Scraping d'Environnements Numériques de Travail et Analyse de Curriculums : un Cas d'Etude

Matthieu Cisel[1], et Camille Bidaud[2]

[1] CY Cergy Paris Université, Laboratoire AGORA, Cergy, France
matthieu.cisel@cyu.fr
[2] ENSA Paris-Val de Seine, Laboratoire EVCAU, Paris, France
camille.bidaud@paris-valdeseine.archi.fr

Résumé. Les recherches sur les curriculums sont souvent limitées par le caractère manuel de la collecte de données. Mobiliser des méthodes issues de l'informatique pour automatiser la collecte et le traitement de données sur des contenus d'enseignement offre un certain nombre d'opportunités en termes de passage à l'échelle des analyses, considération que nous nous proposons d'illustrer avec le cas des Ecoles Nationales Supérieures d'Architecture. Via des techniques fondées notamment sur le scraping de données issues d'Environnements Numériques de Travail, nous nous penchons dans cette réflexion méthodologique sur l'emploi du terme « numérique », dans les intitulés des cours de onze établissements. Nous suivons les évolutions de l'emploi des termes sur plus d'une décennie, et contrastons leur fréquence selon les établissements.

Mots-clés : Environnements Numériques de Travail, textométrie, curriculums, analyse de contenu, scraping.

Abstract. Curriculum research is often limited by the manual nature of data collection. Automating through computer science the collection and processing of data on educational content offers numerous opportunities, a consideration that we propose to take into account. illustrate with the case of the French Higher Schools of Architecture. In this article focused on methodology, we scraped data from Virtual Learning Environments to describe the use of the term "digital", in the course titles of eleven establishments. We follow developments in the use of terms over more than a decade, and contrast the situations across institutions.

Keywords: Virtual Learning Environment, textometry, curricula, content analysis, scraping.

1 Introduction

En sciences de l'éducation, l'analyse des curriculums et de leurs évolutions constitue un champ de recherche investi depuis plusieurs décennies, tant dans la littérature francophone que dans la littérature anglo-saxonne [4,5,6]. Dans le domaine de

l'enseignement scolaire, les chercheurs se sont focalisés sur des objets comme les manuels, ou les programmes tels que publiés au journal officiel [10,11]. Dans l'enseignement supérieur, il existe un nombre considérable de travaux [12], mais les analyses se heurtent notamment à la grande diversité des formations qui caractérisent l'offre d'une université ou d'une grande école. Là où au niveau du collège ou de l'école primaire, le programme des sciences de la vie et de la terre est identique en tout point du territoire national, celui des formations de biologie cellulaire, par exemple, pose davantage de problèmes pour des analyses comparatives à grande échelle entre établissements.

L'étude des différences de curriculums qui existent entre institutions se heurte à l'hétérogénéité des documents sur lesquels sont présentés les programmes pédagogiques : plaquettes, sites internet aux structures variées, programmes au format PDF. Constituer manuellement une base de données sur la base de tels supports représente un travail laborieux, d'où l'intérêt que présente des stratégies d'automatisation d'une telle tâche. Pour ce faire, nous défendons dans cet article l'idée selon laquelle les Environnements Numériques de Travail (ENT) représentent pour l'étude des curriculums une alternative pertinente aux supports de communication destinés à faire la promotion des formations.

D'après la CNIL, un ENT représente « tout ensemble intégré de services numériques choisis et mis à disposition de tous les acteurs de la communauté éducative d'un ou plusieurs établissements de l'enseignement scolaire ou de l'enseignement supérieur […] l'ENT constitue un point d'entrée unifié permettant à l'utilisateur d'accéder, selon son profil et son niveau d'habilitation, aux services et contenus numériques dont il dispose » [1].

Dans le champ des EIAH, les ENT ont fait l'objet d'une littérature abondante, centrée essentiellement sur le secondaire [2]. A notre connaissance, il n'existe pas de travaux portant sur l'utilisation de ces outils pour étudier des curriculums, à plus forte raison à l'échelle de plusieurs établissements. Pourtant, ils constituent des objets de choix pour ce type d'analyse, et ce pour plusieurs raisons. En premier lieu, toutes les formations de l'établissement sont accessibles via la même technologie ; l'ENT permet donc de constituer une base sur l'ensemble des cours sur plusieurs années, ce qui ouvre la voie à des analyses diachroniques. En second lieu, les établissements d'enseignement supérieur mobilisent en France un nombre relativement réduit de technologies différentes (Moodle, le plus fréquent, Sakai, Blackboard, etc.). Cela signifie que si l'on développe pour une université une méthodologie pour constituer une base sur les enseignements – dans le cas fréquent où l'accès à une base de données brutes avec tous les cours serait impossible, l'on peut facilement l'extrapoler à de nombreuses autres universités mobilisant la même technologie. Se pose toujours la question des droits d'accès aux données, mais cet obstacle administratif peut être surmonté, à condition d'avoir dans les établissements d'intérêt des contacts de collègues disposant eux de ce droit d'accès, et qu'ils soient disposés à collaborer avec les responsables de l'étude.

Parmi les nombreuses opportunités que présente pour la recherche une base de données d'intitulés de cours, et fondée sur l'exploitation d'ENT, l'on peut nommer l'étude de l'évolution des termes mobilisés pour décrire les enseignements. Celle-ci permet de capturer des tendances, et de se pencher sur les liens entre les injonctions politiques, les

[1] https://www.cnil.fr/fr/definition/ent-espace-numerique-de-travail

débats dans la sphère publique et leurs traductions éventuelles dans les curriculums. Par ailleurs, l'on peut s'intéresser aux différences entre curriculums selon une logique d'analyse comparative.

Le présent article décrit une étude exploratoire visant à répondre à une question de nature méthodologique illustrée par un cas d'étude, les Écoles Nationales Supérieures d'Architecture (ENSA) : Dans quelle mesure l'analyse de curriculums constitués sur la base de l'extraction de données depuis des ENT permet-elle de visualiser des évolutions de la popularité de certains sujets dans les intitulés de cours d'une part, et d'appréhender des différences entre établissements d'autre part ?

Nous nous proposons de répondre à cette question sur la base de l'analyse de données issues de Taïga, en nous focalisant sur l'enseignement de matières liées au numérique. Taïga est un ENT développé exclusivement pour les ENSA, et utilisé dans l'intégralité de ces institutions. Ce faisant, nous illustrons comment des méthodes issues de l'informatique – et en particulier le scraping et l'analyse textométrique – peuvent être utilisées dans les recherches sur les curriculums. Celles-ci mobilisent traditionnellement des analyses de contenu largement manuelles, alors que l'automatisation au cœur de notre étude nous a permis de couvrir aisément l'évolution de plus d'une décennie d'intitulés de cours pour onze ENSA.

2 Méthode

2.1 Un Environnement Numérique de Travail riche en informations

L'ENT Taïga a déjà été mobilisé pour caractériser l'évolution d'enseignements en architecture, mais uniquement sur la base de l'utilisation de son moteur de recherche [3]. Celui-ci permet d'effectuer des requêtes par mots-clés, et donc d'obtenir des données chiffrées sur le nombre d'occurrences d'un terme en particulier dans les fiches pédagogiques associées aux cours. Utiliser directement un tel moteur présente plusieurs limites : en premier lieu, elle impose de réaliser des analyses année par année, et donc de compiler manuellement les résultats si l'on souhaite appréhender des évolutions au fil du temps. En second lieu, le moteur de recherche le terme d'intérêt que dans des fiches descriptives pédagogiques remplies de manière inégale entre écoles, et qui comprennent le titre de l'unité d'enseignement (U.E.), mais aussi parfois une partie dédiée à la bibliographie.

Ainsi, il suffit qu'une référence bibliographique contienne le mot « numérique » dans le syllabus détaillé d'une U.E. portant sur un sujet qui n'est pas nécessairement connecté pour que le décompte la prenne en compte au même titre que la création de nouveaux enseignements mobilisant effectivement le numérique. Dans le meilleur des cas, un nettoyage laborieux est nécessaire. Par conséquent, ne pouvant accéder au jeu de données brutes sur lequel repose le moteur de recherche, nous avons reconstitué une base de données en capitalisant sur notre accès aux ENSA.

Certaines personnes disposant d'un accès administratif à Taïga peuvent en effet, pour leur institution, extraire l'intégralité des enseignements, et ce jusqu'à l'année scolaire 2013-2014. Nous avons contacté des collègues des différentes ENSA pour obtenir

ces informations via des documents au format html ou csv. Un html est typiquement produit pour un niveau (licence ou master), pour une année donnée. Ce faisant, nous avons compilé les données de 11 ENSA, nommément Belleville, Bordeaux, Bretagne, La Villette, Lyon, Malaquais, Marne, Normandie, Nantes, Val de Seine, Toulouse. Les documents html reprennent l'ensemble des titres des U.E. et des différents enseignements. Les documents csv peuvent venir compléter la base, après jointure, lorsque les enseignements sont peu détaillés dans l'html.

2.2 Choix des termes constituant la focale de l'étude

Nous avons ensuite compté les occurrences du terme « numérique » dans les enseignements, selon une méthodologie que nous précisons par la suite, après avoir exploré d'autres termes comme « digital », et « ordinateur ». Ce terme a été choisi suite à l'exploration des données. Une exploration préliminaire de termes synonymes ou appartenant au même champ sémantique, tels que *digital* ou *ordinateur*, n'a en effet pas permis d'observer de tendance significative : leurs occurrences étaient trop rares pour fournir des résultats exploitables ou pour révéler une évolution comparable. En ce sens, le terme *numérique* apparaît comme le plus représentatif de mutations des enseignements dans le contexte francophone, notamment pour suivre les évolutions discursives et institutionnelles autour de l'enseignement de l'architecture.

S'agissant du choix du terme au centre de l'étude, il convient de noter que l'une des difficultés posées par la focalisation sur les intitulés de cours réside dans le fait qu'il est difficile d'identifier si des évolutions – ou des différences entre établissements – relèvent de modifications substantielles des maquettes pédagogiques ou s'il s'agit avant tout de nuances terminologiques. Il peut y avoir simplement substitution des termes dans les curriculums, résultat constaté dans un autre article pour le terme *transition*, qui a remplacé peu à peu le terme *durable*, présent notamment dans *développement durable* [1]. Or une telle évolution terminologique reflète sans doute une question de tendance, de « buzzword », qu'il est difficile de distinguer, par ces méthodologies, d'un changement significatif dans les contenus présentés aux étudiants.

En théorie, une évolution du nombre des enseignements avec l'intitulé *numérique* peut aussi bien refléter une fluctuation, à l'échelle nationale, de la priorité donnée à la maîtrise d'outils digitaux dans les cours des ENSA, que de changements dans la manière de décrire les cours. Néanmoins, en recherchant *digital*, *informatique*, *ordinateur*, *modèle*, aucun terme ne semble être très utilisé, ou compenser d'évolution de *numérique*, contrairement au phénomène de vase communicant entre *durable* et *transition*. Il existe des termes spécifiques à l'architecture comme « AutoCAD », logiciel dédié à la conception en 3D, mais seulement de manière anecdotique. L'expérience acquise par l'un des co-auteurs suggère que dans les cours en architecture, les étudiants peuvent être amenés à utiliser des outils « numériques » dans des enseignements centrés, dans leur intitulé, sur la géométrie, sans qu'aucun emploi d'un synonyme de *numérique* ne puisse déceler une telle pratique. C'est là la limite des analyses par intitulé de cours.

En se contentant de l'intitulé de cours, l'on ne peut pas voir des évolutions en profondeur d'un curriculum. Mais comprendre l'évolution en profondeur du contenu d'un cours n'est pas possible à l'échelle d'une décennie, sur l'intégralité des cours d'une

dizaine d'écoles. Un décompte de mots-clés permet de construire une première série d'indicateurs. C'est une première étape pour appréhender des évolutions en cours, qu'il faut bien évidemment approfondir par la suite pour éviter les erreurs d'interprétation.

2.3 Collecte et analyse des données

R 4.0 a été mobilisé tant pour extraire des fichiers html et csv, par scraping (librairie *rvest*) le contenu d'intérêt que pour l'analyse textométrique qui a suivi. R est un langage de programmation généraliste particulièrement pertinent pour les statistiques, mais aussi efficace pour les tâches simples de scraping, au même titre que Python.

Le scraping (ou moissonnage de données) désigne l'ensemble des techniques permettant d'extraire automatiquement des informations accessibles sur le web ou dans des bases de données [7]. En contournant l'accès manuel, le scraping facilite la collecte massive de contenus. Il faut interagir avec la structure HTML d'une page pour en extraire textes, liens, métadonnées, etc. Lors d'une opération de scraping, le choix des balises HTML est crucial : ce sont elles qui structurent l'information. Sélectionner les bonnes balises permet de cibler précisément les contenus pertinents et d'éviter de récupérer du bruit (menus, bas de page). Toutes années confondues, nous avons collecté par cette voie des données sur 10807 enseignements en licence et 20915 en master.

Une fois les données collectées, l'analyse textométrique offre des outils pour explorer systématiquement les contenus textuels. Issue de la linguistique quantitative et de l'analyse de corpus, la textométrie vise à objectiver l'étude des textes en quantifiant leur structure et leur lexique. L'un des indicateurs de base est la fréquence des termes (term frequency, TF), qui mesure combien de fois un mot apparaît dans un corpus donné, en l'occurrence une base de données d'intitulés de cours, et qui permet de produire des vecteurs de mots. Ce comptage simple, fait avec R au travers d'une approche de type « expression régulière », permet de repérer les thématiques dominantes, ou l'évolution d'une thématique d'intérêt. De là, l'on peut identifier les mots les plus saillants, ou encore de construire des représentations visuelles comme des nuages de mots ou des cartes de cooccurrence. Nous avons mis en place une approche empêchant des redondances : le décompte du terme « numérique » est fait sur la seule base de l'intitulé des cours, et ne prend pas en compte une éventuelle apparition dans d'autres éléments descriptifs (tels que les ressources bibliographiques), contrairement à ce qui a pu être fait dans des études antérieures [3].

Le calcul de la TF est souvent utilisé seul ou combiné avec d'autres métriques, comme le TF-IDF (term frequency–inverse document frequency), qui pondère la fréquence d'un mot par sa rareté dans l'ensemble du corpus, permettant ainsi de distinguer les termes véritablement spécifiques d'un texte [8]. C'est par exemple ce type de métriques qui a utilisé par Fortino *et al.* [9] pour mesurer l'adéquation entre les termes mobilisé dans un curriculum (intitulés de cours) et les titres de postes auxquels postulent les étudiants de différentes formations d'universités américaines. Néanmoins, dans la mesure où nous souhaitons comparer l'évolution d'un terme donné au fil des ans, la mobilisation de la métrique TF a semblé plus pertinente ici.

3 Résultats

Dans cette section, nous présentons d'une part une analyse diachronique de l'emploi du terme « numérique », tous établissements confondus, et d'autre part une comparaison entre établissements entre licences et masters. En Figure 1, nous constatons un emploi croissant du terme tant en licence qu'en master, avec un pic autour de 2020 correspondant à plus de soixante-dix occurrences, suivie d'une baisse progressive et limité de son emploi. Il y a un décalage d'un an dans les évolutions entre masters et licences.

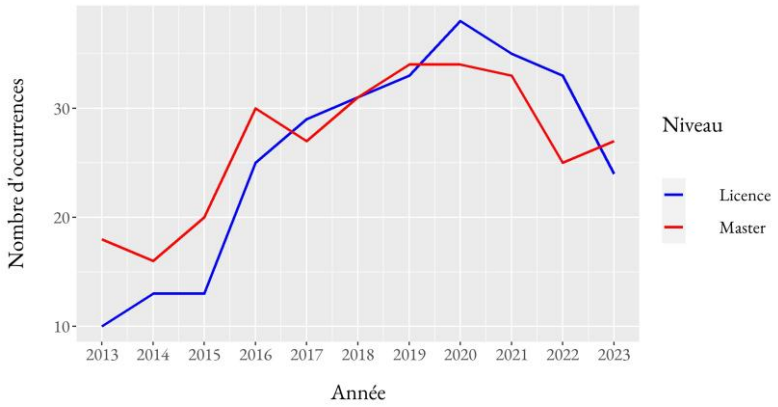

Fig. 1. Analyse diachronique de l'utilisation du terme « numérique » dans les intitulés d'enseignements de 11 ENSA de 2013 à 2023. On notera que 2013 correspond à l'année 2013-2014

L'analyse permet également de visualiser des différences entre établissements (Figure 2). On constate par exemple que les ENSA Bretagne et La Villette n'emploient pas le terme « numérique » dans les cours de licence. Pour certains établissements comme l'ENSA Bordeaux, le nombre d'occurrences est supérieur pour le niveau licence, pour d'autres, comme Val de Seine, il est supérieur au niveau master, reflétant la place que les instances décisionnaires veulent voire le numérique occuper dans les enseignements dans la séquence de cours.

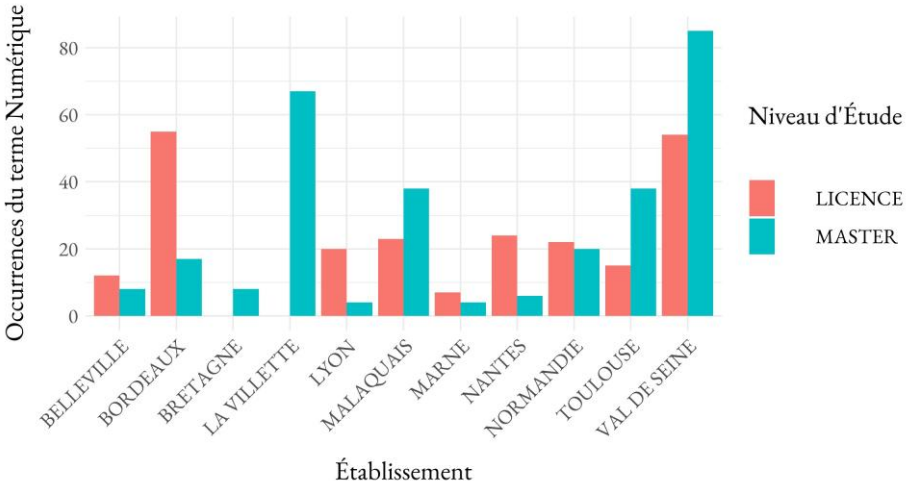

Fig. 2. Utilisation, différenciée selon le niveau et l'établissement, du terme « numérique » dans les intitulés d'enseignements de onze ENSA de 2013 à 2023.

4 Discussion

L'expérience d'enseignement acquise par l'un des co-auteurs à La Villette et dans d'autres ENSA suggère que l'absence ou la faible représentation du terme *numérique* en licence ne découle pas d'une simple évolution terminologique. C'est davantage le reflet d'un choix délibéré des établissements de retarder le plus possible l'utilisation des outils numériques dans le cursus de l'école. Dans la mesure où il s'agit d'une preuve de concept d'une approche méthodologique, nous ne souhaitons pas ici analyser plus avant l'historique de l'utilisation d'un terme comme *numérique*, ni les déterminants des évolutions observées. Pour un travail centré sur l'interprétation des tendances observées, nous renvoyons à un article basé sur la présente base de données, et consacré à l'enseignement de la réhabilitation [1], et aux recherches consacrées aux déterminants des évolutions [4].

Le recours au scraping des Environnements Numériques de Travail (ENT) pour collecter les intitulés de cours constitue une approche méthodologique que nous pensons innovante pour la conception d'indicateurs relatifs à l'évolution des curricula. L'extraction systématique de ces intitulés permet de construire des indicateurs empiriques, fondés sur la fréquence d'occurrence de certains termes (par exemple numérique), témoignant de l'intégration progressive de thématiques émergentes dans les formations, ou d'évolutions dans les termes utilisés. Toutefois, plusieurs limites doivent être soulignées. D'abord, les intitulés captés ne garantissent pas une exhaustivité parfaite : certaines informations peuvent être absentes, obsolètes ou partiellement mises à jour dans

les ENT. Ensuite, comme nous l'avons souligné, l'intitulé seul ne renseigne qu'indirectement sur les contenus effectifs des enseignements, exposant l'indicateur à un risque de surinterprétation. Enfin, l'hétérogénéité des pratiques d'intitulé selon les établissements ou les enseignants introduit une variabilité qui doit être méthodologiquement contrôlée (par normalisation ou codage thématique). Malgré ces limites, le scraping des ENT offre une opportunité rare d'observer, avec une granularité temporelle et institutionnelle fine, la dynamique d'institutionnalisation de nouveaux savoirs, et de documenter les évolutions discursives de l'offre pédagogique.

Toute conception d'indicateur est confrontée à des limites consubstantielles, qui tiennent à la nature même de l'opération de mesure. Un indicateur ne capture qu'une traduction partielle et construite du phénomène étudié : il repose nécessairement sur des choix de définition, de périmètre et de catégorisation qui introduisent un écart entre la complexité du réel et sa représentation. Par ailleurs, tout indicateur simplifie les dynamiques sous-jacentes en les ramenant à des éléments quantifiables, parfois au prix d'une perte de sens contextuel ou qualitatif.

Dans cette perspective, l'indicateur, et les représentations graphiques que nous en avons faites, doivent être compris non comme un aboutissement, mais comme une première étape dans l'analyse : ils permettent de repérer des phénomènes émergents, des inflexions ou des tensions qu'il s'agit ensuite de creuser par des méthodes complémentaires (enquêtes qualitatives, analyses de contenu, observations de terrain).

5 Conclusion

Si notre approche fondée sur les intitulés de cours permet de capturer certaines tendances à l'œuvre, elle reste superficielle dans la mesure où elle ne permet pas nécessairement d'appréhender d'évolutions significatives quant aux contenus enseignés, puisqu'elle n'entre pas dans le détail du contenu des enseignements. Les fiches pédagogiques contiennent de nombreux autres éléments qui peuvent également être exploités, à condition d'une homogénéisation des pratiques de remplissage de ces fiches. La bibliographie recommandée aux étudiants constitue par exemple une rubrique présentant davantage d'informations sur le contenu des enseignements que le simple intitulé du cours. Une analyse à grande échelle d'une base focalisée sur les seules références fournies aux étudiants permettrait ainsi de réaliser des analyses plus approfondies.

6 Références

1. Bidaud, C., et Cisel, M. (2024). Une décennie d'essor des enseignements sur l'existant dans les ENSA vue à travers les intitulés de cours. *Les Cahiers de la recherche architecturale urbaine et paysagère*, *22*. https://doi.org/10.4000/13354
2. Bruillard, É. (2011). Le déploiement des ENT dans l'enseignement secondaire: entre acteurs multiples, dénis et illusions. *Revue française de pédagogie*, *177*, 101-130.
3. Cremnitzer, J.-B. (dir.), Valter Balducci (coord*.), Former à la réhabilitation, opus cit.,* p. 14
4. Krause, K.-L. D. (2022). Vectors of change in higher education curricula. *Journal of Curriculum Studies*, *54*(1), 38–52. https://doi.org/10.1080/00220272.2020.1764627

5. Lemaître, D. (2009). Le curriculum des grandes écoles en France: Un modèle d'analyse inspiré de Basil Bernstein. *Revue française de pédagogie. Recherches en éducation*, *166*, Article 166. https://doi.org/10.4000/rfp.1096
6. Martinand, J.-L. (2014). Point de vue V – Didactique des sciences et techniques, didactique du curriculum. *Éducation et didactique*, *8*(1)
7. Mitchell, R. (2024). *Web Scraping with Python: Collecting Data from the Modern Web*. *3rd Edition.* O'Reilly Media.
8. Ramos, J. (2003, Janvier). Using TF-IDF to Determine Word Relevance in Document Queries. Dans *Proceedings of the First Instructional Conference on Machine Learning*. (pp 24-37).
9. Fortino, A., Zhong, Q., Huang, W. C., & Lowrance, R. (2019, Mars). Application of text data mining to stem curriculum selection and development. Dans *2019 IEEE Integrated STEM Education Conference (ISEC)* (pp. 354-361). IEEE.
10. Denizot, N. (2016). Le manuel scolaire, un terrain de recherches en didactique?. *Le français aujourd'hui*, *194*(3), 35-46.
11. Bruillard, É. (dir.) (2005). *Manuels scolaires, regards croisés*. Caen : CRDP de Basse-Normandie.