De la collecte des traces d'interaction à la publication scientifique

Matthieu Cisel

Distances et médiations des savoirs

Distance and Mediation of Knowledge

37 | 2022
Varia
Retour d'expérience

De la collecte des traces d'interaction à la publication scientifique

Un passage à l'échelle semé d'embûches

Scaling up the collection of learning analytics and the publication of research articles: a sinuous path

Matthieu Cisel

https://doi.org/10.4000/dms.6953

Résumés

Français English
Dans cette contribution prenant la forme d'un retour d'expérience, nous revenons sur le rôle que l'essor des MOOC a joué sur la multiplication des travaux mobilisant les traces d'interaction (*learning analytics* en anglais). En atteignant des échelles inégalées en termes de nombre d'apprenants suivis à la trace, les études fondées sur ces données ont gagné en légitimité et ont pu atteindre les revues scientifiques les plus prestigieuses. Nous arguons du fait que ce succès doit inviter à développer des travaux visant à étudier le comportement d'un nombre important d'utilisateurs de plateformes en ligne. De par la richesse et le volume des données qu'elles produisent, des plateformes d'enseignement françaises pourraient donner aux recherches hexagonales une dimension internationale. Après avoir évoqué FUN MOOC, nous illustrons notre propos en nous penchant sur les LMS comme Moodle, prenant l'exemple de M@gistère. Les ambitions dans ce domaine se heurtent néanmoins à une série d'obstacles, notamment techniques, juridiques ou éthiques. Le fil rouge que nous suivons pour illustrer des analyses ambitieuses réside dans l'articulation entre le comportement d'inscription des utilisateurs au sein d'un catalogue de cours, et leur utilisation de la formation une fois l'inscription réalisée.

In this contribution, which takes the form of a return of experience, we return to the role that the rise of MOOCs has played in the multiplication of studies mobilising interaction traces (learning analytics in English). By reaching unprecedented scales in terms of the number of learners tracked, studies based on these data have gained legitimacy and have been able to reach the most prestigious scientific journals. We argue that this success should invite further work to study the behaviour of large numbers of users of online platforms. Because of the wealth and volume of data they produce, French educational platforms could give French research an international dimension. After mentioning FUN MOOC, we will illustrate our point by looking at LMSs such as Moodle, taking the example of M@gistère. Ambitions in this field nevertheless come up against a series of obstacles, notably technical, legal and ethical. The common thread that we follow to illustrate ambitious analyses lies in the articulation between the registration behaviour of users within a course catalogue and their use of the training once they have registered.

Entrées d'index

Mots-clés : traces d'interaction, MOOC, LMS, éthique, Big Data, learning analytics
Keywords: interaction traces, MOOC, LMS, ethics, Big Data, learning analytics

Texte intégral

Introduction

1 Force est de constater que l'avènement des MOOC tout au long de la décennie 2010 a considérablement contribué à banaliser les travaux fondés sur les traces d'interaction tout en offrant des perspectives sur des projets d'une ampleur inégalée (Cisel, 2016). Comme le souligne Cherigny *et al.* (2020), « les traces regroupent l'ensemble des interactions d'un usager avec son environnement d'apprentissage. Sur les principales plateformes d'e-Learning – LMS et plateformes de MOOCs –, les analystes fouillent donc données et traces de la présence d'un apprenant (logs sur les pages, durées de consultation) » (p. 7). Le terme « traces » permet ainsi de désigner toute forme d'interaction, et non uniquement celles qui révèlent un processus d'apprentissage. Des inscriptions sur la plateforme FUN MOOC constituent ainsi des traces, et ne révèlent presque rien de ce qui est appris de l'apprenant. Au passage, quand l'on s'intéresse davantage au processus d'apprentissage, on utilise souvent le terme *Learning Analytics*. Comme le rappellent toujours Cherigny *et al.* (2020), « le terme Analytics désigne de façon usuelle des techniques informatiques, mathématiques et statistiques pour révéler une information pertinente à partir de très larges ensembles de données. Par extension, les Analytics permettent, sur la base d'actions réalisées, de comprendre, voire de prédire, le potentiel de futures actions dans une quête de performance et d'efficience. » (p. 5). Dans la mesure où nous nous intéressons à toute sorte d'interactions, nous préférerons ici le terme « traces », plus générique.

2 Avec les premiers cours en ligne du MIT sur edX (Breslow *et al.*, 2013), plus de cent mille apprenants issus de tous les continents sont suivis à la trace. Lorsque les analyses portent sur la plateforme américaine edX issue de la collaboration de Harvard et du MIT, on franchit rapidement le million d'utilisateurs (Ho *et al.*, 2015), et les publications dans les plus grandes revues scientifiques généralistes anglophones comme PNAS et Science s'ensuivent (Kizilcec *et al.*, 2020a., Kizilcec *et al.*, 2020b). Ce succès est d'autant plus marquant que la question de l'éducation, et *a fortiori* l'éducation en ligne, ne représente traditionnellement pas pour celles-ci une thématique de prédilection.

3 Certes, le prestige des institutions à l'origine de telles publications a probablement joué dans le succès de telles aventures scientifiques, mais il ne faut pas sous-estimer la question de la portée des dispositifs analysés dans la réussite des projets, la portée étant mesurée en nombre d'apprenants concernés. En effet, il est fréquent, lorsque l'on se penche sur le

fonctionnement d'un dispositif de taille réduite comme un cours ou un programme réunissant quelques dizaines ou quelques centaines d'apprenants, que les évaluateurs des revues demandent de justifier l'intérêt scientifique que revêt une telle étude de cas. La préoccupation est légitime. Une étude de la composition sociodémographique des utilisateurs d'un MOOC donné et de l'engagement de ses apprenants peut certes être publiée (Liyanagunawardena *et al.*, 2015), mais son intérêt scientifique peut être considéré à juste titre comme limité par les examinateurs d'une revue au fur et à mesure que la littérature sur ce type de sujet s'étoffe.

4 Lorsque l'on mène une analyse de traces sur des plateformes rassemblant plusieurs dizaines ou plusieurs centaines de milliers d'apprenants, le simple saut quantitatif peut suffire en lui-même à écarter la critique que des examinateurs pourraient faire sur la portée du cas d'étude. Par exemple, lorsque l'on s'intéresse à la manière dont l'ensemble des enseignants d'un pays sont formés à distance, il devient facile de justifier auprès des revues scientifiques l'intérêt d'une telle démarche. La même analyse de traces, réalisée cette fois au sein d'un simple cours d'une université effectué à distance, pourra en revanche être plus facilement rejetée.

5 Par ailleurs, les travaux suscitent d'autant plus l'intérêt des décideurs que le nombre d'apprenants concernés est élevé. Si les recherches ne visent pas à proprement parler à évaluer des politiques publiques, elles peuvent offrir des retours utiles aux administrateurs des plateformes et à leurs financeurs. Partant de là, il devient plus facile d'obtenir des financements et d'enclencher un cercle vertueux favorisant la pérennisation d'un programme de recherche. Néanmoins, les obstacles juridiques, techniques, éthiques sont nombreux et s'opposent à la mise en place de telles dynamiques ascendantes. Sur la base de projets personnels passés ou à venir, nous illustrons les différents obstacles qui peuvent entraver les collaborations entre chercheurs et administrateurs de plateformes.

Traces d'interaction et analyse de données : trois cas d'étude pour illustrer la question du passage à l'échelle

6 La plupart des plateformes de cours en ligne disposent d'outils d'analyse de trafic qui permettent aux gestionnaires de suivre la fréquentation de tel ou tel cours à différentes échelles temporelles, et ce sans avoir à analyser les jeux de traces d'interaction. Des outils comme Xiti, Kibana ou Google Analytics fournissent ce type d'information en temps réel aux administrateurs et aux concepteurs de cours. Notre expérience personnelle suggère que bien souvent ces informations sont considérées comme suffisantes par les administrations, qu'il faut alors convaincre de l'intérêt de travailler avec un laboratoire pour aller plus loin. Ces outils de suivi n'offrent qu'un aperçu superficiel des dynamiques à l'œuvre, d'une part car ils ne permettent pas de suivre un individu à la trace et, d'autre part, car ils ne contiennent pas d'indication quant aux caractéristiques des utilisateurs. Il n'est par exemple généralement pas possible de distinguer les apprenants selon leur âge ou leur catégorie socioprofessionnelle. Or ce type de distinction est crucial pour appréhender les utilisations qui sont faites de la plateforme.

7 Pour illustrer une analyse intéressante que seule une étude approfondie des traces permettrait de faire, nous proposons de prendre comme fil rouge le lien entre comportement d'inscription et comportement dans un cours. Le premier désigne l'analyse des actions effectuées par les participants en termes d'inscription à des cours de la plateforme, date, durée, rythme, nature des cours choisis. Si l'apprenant est inscrit par autrui (administrateur de la plateforme, enseignant, etc.) dans le cadre d'une formation prescrite, on peut émettre l'hypothèse qu'il ne s'investira pas de la même manière dans un

cours que s'il s'inscrit de sa propre initiative, sans contrainte quelconque. Reste à le prouver. Dans les sections qui suivent, nous revenons sur la manière dont les données de FUN MOOC sont susceptibles de fonder des analyses pertinentes du comportement d'inscription tant pour leurs gestionnaires que pour la communauté scientifique.

Recouper utilisation des cours et comportement d'inscription dans les MOOC

8 Les MOOC reposent par définition sur le principe de l'auto-inscription, la majorité des apprenants s'inscrivant au cours de leur propre initiative (Cisel, 2018 ; Vrillon, 2019). L'analyse des comportements d'inscription représentent au sein des MOOC une piste de recherche fertile, car elle permet de mieux appréhender comment des apprenants, dont la plupart ne subissent pas la pression de leurs pairs ou de la hiérarchie, s'approprient un catalogue de cours. Ce type de recherche avait émergé dans les rapports d'activité de Ho *et al.* (2015). Les auteurs avaient analysé les recouvrements d'audience entre cours de Harvard et du MIT sur des MOOC de edX. Ils n'avaient néanmoins pas pu travailler sur l'intégralité des inscriptions réalisées sur la plateforme edX, du fait d'obstacles sur lesquels nous reviendrons. À notre connaissance, avant nos propres travaux Albo *et al.* (2016) représentaient les premiers auteurs à avoir étudié le comportement d'inscription à l'échelle d'une plateforme complète, *MiriadaX*, mais avant tout sous l'angle sociodémographique.

9 Nos recherches portant sur plus d'un million d'inscriptions de la plateforme FUN avaient permis d'analyser le phénomène de dispersion des inscriptions, susceptible d'affecter des indicateurs de performance tels que les taux de complétion (Cisel, 2018 ; Wintermute, Cisel, et Lindner, 2021). La principale faiblesse dans ce type d'analyse réside dans le fait que, à cause de l'absence de jeux de données complets, nous n'avons jamais réussi à croiser le comportement d'inscription à l'échelle de la plateforme, et le comportement au sein d'un cours. Tout au plus disposons-nous de l'information selon laquelle l'apprenant a obtenu l'attestation de réussite. En revanche, nous ne savons pas ce qu'il y a fait, ni même s'il s'y est connecté.

10 Là encore, des considérations de toutes sortes – juridiques et éthiques notamment – font obstacle à ce type d'analyse en France comme outre-Atlantique. En effet, les données de comportement à l'échelle d'un cours ne sont que rarement transmises aux équipes de recherche. Par ailleurs, les MOOC relèvent de l'apprentissage informel ; les résultats que l'on pourrait obtenir sur la base des données des plateformes qui les hébergent sont difficilement extrapolables à des contextes comme ceux de la formation initiale ou de la formation tout au long de la vie traditionnelle, c'est-à-dire au sein de dispositifs institués. Cette considération nous a poussé à rechercher des sources de données permettant de poursuivre le type d'analyse que nous avons suggéré, mais cette fois dans un contexte d'éducation formelle. À défaut d'avoir poussé la démarche jusqu'à son terme, nous avons identifié deux sources de données intéressantes. Les premiers sont les plateformes universitaires basés sur Moodle, les seconds sont les données d'une plateforme s'adressant aux personnels de l'Éducation nationale. : M@gistère. Nous n'avons pas analysé les données à proprement parler, mais les échanges avec le ministère de l'Éducation nationale nous ont permis d'identifier les opportunités de recherche.

Traces d'interaction de Moodle : potentialités et exemples concrets

11 La plupart des universités françaises disposent d'un Système de Gestion de l'Apprentissage (SGA) – *Learning Management System* (LMS) en anglais – basé sur la technologie *open source* Moodle. Ce LMS vise notamment à mettre à disposition les ressources pédagogiques d'un cours. Si Blackboard a longtemps fait la course en tête dans les LMS universitaires, la solution open source Moodle s'est en effet imposée à partir des années 2000, sans jamais pour autant occuper une place hégémonique : des plateformes comme Sakai ou Claroline Connect continuent à exister.

12 À partir des traces de Moodle, il est facile de générer des rapports adaptables aux besoins d'une institution. Les codes de rapports types, modifiables en quelques lignes par le langage SQL, sont mis à la portée de tous (Zdravev *et al.*, 2018). Les réflexions déjà anciennes sur l'apport des LMS universitaires dans le champ de la formation à distance (Bonu et Charnet, 2007) peuvent désormais s'étoffer d'analyses sophistiquées grâce à ces traces. Ainsi, Zhang *et al.* (2020) utilisent les traces laissées par des étudiants sur Moodle pour prédire leurs performances aux examens. Il existe des tables compilant les utilisations pour chaque module de la plateforme – forums, quiz, activités évaluées ou non évaluées, ce qui ouvre de nombreuses possibilités en termes d'analyses.

13 Si l'on veut gagner en ambition en matière d'analyse et que l'on cesse de se concentrer sur les données Moodle d'une institution prise seule, il faudrait que les établissements français mutualisent leurs jeux de données. Certes, nous faisons face au fait que les versions de Moodle diffèrent d'un établissement à l'autre, et de ce fait, la nature des données collectées. Mais cet obstacle a déjà été franchi dans certains projets nationaux. Dans le cas de M@gistère, chaque académie dispose d'une instance Moodle qui lui est propre, ce qui n'empêche pas une centralisation des traces lorsque la direction du projet en manifeste la nécessité.

14 Lancée en 2013, M@gistère propose une offre de cours en ligne à l'ensemble des personnels de l'Éducation nationale ; des centaines de milliers d'utilisateurs sont donc concernés. Si des travaux lui ont été consacrés (Pogent, Albero et Guérin, 2019 ; Pogent, 2020), ils se focalisaient davantage sur les expériences des enseignants que sur les utilisations de la plateforme. Une recherche fondée sur des analyses quantitatives compléterait utilement les perspectives offertes par des recherches qualitatives menées par Frédéric Pogent, et permettrait d'enrichir le corpus de connaissances disponibles sur ce que les Anglo-saxons nomment les « *online teacher professional development (oTPD) programs* » (Dede *et al.*, 2009), ou programmes de formation continue en ligne des enseignants.

15 Les opportunités offertes par Internet pour la formation des enseignants ont été soulignées dès le début des années 2000 dans de nombreux ouvrages (Perraton, 2002 ; Latchem et Robinson, 2003) et travaux de recherche (King, 2002). Comme Dede *et al.* (2009) le rappellent, passer par Internet permet de s'adapter aux calendriers chargés de la profession et de donner accès à des ressources et personnes-ressources qui ne seraient pas disponibles localement. Des travaux de recherche ont été consacrés à des programmes nationaux, voire internationaux (Bof, 2004). La réflexion sur le sujet a été ravivée par l'essor des MOOC au début des années 2010 dans la mesure où de nombreux enseignants ont mobilisé ces cours pour se former (Hodges, Lowenthal, et Grant, 2016). Plus récemment, la crise liée au COVID a incité de nombreux chercheurs à se pencher sur la manière dont les systèmes éducatifs s'adaptaient à la situation sanitaire, notamment au travers d'outils digitaux pour la formation des enseignants (Lockee, 2021).

16 Dans leur revue de la littérature portant sur quarante articles, Dede *et al.* (2009) soulignent l'importance qu'il y a à identifier les bonnes pratiques d'une part, et les impacts

à long terme des formations sur les pratiques dans la classe d'autre part. Ils pointent le caractère anecdotique de beaucoup des travaux réalisés sur la formation en ligne des enseignants ainsi que l'importance de gagner en ambition. Jusqu'à présent, les études à grande échelle sur le sujet se sont fondées sur des questionnaires et ont visé à identifier avant tout l'utilité perçue des programmes (Smith et Sivo, 2011), sans se préoccuper des utilisations effectives des cours. À notre connaissance, il est rare que soient mobilisées des traces pour suivre à grande échelle ce processus d'appropriation. Whitaker *et al.* (2007) ont certes analysé le comportement de 213 enseignants sur la base de ces traces, mais les projets portant sur une proportion significative du corps enseignant d'un pays font à notre connaissance encore défaut.

17 L'importance que M@gistère au niveau national en matière de formation des praticiens fait de cette plateforme un terrain de choix pour imaginer les opportunités offertes par des analyses à grande échelle de traces. Il serait ainsi possible de contraster les modes d'appropriation des ressources selon le niveau considéré (école primaire, collège, lycée), selon l'ancienneté au sein de l'institution, etc. Parmi la multitude des analyses possibles, dans l'éventualité d'une collaboration avec la plateforme, nous proposons de maintenir le fil rouge qu'est la prise d'initiative dans le choix des formations.

18 Si dans M@gistère, les inscriptions dans des cours prescrits sont réalisées le plus souvent par des gestionnaires ou par les concepteurs d'un cours, les enseignants ont également la possibilité de parcourir de nombreux cours de l'offre de formation de leur académie. À défaut de pouvoir obtenir une reconnaissance officielle, ils peuvent y réaliser les activités afférentes (visionnage de vidéos, quiz, etc.). Le comportement d'auto-inscription dans des cours non prescrits pourrait ainsi offrir un point de vue intéressant sur l'appétence des enseignants pour l'apprentissage en ligne sur des ressources institutionnelles, et sur les déterminants de cette appétence. Ceci est d'autant plus vrai que la plateforme est basée sur Moodle, on peut donc distinguer l'identité de l'apprenant et du prescripteur.

19 On voit ici que le cas de M@gistère est un peu particulier, puisqu'il s'agit d'adultes engagés dans la vie professionnelle et suivant des formations prescrites. Le contexte est différent de FUN MOOC ou de celui de la formation initiale. Néanmoins, l'organisation centralisée, notamment en matière de collecte de données, pourrait servir de modèle pour les universités qui voudraient mutualiser leurs efforts en matière d'hébergement et de traitement des données.

20 Les sources de données potentielles ne manquent pas, d'autant que nous n'avons cité ici que quelques exemples parmi la multitude des projets portés par le service public français. Et pourtant, les publications hexagonales sur les traces d'apprentissage sont rares à l'international. Nous proposons dans la section qui suit de mieux appréhender les raisons de cet état de fait.

Des questions de communication aux obstacles juridiques, le chemin de croix de la recherche sur les traces d'interaction

21 Dans cette partie, nous allons dénombrer quatre types d'obstacles fréquents dès lors que l'on s'engage dans des projets à grande échelle d'analyse de traces. L'importance de tel ou tel obstacle varie selon la nature du projet. Certains sont devenus plus saillants au fil des années. Avec le RGPD, la prudence est devenue de mise et la transmission de données par les administrations ne se fait qu'au compte-goutte, après s'être assuré de la légalité des

analyses, et après qu'un lien de confiance se soit établi entre administrateurs et chercheurs. De l'autre côté du traitement des données, en bout de chaîne lors des étapes finales de la publication des données, les revues scientifiques sont de plus en plus nombreuses à exiger une validation de l'étude par un comité d'éthique. Aboutir à une publication relève dès lors du parcours du combattant.

Le partage des traces, une question sensible pour toute institution

22 Concernant les problèmes liés à la communication institutionnelle, nous allons illustrer notre propos avec les MOOC, pour ensuite l'élargir sur d'autres types de dispositifs. Comme nous le soulignions en introduction, dès le lancement des premiers MOOC, les analyses fondées sur leurs traces d'interaction fleurissent, au point de devenir l'une des approches les plus plébiscitées dans les travaux sur ces cours en ligne (Breslow *et al.*, 2013). Un premier problème vient généralement entraver le travail des auteurs de rapports d'utilisation ou d'articles de recherche qui souhaitent agréger les données de nombreux cours. Si les plateformes qui centralisent les données de dizaines d'institutions disposent de l'intégralité des données sur les apprenants, ce sont généralement ces institutions partenaires qui décident si elles transmettent ou non ces données. Un double accord est nécessaire avant toute transmission : celui de la plateforme, et celui de l'institution à l'origine du cours. Par exemple, dans le cas de FUN MOOC, les traces d'interaction laissées par les apprenants au sein d'un MOOC du CNAM ne pourront pas être transmises aux chercheurs de l'Université de Paris, à moins qu'un accord spécifique ne soit signé.

23 Pour avoir une vue globale de l'utilisation des cours au sein d'une plateforme, il faudrait ainsi obtenir l'accord de l'intégralité des dizaines voire centaines d'établissements. C'est une tâche insurmontable sur le plan pratique. Les institutions d'enseignement supérieur peuvent être réticentes à ce que des chercheurs d'établissements jugés comme concurrents aient un regard sur le succès du déploiement de leurs projets de cours en ligne. On se souvient d'ailleurs qu'une logique de bataille de chiffres s'était instaurée au début du mouvement MOOC, et que la connaissance des données était jugée comme stratégique (Cisel, 2016). On constate ainsi que les rapports d'auteurs comme Ho *et al.* (2015) ne portent que sur les données des établissements de rattachement des chercheurs à l'origine de la publication, soit Harvard et le MIT. Ceci est d'autant plus notable que du fait de leurs liens avec edX, ils pourraient avoir techniquement accès aux traces des dizaines d'établissements membres du consortium.

24 Pour publier sur les traces issues de cours hébergés sur Coursera ou Unow (Cisel *et al.*, 2015 ; Cisel, 2018), nous avions dû établir des contacts personnels avec chacun des enseignants à l'origine du cours, et effectuer des retours personnalisés sur les utilisations des cours. Il est impossible de lancer de telles démarches à grande échelle. En ce qui concerne les données d'inscription de FUN MOOC (Cisel, 2019 ; Wintermute, Cisel, et Lindner, 2021), nous avions certes pu travailler à l'échelle de la plateforme, mais nous avions mobilisé les seules données d'inscription, sans pouvoir entrer à l'échelle des traces collectées pour chaque cours[1]. Nous avions en revanche l'interdiction de communiquer sur les performances (en termes de nombre d'inscrits et de certifiés) de tel ou tel cours pris individuellement.

25 Au-delà du simple cas des MOOC, il faut noter que les administrations, qu'elles relèvent de ministères ou d'institutions d'enseignement supérieur, sont par nature rétives à donner à des personnes extérieures les moyens d'évaluer avec précision le succès des initiatives qu'elles lancent. À cet égard, le numérique éducatif ne fait pas exception. Bien évidemment, diverses formes d'évaluation sont généralement mises en œuvre dans les

projets qui brassent d'importantes sommes d'argent public, et la recherche intervient parfois dans le processus. Ainsi, les Universités numériques thématiques, bras armé de l'État depuis des décennies dès qu'il s'agit de financer la conception de ressources pédagogiques, ont fait l'objet de travaux de la part d'Anne Boyer (2011) plusieurs années après leur lancement. Néanmoins, minoritaires sont les projets ambitieux qui vont jusqu'à la quantification des utilisations par les traces au cours de ces évaluations. Cela ne signifie pas que les chiffres ne sont pas connus, mais à notre connaissance ils ne sont que rarement rendus publics ou du moins pas communiqués à des chercheurs. La publication de chiffres décevants par rapport aux investissements consentis risquerait d'entraver la poursuite ou le développement des dispositifs, ou de jeter l'opprobre sur leurs responsables.

26 Face à la pénurie de jeux de données à grande échelle, des projets de mutualisation de traces pourraient faire bouger les lignes. Ceux-ci ont fleuri depuis quelques années. L'Université de Pittsburgh avait été pionnière en la matière en lançant le PSLC Datashop (Cherigny *et al.*, 2020), et des projets analogues ont vu le jour à l'échelle nationale, avec Datahub éducation, ou à l'échelle européenne avec GAIAX. De manière générale, la mutualisation des traces et des analyses via un LRS (*Learning Record Store*) commun et ouvrir le champ à ce type d'analyse. Néanmoins, il est encore rare que chercheurs et institutions versent des jeux de données conséquents sur de telles infrastructures mutualisées, probablement du fait de la sensibilité supposée des traces.

Obstacles techniques : questions de volume et de qualité

27 Les problèmes techniques se font plus saillants lorsque les projets gagnent en ambition et que les quantités de traces sont telles, qu'elles relèvent de données massives, ou *Big Data* en anglais. Des auteurs comme Nikolovska *et al.* (2018) ont contribué à la réflexion sur le dépassement de tels obstacles, notamment pour Moodle. Mais il reste nécessaire de trouver des infrastructures pour stocker dans un même lieu des Teraoctets de données de clics.

28 En sus d'infrastructures susceptibles d'héberger les données, les chercheurs ont besoin d'ordinateurs assez puissants pour en réaliser le traitement, si tant est que les analyses réalisées nécessitent une certaine puissance de calcul et que les jeux de données sont volumineux. Pour des analyses simples, un banal ordinateur portable permet de suivre le comportement de plusieurs millions de personnes. Mais si l'on cherche à sophistiquer l'analyse, il faut parfois s'adjoindre les services d'un ingénieur dédié et payer des serveurs, problématiques classiques au demeurant dans le domaine de l'informatique, aussi ne nous attarderons-nous pas.

29 La qualité des données pêche également parfois. Nous en avions fait les frais au cours de nos recherches doctorales (Cisel, 2016), devant nous concentrer sur les seules traces de Coursera alors que nous disposions de celles d'autres plateformes de MOOC. Néanmoins, un problème de stockage sur les serveurs avait conduit à la perte d'environ un log sur deux, rendant le jeu de données inexploitable malgré son potentiel, à plus forte raison si l'on veut étudier finement le lien entre la structure des cours, et le comportement des apprenants. Le fait que l'on découvre des problèmes de qualité seulement après avoir déployé des efforts pour régler le problème du volume de données traitées peut être décourageant.

Obstacles légaux et considérations éthiques

30 Depuis l'entrée en vigueur du Règlement général de la protection des données (RGPD) à l'échelle européenne, les réglementations sur le stockage et l'exploitation des données se sont renforcées. Les traces d'interaction ne font pas exception. De grands groupes privés américains comme WhatsApp ont été condamnés suite au non-respect du RGPD. Ces événements ne peuvent que sonner comme un appel à la prudence pour les administrations susceptibles de fournir aux chercheurs de vastes jeux de données. En premier lieu, il faut qu'au moment des formalités liées au RGPD, les responsables aient déclaré que les données peuvent être utilisées à des fins de recherche d'une part et que les utilisateurs concernés par les analyses aient donné leur consentement d'autre part. Par ailleurs, il faut s'assurer que le stockage des données sur les différents serveurs dont ils disposent est systématiquement en conformité avec la loi, de simples déclarations ne suffisent pas. Un manquement qui serait passé inaperçu en temps normal peut devenir visible lors de la collaboration avec un acteur extérieur. Ce peut être perçu comme un risque, à plus forte raison à un chercheur qui communique ses résultats librement sur Internet. En effet, une publication scientifique est susceptible d'attirer l'attention sur d'éventuels dysfonctionnements.

31 Par ailleurs, même si les données sont exploitables tant sur le plan légal que technique, une contrainte supplémentaire s'ajoute souvent à la toute fin du processus de recherche, des considérations éthiques sont à prendre en compte. Comme le soulignent Cherigny *et al.* (2020, p. 49), « même si l'objectif des Learning Analytics, s'inscrivant dans un climat naturel de bienveillance à l'égard de l'apprenant, est louable en soi, ses conditions d'exercice peuvent poser question au regard du respect de la vie privée ». L'un des aspects les plus problématiques des traces tient selon les auteurs au fait que l'on ne sait généralement pas, en amont, distinguer les données qui seront les plus probantes et celles qui se révéleront anecdotiques. Ceci est particulièrement vrai pour les recherches exploratoires, qui ne sont pas fondées sur des hypothèses a priori. Il devient dès lors difficile de réaliser une collecte parcimonieuse et minimale de données. Par ailleurs, comme le rappellent toujours les auteurs, il est « possible de savoir, pour chaque apprenant, s'il a réellement suivi le cours, à quel moment il a abandonné, s'il a visionné plusieurs fois la même partie, etc. ». On peut inférer quant à l'état d'attention ou aux états émotionnels des utilisateurs à partir des interactions avec les environnements numériques, ce qui doit amener le chercheur à faire preuve de vigilance, à plus forte raison s'il travaille sur les données de personnes mineures.

32 On comprend aisément dans ces conditions la nécessité de mettre en place d'une série de garde-fous, et les revues scientifiques exigent de plus en plus fréquemment que les protocoles de recherche impliquant des données issues de comportements humains aient été validés en amont de toute publication par un comité d'éthique – *Institutional Review Board* (IRB) en anglais. En France, les bailleurs de fonds de la recherche exigent par ailleurs de plus en plus souvent un Plan de gestion des données[2]. Ainsi, des analyses peuvent être légales du point de vue du RGPD sans être jugées pour autant éthiques par un comité scientifique. Alors que les comités d'éthique représentent la norme du point de vue des recherches médicales, les travaux en éducation sont soumis de manière croissante à cette pratique, même après anonymisation des données. Notre expérience personnelle suggère que les revues francophones en éducation sont à ce jour moins exigeantes sur cette thématique. Néanmoins, un alignement sur les pratiques des revues anglo-saxonnes est à prévoir dans les années à venir. Si elle complique la tâche aux chercheurs intéressés par la mobilisation de traces d'interaction dans leurs travaux, cette évolution de la prise en compte des enjeux éthiques par les différents acteurs de la recherche est louable – elle était attendue depuis plusieurs années.

Conclusion

33 La question de la mutualisation des données et de la mise en place de contrôles qualité n'est pas récente, et la France a mis en place des initiatives en ce sens qui dépassent largement le cadre de l'éducation. Ainsi, le CNRS a financé le déploiement d'OPIDOR. L'objectif de cette initiative est transparent à la lecture des initiales de ce sigle : « Optimiser le Partage et l'Interopérabilité des Données de la Recherche. Il existe même des initiatives spécifiques au domaine de l'éducation. Ainsi, l'initiative *Data Space Éducation* Skills (DASES[3]) est financé dans le cadre du PEPR « Éducation et Numérique »[4] co-piloté par l'INRIA. Mais l'État devrait voir au-delà de la seule question de la mise en place d'infrastructures pour la centralisation de données de recherche. Grâce à son rôle central dans l'éducation et l'enseignement supérieur, il constitue potentiellement l'un des pourvoyeurs les plus importants de jeux de données.

34 Du fait de sa longue tradition centralisatrice, le service public français dispose d'atouts de taille s'il souhaite contribuer à élever les ambitions de la recherche hexagonale. Parmi les collaborations fructueuses dont nous avons eu connaissance, citons la recherche doctorale de Léonard Moulin sur les frais d'inscription à l'université (Moulin, 2014), et qui n'aurait pas pu voir le jour sous cette forme sans la mise à disposition de la base nationale SISE (Système d'information sur le suivi de l'étudiant). Ce propos vaut également en matière de traces d'interaction issues d'environnements numériques : les jeux de données issus de collectes à l'échelle nationale sont nombreux. Néanmoins, du fait d'un manque de communication entre la communauté scientifique et les administrations, qu'elles soient d'universités ou de ministères, cette richesse demeure largement inexploitée. Trop souvent, les chercheurs doivent déployer une énergie considérable pour produire des dispositifs de collecte de données à la portée souvent modeste, quand des jeux de données d'une grande richesse possédés par des administrations restent largement sous-exploités.

35 Dans cette contribution, nous nous sommes longtemps appesanti sur les éléments qui font obstacle aux recherches. Il existe des solutions concrètes pour chacun de ces problèmes et les chercheurs ne les ignorent pas. Le problème n'est pas tant la méconnaissance des solutions, mais leur caractère chronophage. Mises bout à bout, les différentes étapes à franchir grignotent le temps dédié stricto sensu à l'analyse, de sorte qu'une fois tous les obstacles franchis, la recherche est réduite à sa portion la plus congrue. Cette considération nous amène à souligner l'importance qu'il y a à soulager autant que faire se peut la recherche de ces tâches, afin qu'elle puisse se concentrer sur leur cœur de métier. Reste à établir ce qui peut entrer dans le périmètre d'action des administrations qui fournissent les jeux de données. Néanmoins, pour entamer de telles démarches au profit de la recherche, encore faut-il que celles-ci voient l'intérêt d'un tel investissement. À cet égard, une meilleure compréhension respective des attentes des différents acteurs serait nécessaire pour aller de l'avant.

36 En effet, si les chercheurs peuvent s'attaquer à des problématiques pointues, ils doivent aussi produire des résultats opératoires dans le cadre du pilotage des politiques publiques. Les ministères et administrations sont friands de chiffres. Le célèbre sociologue Alain Desrosières a écrit une série de livres sur le sujet qui font encore référence, au-delà de l'Hexagone. Des ouvrages comme « Prouver et gouverner » (Desrosières, 2014) et « La politique des grands nombres » (Desrosières, 1993) analysent en profondeur le rapport entre pouvoir et statistiques et permettront aux chercheurs qui le souhaitent de mieux appréhender la nature des liens qu'ils peuvent tisser avec les décideurs.

37 Inversement, les administrations doivent apprécier le rôle qu'elles pourraient jouer dans le succès de la recherche française, au-delà de leurs intérêts à court terme. Quel que soit le domaine concerné, le fait que la France puisse tenir sur la scène internationale son rang sur le plan scientifique devrait être l'affaire de tous. Nous nous sommes exprimé ici sur la question des traces d'interaction, mais l'on pourrait sans doute tenir des discours analogues dans d'autres domaines. Dans les administrations, des changements sont en cours à différents niveaux de prise de décision, et à l'aune de nos expériences personnelles,

une prise de conscience semble avoir eu lieu quant à la richesse des données disponibles. Reste à traduire le plus rapidement possible cette prise de conscience en publications ambitieuses. Au regard de la multiplicité des obstacles à franchir, le chemin apparaît encore semé d'embûches.

Bibliographie

Albo, L., Hernández-Leo, D. et Oliver, M. (2016). Are higher education students registering and participating in MOOCs ? The case of MiríadaX. Dans *Proceedings of the European MOOCs Stakeholders Summit (EMOOCS 2016)* (p. 197-211), Norderstedt, Allemagne.

Bof, A. M. (2004). Distance Learning for Teacher Training in Brazil. *The International Review of Research in Open and Distance Learning, 5*(1), 1-14.

Bonu, B. et Charnet, C. (2007). Les espaces numériques de travail favorisent-ils le changement de la formation à distance dans l'enseignement supérieur français ? *ISDM, International Journal of Info et Com Sciences for Decision Making*, *29*.

Boyer, A. (2011). Les universités numériques thématiques : Bilan. *STICEF*, *18*(1), 39-52.
DOI : 10.3406/stice.2011.1014

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., et Seaton, D. T. (2013). Studying learning in the worldwide classroom research into edX's first MOOC. *Research et Practice in Assessment*, *2013*(8), 13-25.

Cherigny, F., El Kechai, H., Iksal, S., Lefevre, M., Labarthe, H., et Luengo, V. (2020). *L'analytique des apprentissages avec le numériqueGroupes thématiques de la Direction du numérique pour l'Éducation (DNE -TN2)* [Rapport de recherche]. Direction du numérique pour l'éducation. https://hal.archives-ouvertes.fr/hal-02912386

Cisel, M., Mano, M., Bachelet, R., Silberzahn, P. et Bruillard, E. (2015). A Tale of Two MOOCs : Analyzing Long-Term Course Dynamic. Dans *Proceedings of EMOOCs conference 2015* (p 191-199), Mons, Belgique.

Cisel, M. (2016). Utilisation des MOOC, éléments de typologie [thèse de doctorat, ENS Paris Saclay, Paris, France]. https://tel.archives-ouvertes.fr/tel-01444125/document

Cisel, M. (2018). Une analyse de l'utilisation des vidéos pédagogiques des MOOC par les non-certifiés. *STICEF, 2018*(24).
DOI : 10.3406/stice.2017.1744

Cisel, M. (2019). Le choix d'un MOOC, un processus influencé par l'organisation des cours en catalogues ? *STICEF, 2019*(25). https://doi.org/10.23709/sticef.25.1.8
DOI : 10.23709/sticef.25.1.8

Dede, C., Jass Ketelhut, D., Whitehouse, P., Breit, L. et McCloskey, E. M. (2009). A research agenda for online teacher professional development. *Journal of teacher education*, *60*(1), 8-19.
DOI : 10.1177/0022487108327554

Desrosières, A. (1993) *La politique des grands nombres*. La Découverte
DOI : 10.3917/dec.desro.2010.01

Desrosières, A (2014). *Prouver et gouverner*. La Découverte
DOI : 10.3917/dec.desro.2014.01

Ho, A. D., Chuang, I., Reich, J., Coleman, C. A., Whitehill, J., Northcutt, C. G. et Petersen, R. (2015). *HarvardX and MITx : Two Years of Open Online Courses Fall 2012-Summer 2014* (SSRN Scholarly Paper No. ID 2586847). Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm ?abstract_id =2586847

Hodges, C., Lowenthal, P. et Grant, M. (2016, Mars). Teacher professional development in the digital age : Design considerations for MOOCs for teachers. Dans les Actes de la *Society for information technology et teacher education international conference 2016* (pp. 2075-2081), San Antonio, Texas, USA.

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., … et Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, *117*(26), 14900-14905.
DOI : 10.1073/pnas.1921417117

Latchem, C. R. et Robinson, B. (dir.). (2003). *Teacher education through open and distance learning* (Vol. 3). Psychology Press.

Liyanagunawardena, T. R., Lundqvist, K. Ø. et Williams, S. A. (2015). Who are with us : MOOC learners on a FutureLearn course. *British Journal of Educational Technology*, *46*(3), 557-569.
DOI : 10.1111/bjet.12261

Lockee, B. B. (2021). Shifting digital, shifting context :(re) considering teacher professional development for online and blended learning in the COVID-19 era. *Educational Technology Research and Development*, *69*(1), 17-20.

Moulin, L. (2014). Frais d'inscription dans l'enseignement supérieur : Enjeux, limites et perspectives. *Revue de la régulation. Capitalisme, institutions, pouvoirs*, *2014*(16). https://journals.openedition.org/regulation/10855
DOI : 10.4000/regulation.10855

Nikolovska, A., Velinov, A., Spasov, S. et Zdravev, Z. (2018, Septembre). Framework for Big Data Analytics of Moodle Data Using Hadoop in the Cloud. Dans *Actes de la conférence International Scientific Conference Computer Science 2018*, (p. 3-8), Kavala, Grèce.

Perraton, H. (2002). *Distance Education for Teacher Training*. Routledge.

Pogent, F., Albero, B. et Guérin, J. (2019). Transformations professionnelles et personnelles en situation de formation hybride. Le cas d'une professeure des écoles aux prises avec la plateforme M@gistère. *Distances et médiations des savoirs, 2019*(26). 10.2139/ssrn.3533833
DOI : 10.2139/ssrn.3533833

Pogent, F. (2020) *Construction de l'expérience et formation hybride : transformations de l'activité de professeur.e.s des écoles instrumentée par la plateforme M@gistère* [thèse de doctorat, Université de Brest, France]. https://theses.fr/2020BRES0032.
DOI : 10.4000/ree.9578

Smith, J. A. et Sivo, S. A. (2012). Predicting continued use of online teacher professional development and the influence of social presence and sociability. *British Journal of Educational Technology*, *43*(6), 871-882.
DOI : 10.1111/j.1467-8535.2011.01223.x

Vrillon, E. (2019). Une nouvelle évaluation de la réussite dans les MOOC à partir de registres d'usages individuels. *Questions Vives. Recherches en éducation*, *2019*(31). https://journals.openedition.org/questionsvives/3933
DOI : 10.4000/questionsvives.3933

Whitaker, S., Kinzie, M., Kraft-Sayre, M. E., Mashburn, A. et Pianta, R. C. (2007). Use and evaluation of web-based professional development services across participant levels of support. *Early childhood education journal*, *34*(6), 379-386.
DOI : 10.1007/s10643-006-0142-7

Wintermute, E. H., Cisel, M., et Lindner, A. B. (2021). A survival model for course-course interactions in a Massive Open Online Course platform. *PLOS ONE*, *16*(1), e0245718. https://doi.org/10.1371/journal.pone.0245718
DOI : 10.1371/journal.pone.0245718

Zdravev, Z., Velinov, A., Spasov, S., et Krstev, A. (2018). *Analytics and Report Plugins in Moodle*. 163-168. http://www.conf.cceng.eu/

Zhang, Y., Ghandour, A., et Shestak, V. (2020). Using Learning Analytics to Predict Students Performance in Moodle LMS. *International Journal of Emerging Technologies in Learning*, *15*(20), 102-115.

Notes

1 Les données anonymisées sont d'ailleurs disponibles à cette adresse : https://zenodo.org/record/3969240#.YhPa3-jMLIU

2 Modèle de Plan de Gestion de Données fourni par l'ANR https://anr.fr/fileadmin/documents/2019/ANR-modele-PGD.pdf

3 https://prometheus-x.org/

4 https://www.inria.fr/fr/programme-education-numerique-muriel-brunet

Pour citer cet article

Référence électronique

Matthieu Cisel, « De la collecte des traces d'interaction à la publication scientifique », *Distances et médiations des savoirs* [En ligne], 37 | 2022, mis en ligne le 08 mars 2022, consulté le 28 novembre 2023. URL : http://journals.openedition.org/dms/6953 ; DOI : https://doi.org/10.4000/dms.6953

Auteur

Matthieu Cisel
Institut des humanités numériques, CY Cergy Paris Université,
matthieu.cisel@cyu.fr

Articles du même auteur

Interactions entre utilisateurs de MOOC : appréhender la partie immergée de l'iceberg [Texte intégral]
Interactions between MOOC users: comprehending the hidden part of the iceberg
Paru dans *Distances et médiations des savoirs*, 20 | 2017

MOOC : les conditions de la réussite [Texte intégral]
Paru dans *Distances et médiations des savoirs*, 8 | 2014

Droits d'auteur