Data4Good ou l'Ecolab du Commissariat Général du Développement Durable (5), face aux impacts encore mal connus de l'IA générative (6).

Discussion

Limite immédiatement perceptible des approches de labellisation proposées : l'unicité du label obtenu. Par construction, a été demandée une étiquette unique, alors que certaines thèses se placement probablement à l'interface de plusieurs des labels du Tableau 2. Un proposition de processus, décrit en annexe, donne une piste pour décrire chaque abstract avec plusieurs labels, en se basant sur un calcul simple de proximité sémantique (*cosine similarity*).

L'une des approches, non présentée ici, consiste à proposer aux auteurs de l'étude une application en ligne pour labelliser les thèses à partir d'une ontologie prédéfinie. Cette labellisation *humaine*, si elle permet un degré supérieur de certitude sur la qualité de la donnée, reste consommatrice de temps pour les opérateurs, et demande un travail de mise en ligne et de gestion du consensus à anticiper. Nous la mentionnons ici à titre d'information, et montrons sur la Figure une maquette d'interface conçue pendant la mission. Elle pourrait être applicable pour des jeux de données plus petits, ou pour des thèses difficilement discriminables d'un point de vue sémantique. Des solutions payantes comme Rayyan sont également une piste à explorer – sous réserve de pouvoir fournir ou atteindre un corpus exclusivement composé de thèses.

Figure 14 : (A) Maquette de *TindArchi*, application de classification rapide des thèses, (B) application commercialisée *Rayyan* pour trier un corpus documentaire d'un simple geste de balayage d'écran.

Le travail de regroupement thématique peut encore être approfondi. A ce stade, nous représentons, en deux dimensions, des clusters de doctorats labellisés selon une liste de mots clés identifiés comme représentatifs. Le rajout d'une dimension - et donc la représentation de clusters en volume - pourrait par exemple mettre en évidence d'autres relations entre thèses, de dégager d'autres tendances. En guise de piste, nos derniers efforts ont porté sur une meilleure **explicitation des thématiques de chaque cluster**, via l'intégration d'un modèle de langage avancé dans le processus d'analyse. Sur la Figure 15, la librairie *BunkaTopics* a demandé au modèle GPT de générer des titres de cluster, sur la base des mots clés et des résumés associés à chaque cluster. Le prompt utilisé dans la librairie *BunkaTopics* est explicité en Annexe.

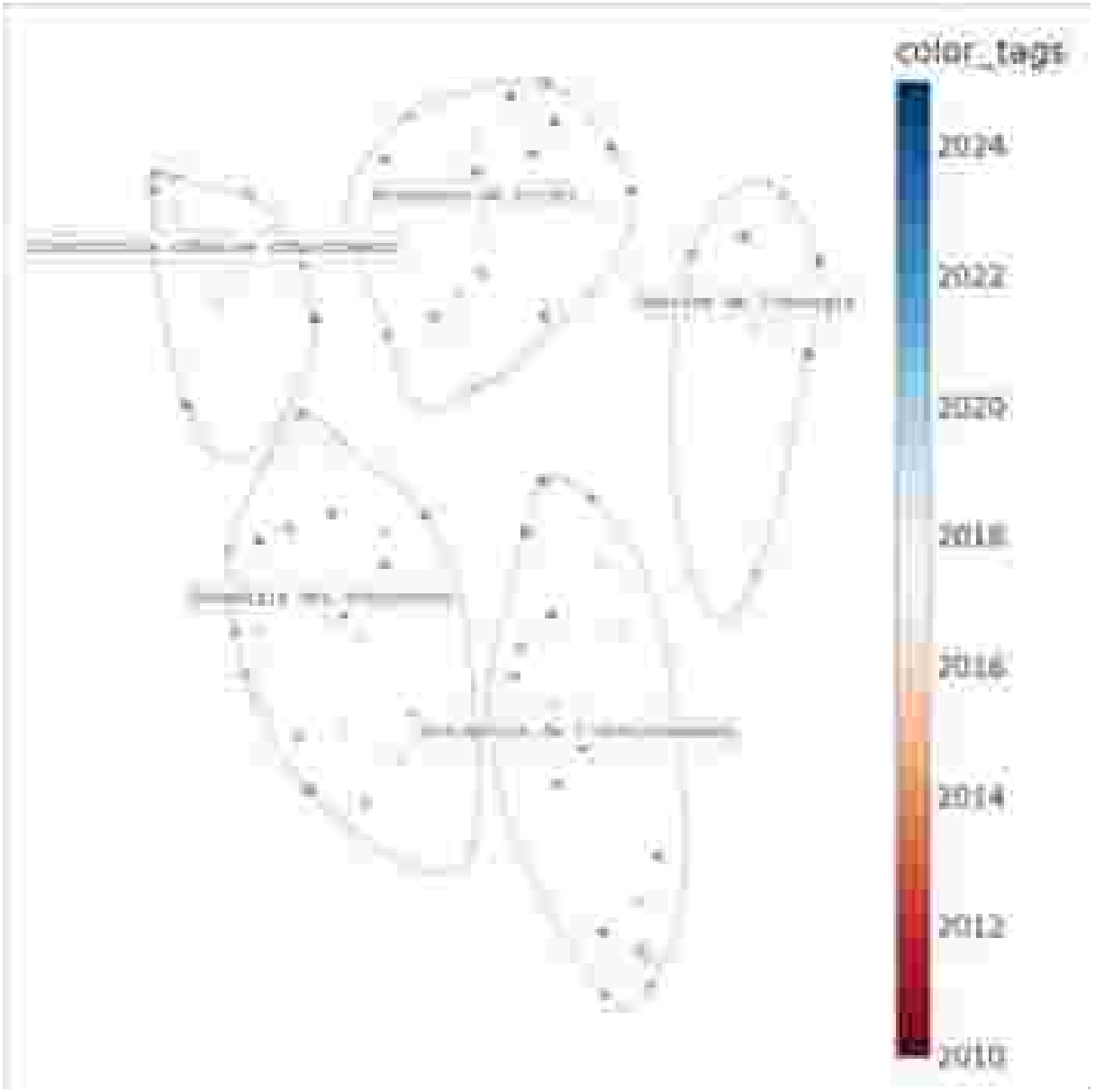


Figure 15 : Labellisation des clusters assistée par GPT. Les clusters d'origine sont ceux décrits Figure 10Figure 13. Le cluster Processus de Projet a été renommé manuellement (à l'origine : « Progrès du Projet »).

Ce travail de caractérisation sémantique pourrait en outre servir de base à un travail de veille, d'observatoire du travail de recherche doctorale. L'étiquetage des doctorats apparaît alors comme une **clé d'identification des thématiques phares, ou au contraire des parents pauvres de la recherche,** et pourrait ainsi permettre d'orienter des politiques de soutien, de flécher des financements, de justifier des appels à manifestation d'intérêt ou le montage de chaires spécialisées.

D'un point de vue strictement académique, les outils d'analyse développés pourraient également permettre un suivi d'un sujet dans le temps, en déclenchant par exemple une alerte lors de la sortie d'une thèse mentionnant 'réalité virtuelle' ou 'HBIM', facilitant d'autant l'écriture d'un état de l'art ou d'un travail de revue.

Outils de consultation et de visualisation

Graphe interactif en ligne

L'isolation des thèses du corpus ARCHI-NUM et la structuration de la donnée ayant pris un temps conséquent, il apparaît intéressant de proposer cette donnée à la consultation, dans le respect des politiques de protection des données personnelles. On propose un démonstrateur simple, sous forme d'un graphe dynamique, pour permettre la consultation du corpus accessible via une URL. Ce graphe est une simple page HTML générée par la librairie python *networkx.* Les images fournies en illustration sont directement tirées des manuscrits, et voulues les plus représentatives possibles des thématiques abordées par le doctorat.

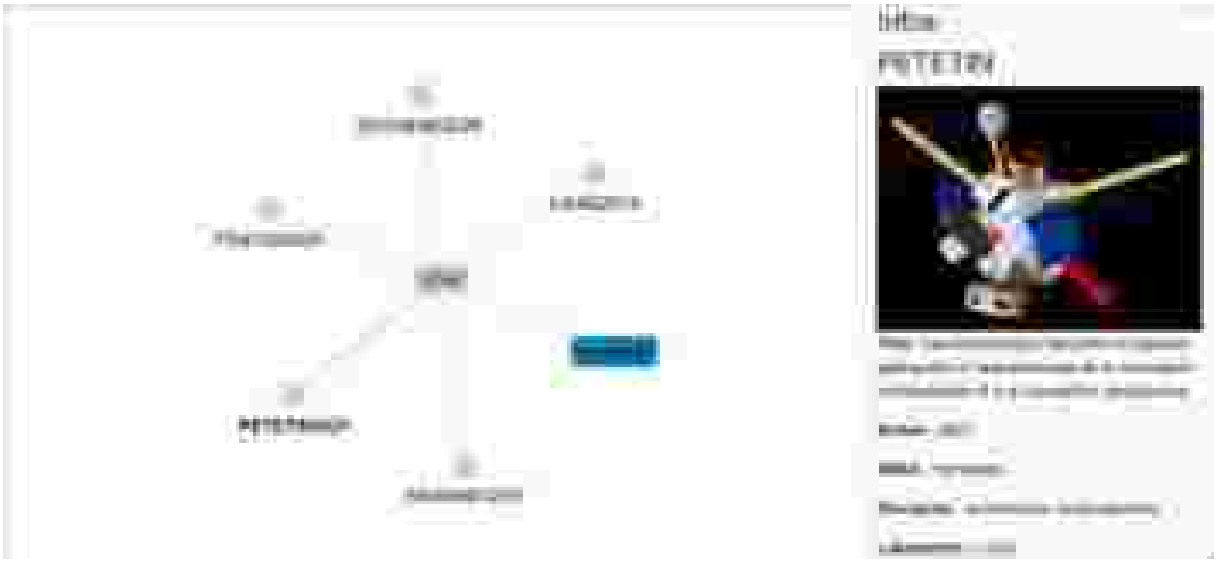


Figure 16 : Capture d'écran de l'interface de consultation, disponible sur HTTPS://ACCN.ARCHI.FR/CARTOGRAPHIE-DOCTORALE

Interface de discussion : ACCNbot

Parmi les propositions d'interface utilisateur, celle de l'agent conversationnel *Chatbot* pourrait permettre une consultation facilité de la recherche doctorale sur la thématique du Numérique en Architecture.

Pour qui a l'habitude de chercher de la documentation scientifique sur un moteur comme Google Scholar, il est difficile d'isoler un travail de doctorat, a fortiori français, a fortiori sur des sujets traités dans des articles de revue, qui sont priorisés par l'algorithme du fait de leur nombre de citations. On peut ainsi essayer une requête du type « maquette numérique thèse france architecture jacquot » sans trouver trace du manuscrit de K. Jacquot (2014). A l'inverse, l'ACCNbot, présenté sur la Figure 17 répond à la question « que lire sur les maquettes numériques » par une liste de trois travaux doctoraux directement reliés au sujet.

Figure 17 : Capture écran du prototype d'agent conversationnel ACCNbot, développé sur la base du modèle Mistral. La partie gauche de l'écran montre les consignes fournies, la partie droit une discussion test sur la base de la question « que lire sur les maquettes numériques ? ».

Cette première expérimentation reste à état un état exploratoire et deux initiatives sont en cours. L'une envisageant la spécialisation d'un modèle fondamental libre, interfacé avec OpenWebUI et dont l'hébergement reste à identifier. L'autre est envisagée dans le cadre de l'expérimentation Assistant IA de la Direction du Numérique (DiNum) et du Ministère de la Culture. Les prochaines fonctionnalités proposées par la DINUM sur cet assistant devraient permettre une évaluation en grandeur.

Par ailleurs, nous pourrions aussi donner davantage de connaissances antérieures à l'assistant. On pense par exemple aux manuscrits des thèses, sous réserve de garantie suffisante sur la protection des données par le modèle support. L'interface permettrait alors de dialoguer directement *avec* une thèse donnée, de consulter ses figures, d'aller directement au chapitre d'intérêt, bref de s'imprégner de son contenu autrement qu'en feuilletant un document de dimensions parfois conséquentes (voir Figure 18).

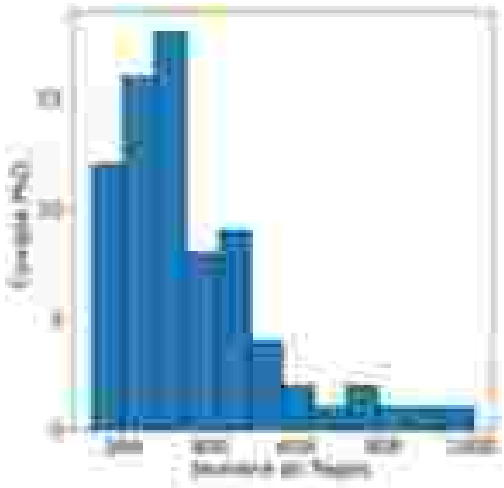


Figure 18 : Tailles de 75 des 100 manuscrits de thèses du corpus ARCHI-NUM.

Annexes et Scripts

Requêter l'API de these.fr et obtenir le résumé

```
# construire la requête
# url_request = 'https://theses.fr/api/v1/theses/recherche/?q=' + df['Nom'][these_num] + '+' +
df['Prénom'][these_num]   +   '+'   +   df['Titre   de   la   thèse'][these_num].replace('é',
'e').replace('’', ' ').replace(' ', '+') + '&debut=0&nombre=1&tri=pertinence'
url request = 'https://theses.fr/api/v1/theses/recherche/?q=auteursNP:(' + df['Nom'][these num]
+'+'+ df['Prénom'][these_num] + ')&debut=0&nombre=3&tri=pertinence'

print(url_request)
```

>> https://theses.fr/api/v1/theses/recherche/?q=auteursNP:(YUAN+Huang)&debut=0&nombre=3&tri=pertinence

```
# récupérer les abstracts pour compléter le tableau des docteurs
#                                        url_request                                        =
"https://theses.fr/api/v1/theses/recherche/?q=modelisation+simulation+microclimat+urbain+etude
+impact+morphologie&debut=0&nombre=2&tri=pertinence"
response = requests.get(url_request)
response.json()['theses'][0]
```


>> {'id': '2010ECDN0010',
'titrePrincipal': 'Methodology of climatic design of urban district for buildings energy efficiency',
'titreEN': 'Methodology of climatic design of urban district for buildings energy efficiency',
'etabSoutenanceN': 'Ecole Centrale de Nantes',
'etabSoutenancePpn': '03063525X',
'dateSoutenance': '01/01/2010',
'datePremiereInscriptionDoctorat': None,
'auteurs': [{'ppn': '256014493', 'nom': 'Huang', 'prenom': 'Yuan'}],
'directeurs': [{'ppn': '19848786X', 'nom': 'Hégron', 'prenom': 'Gérard'},
{'ppn': '168613174', 'nom': 'Li', 'prenom': 'Baofeng'},
{'ppn': '12252568X', 'nom': 'Musy', 'prenom': 'Marjorie'}],
'rapporteurs': [{'ppn': '033866139', 'nom': 'Inard', 'prenom': 'Christian'},
{'ppn': '169681386', 'nom': 'Teller', 'prenom': 'Jacques'}],
'examinateurs': [{'ppn': '19848786X', 'nom': 'Hégron', 'prenom': 'Gérard'},
{'ppn': '168613174', 'nom': 'Li', 'prenom': 'Baofeng'},
{'ppn': '12252568X', 'nom': 'Musy', 'prenom': 'Marjorie'},
{'ppn': '033866139', 'nom': 'Inard', 'prenom': 'Christian'},
{'ppn': '169681386', 'nom': 'Teller', 'prenom': 'Jacques'}],
'president': {'ppn': None, 'nom': None, 'prenom': None},
'nnt': '2010ECDN0010',
'discipline': 'Ambiances architecturales et urbaines',
'status': 'soutenue',
'ecolesDoctorale': [{'ppn': '128696265',
'nom': "École doctorale Sciences pour l'ingénieur, Géosciences, Architecture (Nantes)",
'type': None}],

'partenairesDeRecherche': [{'ppn': '026390310',
'nom': 'Laboratoire central des Ponts et Chaussées (France ; 1949-2011)',
'type': 'Laboratoire'}],
'sujets': [{'langue': 'fr', 'libelle': 'Microclimat urbain'},
{'langue': 'fr', 'libelle': 'Morphologie urbaine'},
{'langue': 'fr', 'libelle': 'Efficacité énergétique des bâtiments'},
{'langue': 'fr', 'libelle': 'Typologie de formes urbaines'},
{'langue': 'fr', 'libelle': 'Indicateurs morphologiques'}],
'sujetsRameau': [{'ppn': '178479934',
'libelle': "Constructions -- Économies d'énergie"},
{'ppn': '145909476', 'libelle': 'Urbanisme durable'},
{'ppn': '029459370', 'libelle': 'Microclimat urbain'}]}

```
# procéder au scrapping du résumé via l'identifiant
url_these='https://theses.fr/'+response.json()['theses'][0]['id']
print(url_these)
response these = requests.get(url these)
html = response_these.content
soup = bs(html, "lxml")
abstract = soup.find("div", id="resume-text").get_text()
print(abstract)
```

Requêter un LLM pour obtenir une classification supervisée (prompt)

« Lis le fichier data1.csv. Celui-ci décrit des documents avec 4 colonnes :

- *ID*
- *Titre*
- *Mots-clés*
- *Résumé*

Voici les 7 catégories à utiliser pour classer chaque ligne du fichier :

- *Conception paramétrique, conception générative (IA)*
- *Fabrication, matériaux et construction robotisée*
- *Réalité virtuelle et augmentée*
- *Cartographie, système d'information géographique (SIG)*
- *Maquette numérique (BIM) et simulation*
- *Patrimoine et maquette numérique (HBIM)*
- *Théorie des médias et des études numériques*

Objectif : Associer chaque ligne du fichier data1.csv à l'une de ces 7 catégories. Propose un nouveau fichier CSV contenant uniquement deux colonnes :

- *ID*
- *Catégorie*

Remarque : Le fichier data1.csv est découpé en fichiers de 25 lignes. Chaque ligne contient le titre et les mots-clés de la thèse. »

Partie 2. Addendum : Elargissement de l'analyse aux doctorats en cours et aux doctorats conduits au sein des laboratoires universitaires

Introduction

Cette seconde partie du rapport présente les résultats de la mission de bibliométrie, commandée par le réseau Architecture Conception et Culture Numérique (ACCN) au 1er janvier 2026. Il s'agit d'un complément à l'étude bibliométrique sur les travaux doctoraux en architecture menés entre 2010 et 2025 au sein des ENSA, commandée aux différents réseaux RSPA par le BER du ministère de la Culture en 2025.

Cet addendum se concentre d'abord sur l'étude des travaux doctoraux en cours au sein des ENSA. Issu d'un travail de recensement et de veille du BER, le dataset étudié permet d'obtenir un « instantané », de la recherche doctorale en architecture au 1er octobre 2025, instantané qu'il serait intéressant de pouvoir actualiser régulièrement pour lui conserver sa fraicheur.

Dans la continuité du travail demandé aux réseaux, le rapport propose également un élargissement du scope de l'étude de cartographie, en recherchant sur l'intégralité de la base Thèses.fr des travaux doctoraux relatifs à l'Architecture et liés à la culture du Numérique. Nous proposons une méthode de récolte de donnée, de classification semi-supervisée, et un pipeline d'analyse qui pourrait être réutilisé sur un thème différent de celui du Numérique pour l'Architecture.

Une partie des méthodes et outils numériques utilisés ont été décrits dans le premier volet de l'étude, et seront rappelés sous forme de renvois. Comme pour ce premier volet, l'usage des outils numériques génératifs a été strictement réservé aux analyses bibliométriques de l'étude, et tout le texte de ce rapport a été écrit et corrigé humainement.

Nous adressons une fois encore nos remerciements à toute l'équipe du réseau ACCN, ainsi qu'aux personnes du BER qui ont fourni les données ayant servi de base à ce travail.

Doctorats En cours dans les ENSA

Données et méthodes

La donnée brute est contenue dans le fichier « 250916_Liste doctorants des ENSA 2025_diffusion RSPA » distribué aux RSPA. Il liste les 397 doctorats en préparation dans les ENSA au 16 septembre 2025. Par simplicité, nous n'avons pas mis à jour cette liste au gré des soutenances et des nouvelles inscriptions de la fin d'année 2025, de sorte que l'analyse qui suit constitue un instantané de la recherche doctorale à la date donnée plutôt qu'un tableau de bord utilisable dans le temps.

Comme souvent, la donnée a dû être homogénéisée, corrigée et complétée pour passer par le processus d'analyse. Le Tableau 3 donne la structure du dataset des doctorats en cours à l'issue de

l'étude. Comme pour les précédentes études, chaque ligne a été parcourue et annotée manuellement ; l'attribution du label booléen *isNum* dépend de la réponse de l'évaluateur à la question '*Vrai ou Faux : cette thèse est-elle en lien avec le numérique ?*'. Le genre des doctorants *Gender* est évalué automatiquement par appel de l'API Genderize.io sur la base des prénoms des doctorantes et doctorants. Cette analyse est probablement imparfaite, en particulier pour les prénoms mixtes, mais fait gagner beaucoup de temps et permet d'avoir à grands traits la répartition femme/homme dans la population doctorante.

ANNEE D'INSCRIPTION EN THESE
NOM
PRENOM
GENRE
MAIL
ENSA
NATURE INSCRIPTION
CO-TUTELLE
LABORATOIRE
ECOLE DOCTORALE
DIRECTEUR.ICE DE THESE
FINANCEMENT
TITRE DE LA THESE
TITRE DE LA THESE_FR
DOMAINE
MOTS CLES
isNum
AUTRES INFORMATIONS

Tableau 3 : Structure de la donnée sur les doctorats en cours dans les ENSA.

Analyse des données

On donne ici quelques statistiques générales sur la population doctorale étudiée. Comme le montre la Figure 19, le recrutement de doctorantes reste majoritaire, donnant chez les inscrits un total de 240 femmes pour 152 hommes, en accord avec la tendance observées dans les thèses soutenues ces quinze dernières années. La répartition de ces étudiantes et étudiants dans les ENSA est montrée dans la **Erreur ! Source du renvoi introuvable.**, qui informe également sur la durée typique des doctorats en architecture. 64% des étudiants sont ainsi en thèse depuis plus de 3 ans, 34% depuis plus de 5 ans. Un bon tier (37%) des thèses en cours affichent l'architecture comme leur domaine principal, les autres travaux étant rattachées à des domaines plus précis de l'architecture (architecture et ville, architecture et paysage…) ou à d'autres domaines comme l'aménagement, la géographie ou les SHS. Parmi ces travaux doctoraux en cours, ont été pointés ceux en lien avec les thématiques du Numériques et pouvant intéresser les membres du réseau ACCN. La Figure 20 souligne leur place relativement étroite, 29 thèses sur 397 au total, soit moins d'une thèse en cours sur 14 pouvant être reliées au Numérique. Ces recherches sont en revanche bien représentées sur le territoire, avec 12 des 20 ENSA listées précédemment qui porte au moins un doctorat sur une thématique du Numérique.

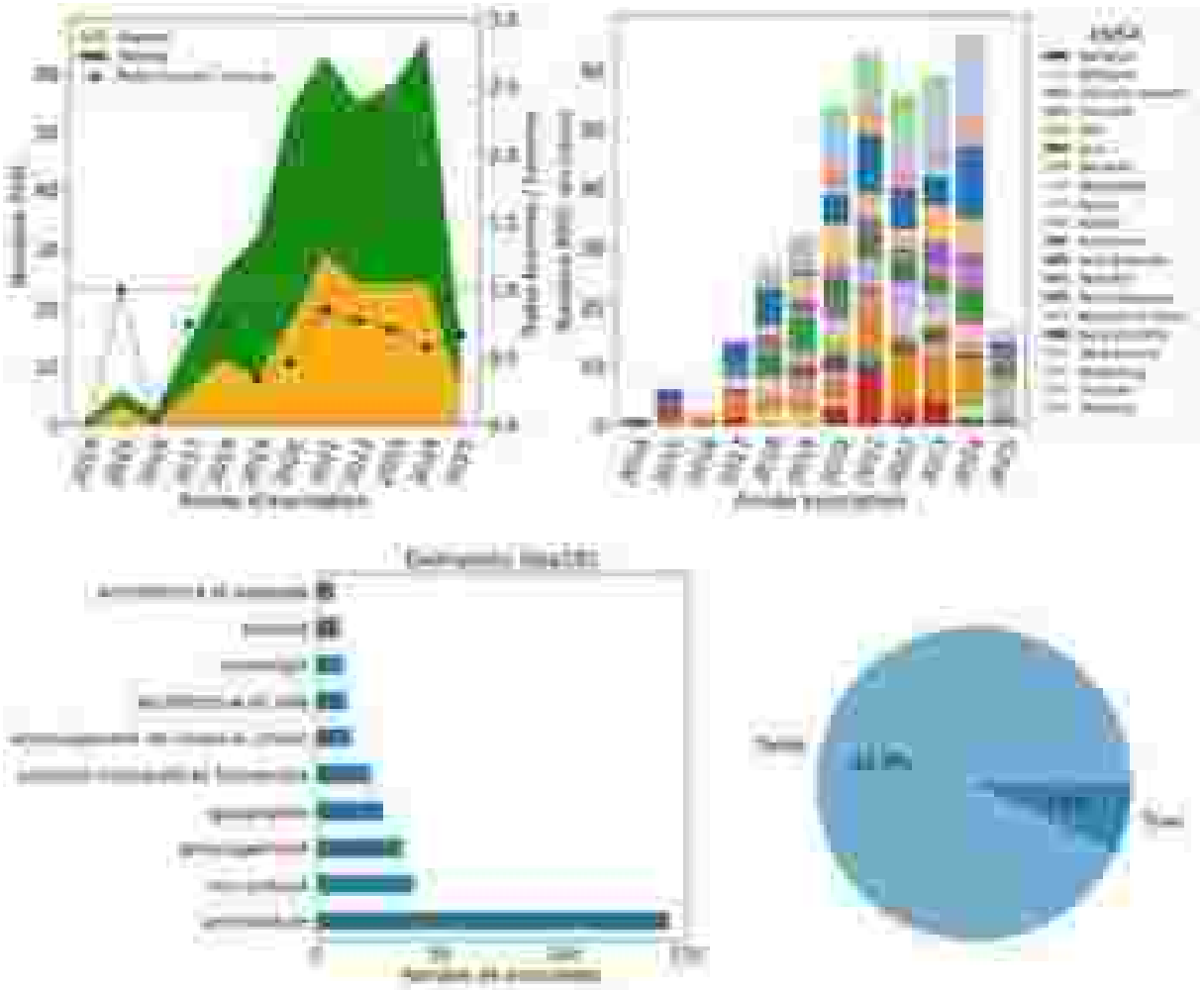


Figure 19 : Répartition en genre dans la recherche doctorale en cours ; Répartition de la recherche doctorale en cours dans les différentes ENSA ; Domaines associés au recherche doctorales des doctorats en cours (top 10) ; Place des thématiques du numérique dans les thèses en cours.

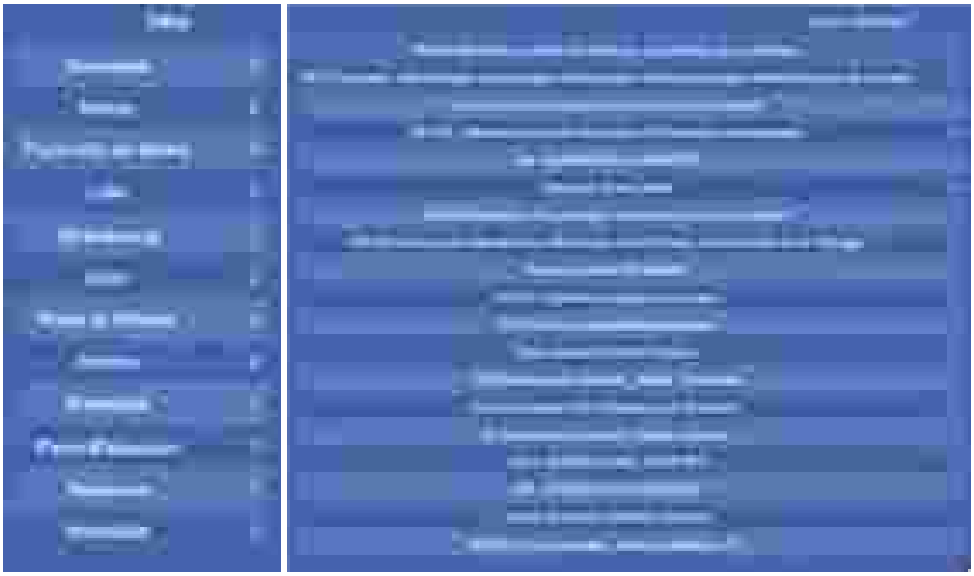


Figure 20 : Acteurs de la recherche doctorale liée au Numérique parmi les travaux doctoraux en cours.

Pour reprendre l'approche sémantique conduite sur les thèses soutenues, on utilise l'association des titres et des mots clés associés à chaque thèse – n'ayant pas encore de résumé disponible pour ces travaux inachevés. En nombre brut de mots, on parle ici de faire travailler nos algorithmes de traitement du langage sur typiquement 10% de la quantité habituelle de donnée textuelle. A très court terme, il serait intéressant de compléter la donnée par un simili-résumé, fourni par les doctorantes et doctorants, ou leurs encadrantes et encadrants. Comme dans la première partie du rapport, les paramètres de projection (nombre de voisins, distance minimale) et de clusterisation (nombre typique de clusters, nombre minimal d'éléments par cluster) sont explorés de façon systématique.

Sur cette exploration, les scores des partitions et la stabilité de l'appariement de deux thèses entre elles, sont relativement bas (typiquement autour de 0.30), certains clusters contenant ainsi des thèses attribuables à d'autres clusters, sans qu'il soit possible de déterminer avec certitude ce qui génère de la confusion dans le processus de classification. L'information à l'échelle des groupes constitués est en revanche de très bonne qualité, avec un indice *Normalized Mutual Information* (NMI) élevé (systématiquement supérieur à 0.7) qui traduit une conservation de l'information globale et une stabilité des grandes tendances relevées.

Sur la Figure 21 on montre un exemple de regroupement sémantique sur l'ensemble des thèses en cours jugées en lien avec le numérique (paramètres : 2 ; 0.6 ; 5). On y distingue cinq groupes sémantiques disjoints. Dans le sens de lecture de la figure : (i) le groupe « Evolution de l'Intelligence », 4 thèses sur les nouvelles approches et outils du numérique dont ceux liés à l'IA ; (ii) « Réhabilitation de l'Héritage », 3 thèses, sur des questions de reconstruction d'images, de mémoire, d'intégration des enjeux énergétiques dans les héritages architecturaux ; (iii) « Construction et Echanges », 5 thèses orientées prévention des risques, interopérabilité des données du bâtiment ; (iv) « Apprentissage Professionnel », une dénomination douteuse pour regrouper 3 thèses en lien avec la réception de l'architecture dans divers cadres et via différents média ; (v) « Environnement et Développement Durable » lui aussi mal nommé, mais regroupant les thèses en lien avec la phase de construction, dans ses aspects matériels (réemploi, matériaux bois, qualité environnementale) et dans ses dynamiques de co-création, de gouvernance, d'échange d'information.

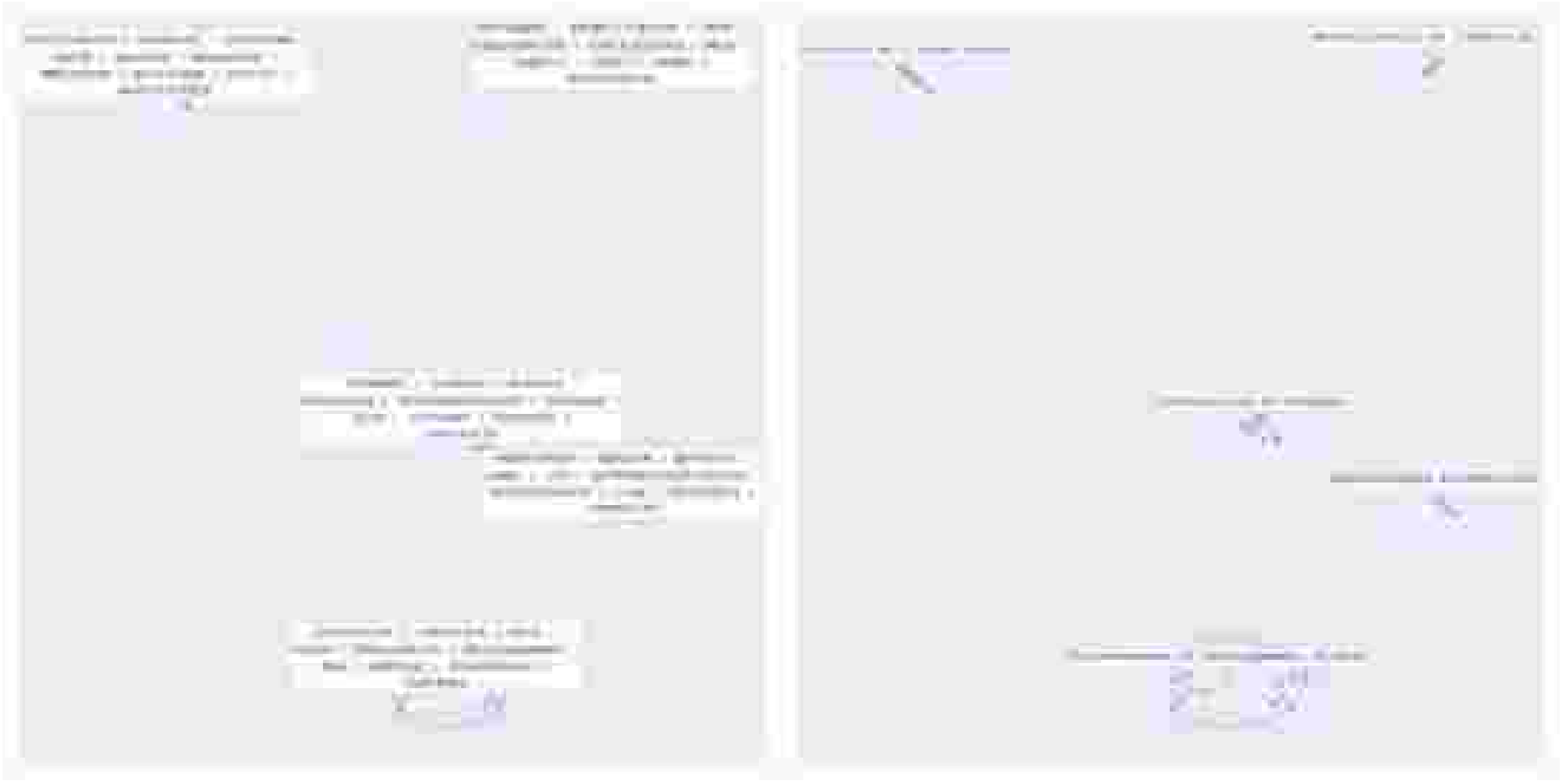


Figure 21 : Tentative de regroupement sémantique sur la base de titres des thèses en cours, avec étiquetage par des mots clés (gauche), ou par l'action d'un LLM sur la base des mots clés et des titres (droite).

Doctorats « Architecture et Numérique » en France depuis 2010

Cette partie a fait l'objet d'une publication dans la conférence SCAN'26 devant se tenir à Montpellier en novembre 2026. L'article décrit la constitution et l'analyse d'un jeu de données étendu, comprenant un ensemble de thèses relatives à l'architecture sur la période 2010-2025, dans lequel sont pointées les thèses relatives au numérique. Nous incrustons ici le manuscrit de l'article soumis au 17 février par l'équipe ACCN. Il présente l'analyse du jeu de donnée et des conclusions partielles sur la recherche doctorale en Architecture et Numérique en France. Il est suivi d'un paragraphe complémentaire sur une autre approche de classification des thèses de doctorat, développée pendant la mission sans être intégrée à la publication.

Quelle place pour les thématiques du numérique dans la recherche doctorale française en architecture ?

Adrien Saint-Sardos[1,1], *Isabelle* Fasse[2], *Kévin* Jacquot[3], *Armelle* Le Mouëllic[4], *Camilo* Cifuentes[1], et *Philippe* Marin[1]

[1] Univ. Grenoble Alpes, ENSAG, MHA, 38000 Grenoble, France

[2] UPR CNRS 2002 MAP, ENSA Marseille, 13009 Marseille, France

[3] URM MAP-Aria, ENSA Lyon, 69120 Vaulx-en-Velin, France

[4] Ministère de la Culture, BER, Paris, France

Abstract. Over the past fifteen years, architecture has established itself as an academic scientific discipline, structured by specialised literature, journals and research methodologies. Doctoral theses in architecture, which often reflect societal concerns, are an interesting source of data, despite their limited accessibility. This study analyses doctoral research in architecture and urban planning conducted between 2010 and 2025 in French universities and schools. A corpus of 1,838 theses, collected through systematic queries, is subjected to statistical and textual analysis to identify the actors involved in doctoral research related to Digital Technology for Architecture. Natural Language Processing (NLP) enables the analysis of titles, abstracts and keywords to identify and label work related to digital transformations. The trends identified are compared with existing diagnoses of teaching and research in architecture. The results, including the corpus and visualisations in the form of interactive graphs, are accessible as open data according to FAIR principles.

Keywords. Bibliometrics; NLP; PhD; Digital Architecture;

Résumé. Depuis une quinzaine d'années, l'architecture tend à se présenter comme une discipline scientifique académique, structurée par une littérature spécialisée, des revues et des méthodologies de recherche. Les thèses de doctorat en architecture, reflétant souvent les préoccupations sociétales, constituent une source de données intéressante, malgré leur faible accessibilité. Cette étude analyse la recherche doctorale en architecture et urbanisme menée entre 2010 et 2025 dans les universités et

[1] Corresponding author: saint-sardos.a@grenoble.archi.fr

écoles françaises. Un corpus de 1838 thèses, collecté par requêtes systématiques, est soumis à une analyse statistique et textuelle pour identifier les acteurs de la recherche doctorale en lien avec le numérique pour l'architecture. Le *Natural Language Processing* (NLP) permet une analyse des titres, résumés et mots-clés pour identifier et labelliser les travaux liés aux transformations numériques. Les tendances dégagées sont confrontées aux diagnostics existants sur l'enseignement et la recherche en architecture. Les résultats, incluant le corpus et les visualisations sous forme de graphes interactifs, sont accessibles en open data selon les principes FAIR.

Mots-clés. Bibliométrie ; NLP ; Doctorat ; Architecture numérique ;

1 Introduction

L'architecture s'institutionnalise progressivement comme discipline académique, produisant et mobilisant des ressources documentaires, rapports techniques, articles de revue [1–3]. Parmi ces ressources, on compte les thèses de doctorat [4], qui se sont se démocratisées en France depuis 2005 suite à la réforme LMD et de l'alignement des systèmes éducatifs européens [5].

Ces thèses constituent une source d'information originale et riche, issue d'un travail de recherche continue d'au moins trois ans [4,6,7]. Des efforts nationaux ont été déployés dans de nombreux pays pour rendre cette donnée accessible et interopérable, en rendant systématique la diffusion des manuscrits sur dépôts de données publics [8–10] : on citera, pour l'Europe, TESEO en Espagne, OpenAIRE aux Pays-Bas, Zenodo en Suisse, Hal et theses.fr en France.

Ces dépôts constituent des bases de données opérables, qui ont par exemple rendu possible un suivi des travaux doctoraux en Architecture à l'échelle universitaire [4], et à l'échelle nationale dans les pays d'Europe de l'Ouest [11], en Turquie [12], au Nigéria [13], et en Australie [14]. Au-delà du recensement, cette accessibilité de la recherche doctorale permet de mieux éclairer les politiques publiques [15] et de suivre les grandes tendances de la recherche ; on citera notamment le travail de Zinilli et al. [16] pour son analyse de la recherche doctorale autour des thématiques du changement climatiques en Italie.

Dans le présent article, l'objectif est de mettre en valeur la production doctorale française en architecture, à travers la constitution puis l'étude d'un jeu de données listant des thèses de doctorat en architecture soutenues en France entre 2010 et 2025. Dans le contexte de la transition numérique, nous s'intéressons plus particulièrement à la présence des thématiques numérique (*Digital Architecture*) [17], et tentons d'évaluer la place du numérique dans la recherche doctorale française en architecture au cours de ces quinze dernières années.

Pour répondre à cette question de recherche, nous nous reposons sur des outils d'analyse statistiques et sur des approches de traitement du langage naturel (TLN), en s'inspirant de travaux récents de Nanni et al. sur la détection d'interdisciplinarité dans les abstracts de thèses [18], ou ceux de Zinilli et al. sur le regroupement sémantique d'abstracts de thèses [16]. Les résultats proposent tout d'abord quelques caractéristiques de la recherche doctorale en embrassant l'ensemble des thématiques du jeu de données ; puis nous considérons plus spécifiquement les travaux traitant du numérique, et tentons de faire émerger les familles thématiques explorées sur la période 2010-2025 ; et nous terminons en proposant un outil de consultation en ligne pour faciliter l'accès, la prise de connaissance et le partage de ces données.

2 Matériel et Méthodes

Donnée. Le corpus considéré est une liste des thèses de doctorat en architecture, soutenues dans les institutions françaises depuis 2010, jusqu'au mois d'octobre 2025. Un jeu de données du ministère de la Culture contenant 650 thèses a été complété par requête systématique sur le site de dépôt theses.fr, pour collecter les thèses en lien avec l'architecture. Pour entrer dans le corpus, une thèse doit être en français, dans les bornes de temps considérées, mentionner explicitement l'architecture dans sa discipline de rattachement, ou, dans ses mots-clés associés, les termes 'architecture', 'urbanisme', 'patrimoine', 'bâtiment'. A la suite d'une requête, on filtre automatiquement les thèses identifiées sur la base du nom de l'auteur pour éviter les doublons, puis vérifions manuellement le lien

entre la thèse et le domaine de l'architecture. Une étape complémentaire d'extraction automatique de contenu depuis internet (*scrapping*), basée sur l'utilisation de la librairie python *beautifulSoup*, permet de récupérer le résumé de la thèse. La labellisation par un booléen (vrai/faux) pour marquer le rattachement d'une thèse aux thématiques du Numérique pour l'architecture est réalisée humainement, avec relecture par un pair et discussion en cas d'ambigüité.

Fig. 1. Les différents ensembles de données considérés : ensemble des thèses en lien avec l'architecture soutenues en France entre 2010 et (oct.) 2025.

Genre. Le genre des doctorants est évalué sur l'ensemble du corpus ARCHI par la librairie *genderize* avec un seuil de certitude à 0,7 pour déterminer automatiquement le genre des doctorantes et doctorants. Les lignes laissées vides par l'API sont vérifiées manuellement par recherche internet.

Traitement du Langage Naturel. L'analyse et la classification des différents textes liés au thèses repose sur les outils de TLN de la librairie populaire *sklearn*, ainsi que sur une librairie *BunkaTopics*, présentée dans Dampierre et al. [19]. La vérification de la stabilité des groupements sémantiques est effectuée par la méthode de *bootstrapping*, i.e. en vérifiant que les textes issus d'un sous ensemble de textes du corpus tirés au hasard conservent leurs proximités sémantiques. Le modèle de langue utilisé est un BERT « all-MiniLM-L6-v2 » disponible sur le dépôt Huggingface[2]. L'algorithme de projection utilisé est UMAP, et l'on compare les performances de trois méthodes de clustering classiques (KMeans, HDBSCAN et AgglomerativeClustering). Un post traitement des clusters, basé sur les mots clés identifiés et sur les documents analysés, utilise l'API du modèle gpt-4o pour nommer les clusters.

Datavis. L'ensemble des graphiques sont générés par la librairie *matplotlib*. L'outil de visualisation repose sur la librairie *Networkx* permettant de générer et d'afficher des graphes de connaissance dynamiques. Un exemple de graphe est déposé sur le Github du réseau Architecture, Conception et Culture Numérique à l'adresse suivante : https://adrien341.github.io/. L'ensemble du jeu de données sur les thèses en architecture sera déposé sur HAL.

Tableau 1. Structure de la donnée sur les doctorats en cours en Architecture et Numérique.

Entrée	Contenu ; format
SOUTENANCE	Date de soutenance ; format dd/mm/yyyy
ANNEE	Année de soutenance ; date object, format yyyy
NOM	Nom du docteur ; str

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

PRENOM	Prénom du docteur ; str
GENRE	Genre du docteur ; “female”, “male”, “ ”
INSTITUTION	Nom de l’institution de rattachement (ENSA, Ecole, Univ.) ; str
LABORATOIRE	UMR de rattachement ; str
ECOLE DOCTORALE	ED de rattachement ; str
DIRECTEUR.ICE DE THESE	Nom prénom directeurice de thèse ; liste de str
TITRE DE LA THESE	Titre de la thèse ; str
TITRE DE LA THESE_FR	Titre traduit en français lorsque nécessaire ; str
DISCIPLINE	Discipline de rattachement principal ; str
MOTS CLES	Liste des mots clés attachés ; [str ; str ; …]
CATEGORIE	Catégorie CMA Archi 2023 attribuée automatiquement ; str
THESE_ID	Numéro de référencement sur theses.fr ; str

3 Résultats

3.1 Statistiques

Ce premier paragraphe présente quelques statistiques tirées de l’analyse du corpus. La Figure 2A montre ainsi l’évolution du nombre de doctorats en lien avec l’architecture soutenus pendant les quinze dernières années. Après un doublement entre 2011 et 2013, ce nombre reste stable autour de 125 doctorats par an. La proportion de thèses en lien sur l’ensemble de la période est de 12%±4%, mais semble croître au cours du temps, passant de 5% en 2010 à 20% en 2024. Une analyse plus fine est possible sur les institutions de type École Nationale Supérieur d’Architecture et de Paysage (ENSA-P). La proportion de thèses en lien avec le numérique y reste quasi-constante sur la période considérée, à 16%±6%.

La Figure 2B présente les tendances en termes de genre des doctorantes et doctorants. On compte 967 docteures, dont 92 sur des thématiques numériques, pour 826 docteurs, dont 136 sur des thématiques Numériques. Le rapport femme-homme se trouve à 1.18±0.31 pour l’ensemble du corpus, et tombe à 0.72±0.42 pour les doctorats en lien avec le Numérique. Côté ENSA-P, le rapport Femme-Homme est de 1.35± 0.49 sur la période considérée, contre 0.93± 0.97 pour les thèses sur le numérique.

L’analyse du jeu de données permet aussi d’identifier les acteurs académiques de la recherche doctorale en numérique pour l’architecture, à l’échelle des institutions de rattachement, des écoles doctorales et des laboratoires. Les Figure 2C présente pour illustration les 25 institutions de rattachement les plus représentées dans le corpus. L’ENSA de Nantes y apparait comme les premiers producteurs de thèses en lien avec le Numérique (13% des thèses soutenues depuis 2010), suivie par les ENSA de Paris-Malaquais et de Nancy (respectivement 8 et 6,5% des thèses soutenues depuis 2010). L’HESAM y apparait comme une institution non-ENSA-P émergente en termes de recherche doctorale sur l’Architecture en lien avec le Numérique, où ont été soutenues neuf thèses au cours des cinq dernières années. Exception faite de l’Université Paris-Est, qui compte une dizaine de doctorats sur la période considérée, les autres universités de la Figure ont permis la soutenance de trois et cinq thèses sur le sujet, sans accélération notable sur la période. Le jeu de données permet une analyse comparable à l’échelle des laboratoires de rattachement. La Figure des 25 premiers laboratoires de rattachement (non montrée ici) est le reflet de la Figure 2C, avec en tête les laboratoires de Nantes (AAU-CRENAU), de Paris-Malaquais (GSA), de Nancy (MAP-CRAI), mais aussi celui de Paris-La-Villette (MAP-MAACC). La cinquième place est occupée par le laboratoire LIRIS, qui héberge les doctorats de plusieurs institution (EC Lyon, INSA de Lyon, Université de Lyon).

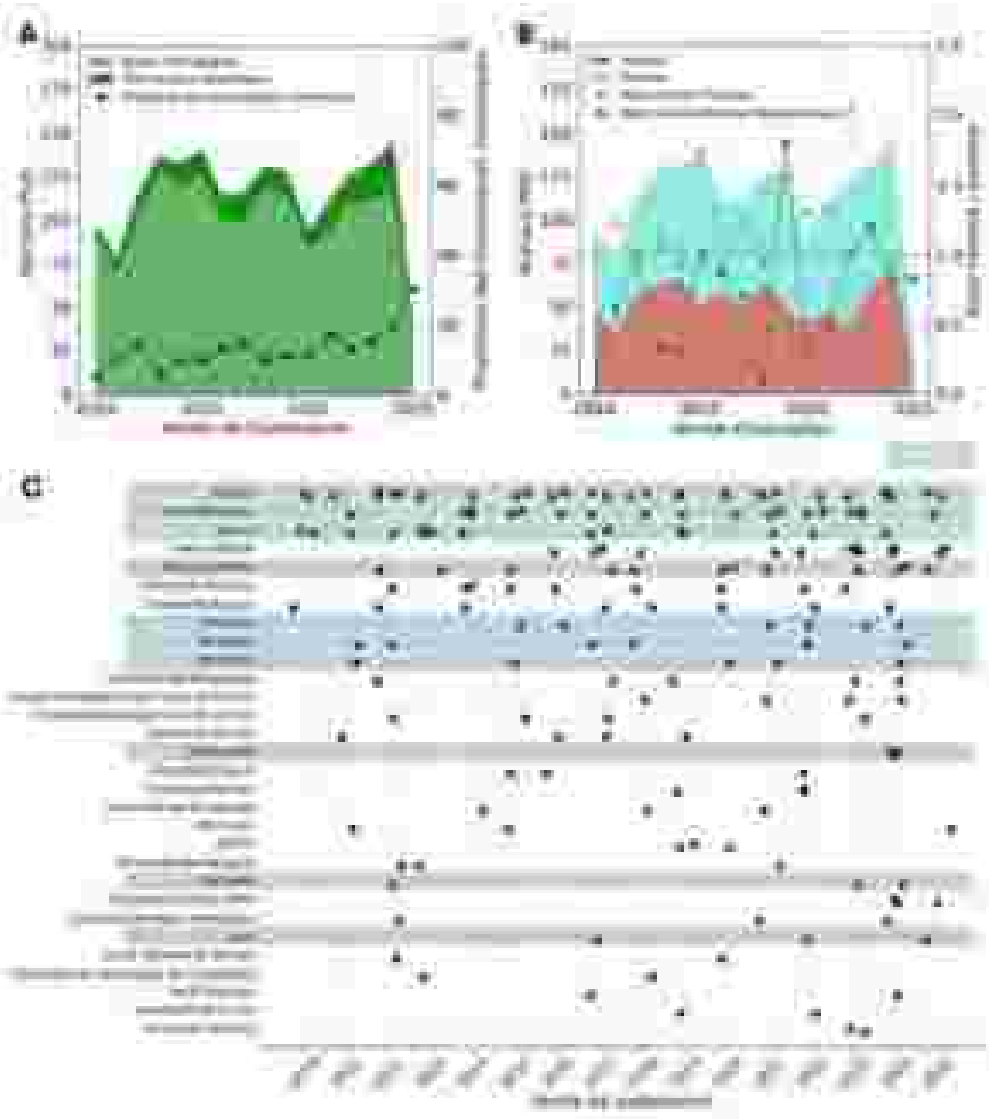


Fig. 2. Statistiques sur le doctorat en architecture et la place du numérique. (A) Place des thèses en lien avec le Numérique dans le corpus (B) rapport Femme-Homme des docteurs en architecture diplômées depuis 2010 dans le corpus (C) Répartitions par institution des thèses du corpus ARCHI-NUM, au cours du temps (top 25 sur 67 institutions représentées).

3.2 Etude des mots-clés et résumés

Une première analyse du corpus des thèses sur le numérique pour l'architecture peut être menée à l'échelle des mots-clés utilisés dans leurs résumés. Sur la Figure 3, on représente ainsi deux nuages de mots pour deux thèses du corpus ARCHI-NUM, qui montrent assez directement l'appartenance de chacune des thèses, à l'utilisation de données cartographiques, et au travail en projet dans un cadre BIM, respectivement.

Fig. 3. Nuages de mot associés aux résumés de deux thèses du corpus ARCHI-NUM, celle de Nicolas BIORET (2010) et celle de Hana REZGUI (2024).

La Figure 4 montre quant à elle l'évolution temporelle de l'occurrence de ces mots-clés à l'échelle de l'ensemble des thèses d'ARCHI-NUM. La courbe de la Figure 4A associe simplement l'année de soutenance à la présence d'un mot-clé dans un résumé. Un mot clé répété plusieurs fois dans un même résumé compte pour 1. La figure permet par exemple d'observer l'émergence d'un mot-clé comme « BIM » dans les soutenances postérieures à 2015, ou d'affirmer que, chaque année, au moins un doctorat soutenu traite de questions de simulation ou de système d'information cartographique. Cette représentation ayant le défaut de présenter les travaux doctoraux comme des points dans le temps, la figure 4B propose d'intégrer la présence d'un mot-clé sur toute la durée du projet doctoral, en considérant qu'un doctorat en architecture dure typiquement quatre ans. Cette représentation 'lisse' la présence des différentes thématiques dans la recherche doctorale, et permet de la mesurer à un temps donné.

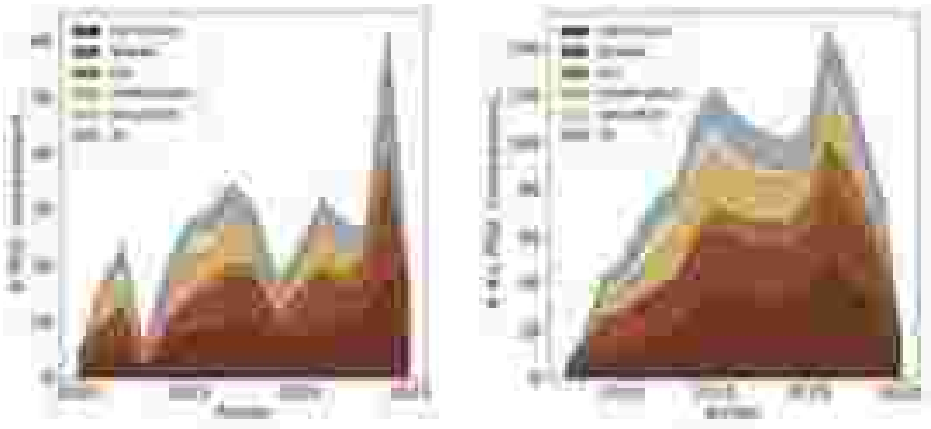


Fig. 4. Occurrences des sept mots-clés relatifs au Numérique les plus cités dans les résumés de thèse du corpus ARCHI-NUM. (Gauche) Occurrences par année de soutenance, un résumé qui contient le mot-clé compte comme +1 sur son année de soutenance. (Droite) Occurrences par période de 4 ans - un résumé qui contient le mot compte comme +1 sur une période de 4 ans finissant à la date de soutenance.

Pour terminer, une approche basée sur le sens des textes associés aux thèses cherche à mettre en évidence des proximités et des liens entre les travaux doctoraux de différentes auteurices à différentes périodes, sans préjuger des thématiques ou des mots-clés qui pourraient y être associés. Les textes considérés sont les résumés des thèses. Sur la Figure 6A, on représente ainsi l'ensemble des résumés du corpus, vectorisés par un modèle de langue léger, puis projetés (algorithme UMAP, réglé avec paramètre de distance minimale 0,1 ; nombre de voisins 5 ;) et regroupés par clustering (algorithme Clustering Agglomératif ; nombre de cluster 7). On voit ainsi apparaître sept clusters thématiques stables (Bootstrapping à échantillon n/2 : ARI_moyen = 0,34 ; NMI_moyen = 0,51), étiquetés avec des mots significatifs tirés des résumés. Chaque cluster contient en moyenne 33±8 thèses. Ces clusters, et les documents qui y sont associés, passent par un modèle de langue massif, pour voir émerger sept thèmes exprimés en langage naturel (Figure 6B), respectivement :

- « Amélioration de la productivité » associé à l'étude de la maquette, du Building Information Modeling, de l'approche Lean en construction ;
- « Energie et Environnement » regroupe les thèses autours des réseaux et des flux (air, eau en particulier), de leur gestion, de leur impact sur la qualité environnementale d'un objet architectural ;
- « Environnement et Gestion » s'intéresse à l'influence des processus humains (décision, cadres méthodologiques politiques, gestions de projet) sur la réussite et la performance environnementale de l'architecture ;
- « Géolocalisation et Imagerie » regroupe les thèses en lien avec les données cartographiques, les photos urbaines ou satellites, des vidéos, les systèmes d'information géographiques (SIG) ;
- « Histoire et Patrimoine » pour des thèses s'intéressant au lien entre l'objet architectural et le temps, interrogeant les notions de modernité, de conservation, de transformations des espaces et des pratiques ;
- « L'Expérience Humaine » s'intéresse à la perception d'une information, d'une sensation liée à l'architecture ou à sa pratique, par les humains, leur cognition ou leur imaginaire ;
- « Technologie de conception » lié aux approches de conception assistée par ordinateur, à la modélisation, aux expériences et aux données sur les matériaux.

La Figure 5C reporte sur la Figure 2C les différents rattachements thématiques identifiés par cette approche non-supervisée. Pour exemple, l'ENSA Bordeaux, affiche six thèses attribuées à deux groupes « Energie et Environnement » (3) et « Environnement et Gestion » (3), intérêt pour les thématiques environnementales qui est confirmé par les mots clés relevés dans les résumés (« quartier, écoquartier, climat, climatique, confort, chaud, énergie, énergétique »). Une observation ligne à ligne révèle une forte teinte thématique de chaque institution sur un ou deux groupes. L'homogénéité thématique de la recherche doctorale est évaluée par un calcul de l'indice de Simpson [20] sur les institutions ayant accompagné plus de deux thèses, indice qui atteint une moyenne de 0.669±0.091 sur la période considérée.

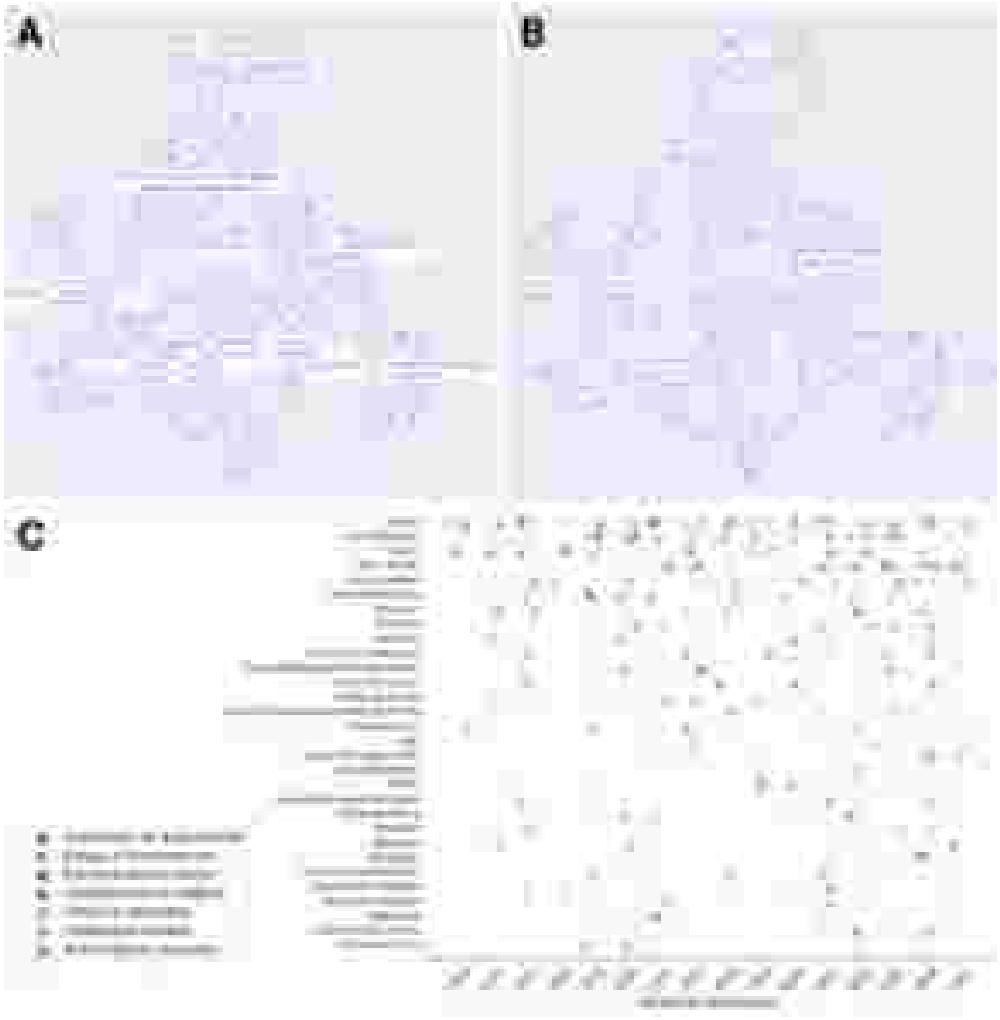


Fig. 5. Projection, regroupement sémantique, filtrage des résumés de thèse du corpus. (A) Résultat de regroupement sémantique des abstracts du Corpus ARCHI-NUM (B) post-traitement des groupes par un modèle de langue pour obtenir des étiquettes en langage naturel. (C) représentation des thèses selon l'année de soutenance, et l'étiquetage du groupe en langage naturel. L'ensemble de ces graphes interactifs sont disponibles à la consultation via le dépôt FigShare (https://doi.org/10.6084/m9.figshare.31305136).

4 Discussion

4.1 Eléments de réponse

Sur la période considérée, l'analyse du corpus montre une place ténue mais stable des thématiques du numérique dans la recherche doctorale en architecture. En ordre de grandeur une thèse en architecture sur quinze se préoccupe de numérique - une proportion conservée lorsque l'on regarde le sous-ensemble des thèses soutenues en ENSA-P. Cette recherche est répartie sur un grand nombre d'acteurs institutionnels, comprenant des ENSA-P (quatorze institutions, 48% des thèses), mais aussi des universités (quarante-trois institutions, 39% des thèses) et des écoles

d'ingénieur (neuf institutions, 13% des thèses). Seules dix de ces soixante-sept institutions ont hébergé plus de cinq thèses en lien avec le numérique sur la période considérée. Du reste, les acteurs de la recherche en architecture et numérique tendent à se diversifier, sept ENSA-P ayant par exemple hébergé leur première thèse sur les thématiques numériques après 2017.

Alors que le doctorat en architecture concerne une majorité d'étudiante (61% de doctorantes en 2023 selon les chiffres du ministère de la Culture), la proportion de genre s'inverse lorsque l'on considère les travaux liés au numérique, avec un rapport Femme-Homme qui tombe régulièrement au-dessous des ½ jusqu'en 2022. Ce constat d'inégalité peut être adressé par des politiques à l'échelle des laboratoires et des institutions [21].

En termes d'intérêt de recherche, le corpus considéré fait facilement ressortir des mots-clés en lien avec les techniques du numérique (simulations, modélisations, visualisations 3D, etc.), permettant d'observer la montée du BIM à partir de 2015, ou l'amenuisement de la place des simulations dans les résumés. A ce stade, le vocabulaire lié à l'intelligence artificielle et aux IA génératives n'apparaît ni sous forme de mot-clé récurrent, ni comme termes sémantiquement significatifs. Une étude des doctorats en cours, présentement menée à l'échelle des ENSA-P, permettra de compléter nos jeux de données, et donnera une vision et une compréhension plus prospectives de la connaissance en train de se construire sur le sujet.

L'analyse sémantique des résumés a permis de rapprocher des thèses soutenues dans des institutions variées, à des dates variées, et de faire apparaître des clusters liés par des thématiques scientifiques, sans biais de connaissance préalable sur la recherche française en architecture. On observe une répartition assez homogène de la recherche entre les différents clusters, mais aussi des recouvrements possibles et des coalescences de clusters qui apparaissent à certaines combinaisons de paramètre. La sémantique rapproche facilement les approches teintées Sciences et Techniques de l'Architecte, ou bien celles en lien avec les performances écologiques, qui tendent à émerger de l'analyse indépendamment des paramètres choisis.

4.2 Limites et Perspectives

La première limitation de l'approche décrite tient à la structuration de la donnée. Il apparaît par exemple que certaines thèses attribuées à des universités, notamment à l'HESAM ou à Paris-Est, pourraient être rattachées à des ENSA-P et modifier les chiffres des Figures 2C et 5C. La clarification du rattachement nécessite un parcours de chaque affiliation du corpus, parfois un téléchargement du manuscrit. Le suivi des écoles doctorales est rendu plus difficile par les changements de noms et de statut de ces entités au cours du temps ; des travaux sont en cours pour croiser nos corpus avec les jeux de données de Datagouv sur les unités de recherche et les écoles doctorales accréditées [22,23].

L'approche proposée repose sur une première labélisation grossière, répondant à cette question « cette thèse concerne le numérique ? Vrai ou Faux ». A ce jour, il a été décidé d'effectuer cette labélisation via des opérateurs humains. Si cette stratégie permet de s'assurer de la pertinence des éléments associés au corpus ARCHI-NUM, elle reste peu adaptée à un passage à l'échelle sur un grand nombre de documents. Des développements basés sur de la décision automatisée « zero-shot » sont en cours, et donnent des résultats prometteurs, notamment en termes de précision (avec très peu de faux-positifs).

Le processus d'analyse et de représentation sémantique utilisé dans l'étude pourrait être complété par des outils clé en main pour analyser la littérature comme Gargantext [24], voire permettre de reconstruire l'évolution des tendances dans une discipline scientifique [25]. Ces approches demandent cependant 1) d'atteindre une masse critique de documents, et 2) de s'assurer de la propreté des données. En l'état, l'approche utilisée a l'avantage de sa malléabilité, l'ensemble de la cartographie sémantique pouvant être produite via une demi-douzaine de cellules dans un notebook accessible et annoté. L'ensemble des codes, données (purgées des données personnelles non publiques), et visualisations produites sont placés dans un dépôt en vue de leur réutilisation, de leur complétion et de leur consultation. Un graphe dynamique, accessible directement par une URL hébergée sur GitHub, permet d'explorer la centaine de thèses du corpus qui ont été hébergées par des ENSA (voir Figure 7).

Au-delà du constat statistique, cette étude est un premier pas vers une cartographie de la recherche en architecture et numérique, une invitation à davantage de collaboration entre les différents acteurs de cette thématique. La proximité sémantique de plusieurs thèses issues de laboratoire différents pourrait déclencher des mises en commun de ressources, d'actions pédagogiques, de réponse à des appels à projet, d'événement liés à la recherche. L'approche décrite ici pourrait être étendue, à moindre coût de temps, sur des périodes et des thématiques différentes. Il serait par exemple possible de constituer un corpus, puis d'étudier les sujets de recherche doctorale

relative à la ruralité, à la notion de projet en architecture, ou à la transition écologique. L'approche constituerait alors un observatoire de la recherche doctorale, et permettrait d'orienter les efforts de recherche à l'échelle du territoire, voire de comparer les productions doctorales avec celles d'autres nations.

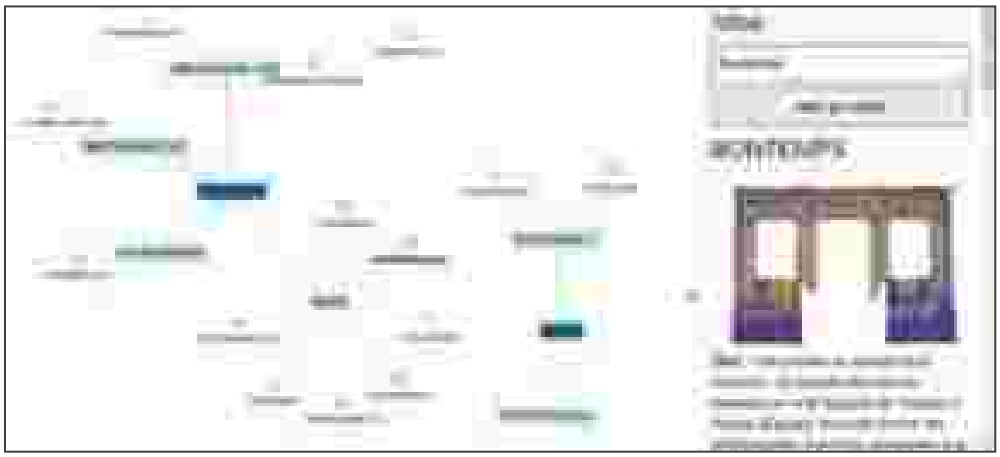


Fig. 7. Visualisation des travaux doctoraux Architecture et Numérique, disponible en ligne

References

1. *The Discipline of Architecture*. (University of Minnesota Press, Minneapolis, 2001).

2. Plowright, P. D. *Revealing Architectural Design*. (Routledge, 2014). doi:10.4324/9781315852454.

3. Sauvé, J.-S., Mongeon, P. & Larivière, V. From art to science: A bibliometric analysis of architectural scholarly production from 1980 to 2015. *PLoS ONE* **17**, e0276840 (2022).

4. Sağlamer, E. G. & Erkök, F. *Doctoral Education in Architecture*. (Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK, 2015).

5. Journal Officiel. *Décret N° 2002-482 Du 8 Avril 2002 Portant Application Au Système Français d'enseignement Supérieur de La Construction de l'Espace Européen de l'enseignement Supérieur*.

6. Boyer, C. J. The Ph.D. Dissertation: An Analysis of the Doctoral Dissertation as an Information Source. (THE UNIVERSITY OF TEXAS AT AUSTIN, Metuchen, N.J., 1972).

7. Baptista, A. *et al.* The doctorate as an original contribution to knowledge: Considering relationships between originality, creativity, and innovation. *FLR* **3**, 55–67 (2015).

8. Stanton, K. V. & Liew, C. L. Open access theses in institutional repositories: an exploratory study of the perceptions of doctoral students. *Information Research* **16**, (2011).

9. Troman, A., Jacobs, N. & Copeland, S. A new electronic service for UK theses: access transformed by EThOS. *Interlending & Document Supply* **35**, 157–163 (2007).

10. Jones, R. & Andrew, T. Open access, open source and e-theses: the development of the Edinburgh Research Archive. *Program* **39**, 198–212 (2005).

11. Foque, R., Van der Voordt, T. & Wegen, H. B. R. *Doctorates in Design and Architecture. Vol. 1, State of the Art.* (1996).

12. Erbil, Y. & Gür, M. A review of doctoral dissertations in architecture in Turkey. *HumanSciences* **15**, 1481 (2018).

13. Ilesanmi, A. O. Doctoral research on architecture in Nigeria: Exploring domains, extending boundaries. *Frontiers of Architectural Research* **5**, 134–142 (2016).

14. Beza, B. B., Zeunert, J., Kilbane, S. & Padgett Kjaersgaard, S. Examining PhD modes in the Australian landscape architecture academy. *Landscape Research* **47**, 679–694 (2022).

15. Damba, F. U., Mtshali, N. G. & Chimbari, M. J. Factors influencing the utilization of doctoral research findings at a university in KwaZulu-Natal, South Africa: Views of academic leaders. *PLoS ONE* **18**, e0290651 (2023).

16. Zinilli, A. *et al.* Anatomy of climate change research in Italian doctoral dissertations using a machine learning approach. *Sci Rep* **15**, 38095 (2025).

17. Leach, N. *Architecture in the Age of Artificial Intelligence: An Introduction to AI for Architects*. (Bloomsbury Publishing, London, 2025).

18. Nanni, F., Dietz, L. & Ponzetto, S. P. Toward a computational history of universities: Evaluating text mining methods for interdisciplinarity detection from PhD dissertation abstracts. *Digital Scholarship in the Humanities* **33**, 612–620 (2018).

19. Dampierre, C. de, Mogoutov, A. & Baumard, N. Towards Transparency: Exploring LLM Trainings Datasets through Visual Topic Modeling and Semantic Frame. Preprint at https://doi.org/10.48550/arXiv.2406.06574 (2024).

20. SIMPSON, E. H. Measurement of Diversity. *Nature* **163**, 688–688 (1949).

21. Guziolowski, C., Thiery, O., Harzallah, M., Rizkallah, M. & Chevallereau, C. Gender Equality in a Digital Science Research Laboratory.

22. Ministère de l'Enseignement supérieur, de la Recherche et de l'Espace. Liste des écoles doctorales accréditées. (2026).

23. Ministère de l'Enseignement supérieur, de la Recherche et de l'Espace. Structures de recherche publiques actives. (2026).

24. Gargantext. ISCPIF (2023).

25. Chavalarias, D., Lobbé, Q. & Delanoë, A. Draw me Science: Multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies. *Scientometrics* **127**, 545–575 (2022).

Vers d'autres approches de regroupement sémantique

L'approche décrite dans le paragraphe précédent repose sur la comparaison de résumés de thèses, et l'identification, par l'ordinateur, de points communs et de différences sémantiques entre eux. Cette approche a au moins deux avantages : Premièrement, elle met à plat l'ensemble du corpus, et cherche à projeter puis regrouper uniquement sur la base des textes fournis et vectorisés, sans d'autre connaissance préalable que celle de la langue utilisée. Deuxièmement, elle ne dépend pas de catégories préexistantes ou présupposées, et apparaît, de fait, comme relativement non-biaisée.

Les regroupements formés varient selon les paramètres utilisés, faisant de cette cartographie 'non-supervisée' un objet à facettes, multi-échelles, et donnant des résultats parfois difficiles à interpréter, en particulier pour des travaux doctoraux aux interfaces de plusieurs disciplines.

Une autre façon de procéder à ces regroupements consiste à demander directement à un modèle de langue une labellisation pour un texte de donné, en lui fournissant une grille de classification. On parle ici d'utiliser des méthodes de Zero-shot, i.e. des tâches de classification pour lesquelles l'algorithme n'a pas directement reçu d'entrainement pour reconnaître spécifiquement chacune des classes [voir par exemple l'article de Moreno-Garcia et al. *A novel application of machine learning and zero-shot classification methods for automated abstract screening in systematic reviews* (2023)].

C'est ce que nous avons montré à la page 17 du présent rapport, par l'interrogation du modèle Ollama gpt-oss:120b en imposant une grille de sept labels possibles (Voir Figure 23). Cette approche a permis d'étiqueter rapidement l'ensemble des cent thèses étudiées, et d'obtenir un graphe d'évolution des tendances de recherche au cours du temps et dans les différentes ENSA-P. Elle peut être généralisée pour nos jeux de données étendus.

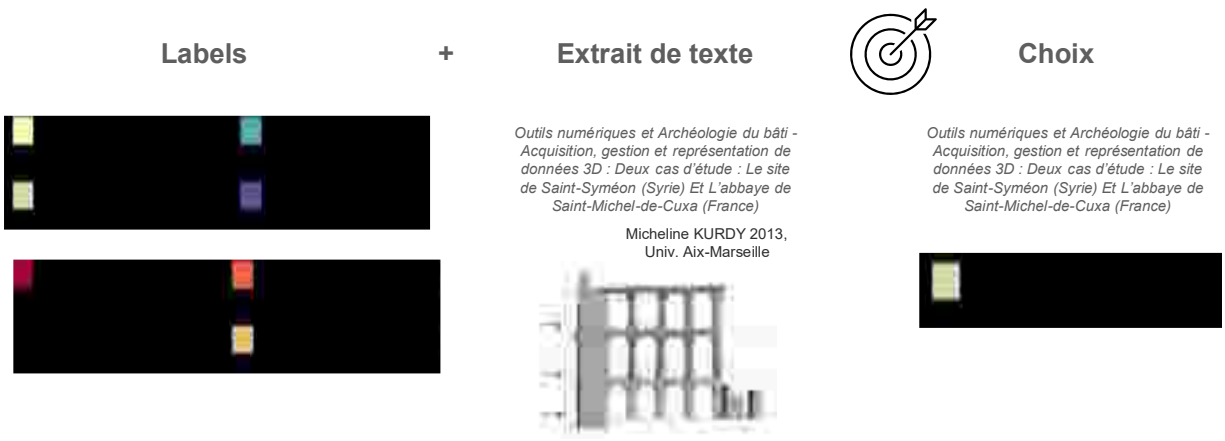


Figure 22 : Schéma explicatif de la classification en zéro-shot utilisée pour classer les cent thèses du dataset original de l'étude (voir Figure 11 de la Partie 1).

L'évaluation de ces approches zéro ou few shots a pu être menée lors de la deuxième partie de la mission. Une sous-partie du jeu de donnée (46 thèses, 20% du corpus) a été étiquetée par un opérateur humain, sur la base de la lecture de l'ensemble des données (titres, mots clés, auteurs, institution, résumé, qui a attribué à chaque thèse l'un des sept labels présentés sur la Figure 23. On demande ensuite à un LLM (gpt-4o) de labelliser ces mêmes thèses sur la base de données plus ou moins fournies, en mesurant le taux d'accord '*matching*', rapporté sur le tableau suivant.

Type de Choix	Labels à attribuer	Base fournie	Taux d'accord (%)
Zero-shot	CMA	Titre	47.83
Zero-shot	CMA	Titre + Mots-clés	41.30
Zero-shot	CMA	Titres + Résumés	36.96
One-Shot (Ex. 1 Titre)	CMA	Titre	54.35
One-Shot (Ex. 1 Titre)	CMA	Titres + Résumés	56.52
Zero-Shot (+ 1 Définition)	CMA	Titres + Résumés	50.00

Une analyse plus fine des différents cas d'accord ou de désaccord entre le LLM et l'évaluateur humain révèle une forte disparité de précision selon le label considéré. Les labels 'Cartographie, système d'information géographique (SIG)', 'Conception paramétrique, conception générative (IA)' ou 'Patrimoine et maquette numérique (HBIM)' sont systématiquement bien identifiés par le LLM, alors qu'à l'inverse, les thèses de 'Fabrication, matériaux et construction robotisée' ou de 'Maquette numérique (BIM) et simulation' sont quasi systématiquement mises sous le label 'Conception paramétrique, conception générative (IA)'. Les labels 'Réalité virtuelle et augmenté' ou 'Théorie des média et des études numériques' sont plus variables en termes de performance.

Il est possible que les modèles de langue ne puissent pas saisir l'ensemble des nuances dans les sujets de thèses, et soient limité à pouvoir distinguer des approches sciences de l'ingénieur, des problématiques patrimoniales ou des sciences humaines et sociales. Une part de la fragilité de ces résultats tient peut-être également de la fragilité de l'évaluation humaine : certaines thèses se rapportent à plusieurs labels, qui n'ont du reste pas de définition consensuelle à ce jour. L'intérêt de l'expérience est davantage de monter les leviers pour rendre opérable cette classification à la volée de travaux doctoraux : 1) proposer des labels explicites et univoques, leurs définitions, et des exemples typiques à l'algorithme 2) fournir des bases textuelles de qualité et de quantité suffisantes, et 3) analyser systématiquement les résultats pour identifier le ou les labels qui génèrent de la confusion lors de la classification.

Conclusions et perspectives

L'ensemble de ce rapport donne une idée de la place du numérique dans la recherche doctorale française en architecture, en s'attachant à caractériser :

- *Les acteurs de cette recherche.* A travers l'étude des laboratoires, des institutions de rattachement, des écoles doctorales, mais aussi du genre des doctorants, ce rapport répond en partie à la question *Qui produit la recherche sur l'architecture numérique en France ?* Si le doctorat en architecture est majoritairement porté par des doctorantes, cette tendance tend à s'inverser lorsque l'on parle de numérique. Les ENSA-P comptent parmi les institutions les plus impliquées, malgré des contributions parfois invisibilisées faute de détenteur d'HDR pour revendiquer les travaux doctoraux. Les structures de type UMR apparaissent comme les locomotives de la recherche en numérique, avec des laboratoires historiquement intéressés par les thématiques numériques qui tendent à être rejoints par de nouveaux acteurs depuis 2017, traduisant une appropriation de la culture numérique dans l'ensemble de la communauté de recherche.
- *La temporalité de cette recherche*. La place du doctorat en Numérique pour l'Architecture croît en proportion de la recherche doctorale en architecture, avec une répartition relativement stable et homogène des différentes sous-thématiques liées au numérique. Les doctorantes et doctorants conduisent leurs recherches sur typiquement 5 ans. L'intérêt pour les IA génératives ne se fait sentir qu'à l'échelle des doctorats en cours, traduisant peut-être une latence entre les thèmes de recherche et l'ouverture d'opportunités technologiques.
- *Les thématiques que ces acteurs abordent.* Le passage par une étape d'analyse de texte permet de s'abstraire des dimensions géographiques et politiques, pour dégager des tendances qui dépassent les frontières des laboratoires et les époques des doctorants. Si les analyses sémantiques proposées présentent leur part d'aléatoire - les noms de groupes thématiques pouvant changer au gré des modélisations – ces groupes font ressortir un intérêt scientifique ténu pour le lien entre numérique et conception paramétrique, simulation, représentation de l'information notamment spatiale et cartographique, notions de projet et de collaboration interacteurs, conservation du patrimoine, performance environnementale, gestion des risques,

Une étude approfondie des manuscrits de thèses en lien avec le Numérique pour l'architecture permettrait sans doute d'affiner la cartographie des thématiques de recherche. En ce sens, ont été rassemblés 78 manuscrits associés à la liste des thèses en numérique soutenues en ENSA. A ce stade, ils n'ont servi qu'à évaluer la longueur moyenne des manuscrits – qui semble significativement plus petite que celle des manuscrits d'autres thématiques de l'architecture – mais il paraitrait pertinent d'explorer le contenu textuel et iconographique de ces thèses, aujourd'hui considérées comme une « littérature grise » difficilement mobilisable. Le corpus de ces thèses sera mis à disposition sur demande pour poursuivre un éventuel travail d'analyse.

Cette étude reste un instantané des efforts de recherche, et tendra à devenir moins pertinente avec le temps. Il serait intéressant de mettre à profit certains des codes et méthodes développées ici pour construire un observatoire automatisé des travaux en Numérique pour l'Architecture. La recommandation serait (i) d'identifier les marqueurs que l'on souhaiterait suivre et de convenir d'un taux de rafraîchissement de la donnée (ii) d'assembler un *pipeline* simple pour récolter et analyser les thèses publiées en récupérant lesdits marqueurs, (iii) de proposer des modes de diffusion adéquats pour profiter et faire profiter de l'information dégagée sur les tendances de la recherche en numérique. A ce stade, les efforts de diffusion ont pris la forme d'une intervention de clôture à la conférence EDUBIM (nov. 2025) et d'une soumission d'article à la conférence SCAN (nov. 2026).

Pour transférer l'approche à d'autres sous-thématiques de l'architecture, et réaliser une cartographie comparable sur d'autres thématiques liées à l'architecture, l'investissement en temps pourrait être minime, et nécessiterait simplement d'adapter les critères d'agrégation d'un corpus, de lister les mots clés et sous-thématiques à suivre, et d'optimiser les paramètres des algorithmes de traitement automatique de la langue pour voir émerger des regroupements stables. En ce sens, l'ensemble des codes produits pendant la durée de ces deux missions est disponible à la réutilisation – et les auteurs se tiennent disponible pour aider à leur prise en main.

La lecture du Rapport ENSAECO ayant été un point de départ du présent travail, les auteurs seraient tout particulièrement curieux de tester la méthodologie sur les thématiques de la Transition Ecologique. Il est fort à parier que des sous-thématiques comme l'éco-conception, la nature en ville, les transformations du territoire, les ambiances climatiques, la gestion des ressources, pourraient émerger d'une analyse sémantique. Un regard croisé sur les différents corpus révèle des zones de recouvrement, entre la recherche sur les outils et la culture numériques et celle sur la transition écologique. Dans la lignée des initiatives Data4Good, côté associatif, ou de l'Ecolab /Greentech Innovation, côté institutions, cette zone de recouvrement pourrait être cultivée et étendue, pour mettre à profit les avancées du numérique dans nos efforts de protection de l'environnement.

Annexes et Scripts

L'ensemble des codes utilisés au cours des deux missions successives a été segmenté, commenté, mis au propre en vue d'éventuelles vérifications et réutilisation par des pairs. Pour chacun des segments, nous proposons ici un lien vers le code, que les auteurices du présent rapport pourront compléter par un temps de discussion à la demande. Chaque *notebook* présente une partie Prérequis, i.e. les packages à installer, puis une partie Chargement des données, qu'il conviendra de faire pointer vers le bon dossier, et enfin d'une ou plusieurs parties d'Expériences et d'Analyses sur les données, les plus structurées possibles.

Lien vers le travail d'analyse de la donnée :

https://colab.research.google.com/drive/1lw25sOiav7A9f4lcu6398oFlpn5n9l1t?usp=sharing

Lien vers le travail de classification non-supervisée :

https://colab.research.google.com/drive/1t1-Bbq9aLh5BTO9MbSjUjq0yxKwk8ggn?usp=sharing

Lien vers le travail de classification assistée par les LLM :

https://colab.research.google.com/drive/1Rv1N4a8hb9wVNuA2VRTHKKN5A9cr98rQ?usp=sharing

Références

1. Beza BB, Zeunert J, Kilbane S, Padgett Kjaersgaard S. Examining PhD modes in the Australian landscape architecture academy. Landsc Res. 4 juill 2022;47(5):679-94.

2. Sağlamer EG, Erkök F. Doctoral Education in Architecture. Cambridge Scholars Publishing. 2015.

3. Ilesanmi AO. Doctoral research on architecture in Nigeria: Exploring domains, extending boundaries. Front Archit Res. mars 2016;5(1):134-42.

4. Erbil Y, Gür M. A review of doctoral dissertations in architecture in Turkey. J Hum Sci. 30 juill 2018;15(3):1481.

5. ECOLAB. Référentiel général pour l'IA frugale. 28 juin 2024;

6. Luccioni S, Jernite Y, Strubell E. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In: The 2024 ACM Conference on Fairness, Accountability, and Transparency [Internet]. Rio de Janeiro Brazil: ACM; 2024 [cité 1 oct 2024]. p. 85-99. Disponible sur: https://dl.acm.org/doi/10.1145/3630106.3658542