L’analyse des traces dans un environnement OpenEdX : validation des profils d’utilisateurs et d’un modèle prédictif

Normand Roy*, Bruno Poellhuber**, Ibtihel Bouchoucha***
*Normand.Roy@umontreal.ca, Université de Montréal
**Bruno.Poellhuber@umontreal.ca, Université de Montréal
***Ibtihel.bouchoucha@umontreal.ca, Université de Montréal

Résumé

La persévérance dans les cours en ligne reste une préoccupation des différents établissements postsecondaires (Dussarp, 2015). Le cas des MOOC représente une situation particulière des cours en ligne, avec un taux d’abandon encore plus élevé (Jordan, 2015). L’exploitation des traces informatiques a donné lieu à des approches innovantes en éducation, comme le *Academic Analytics*, le *Educational Data Mining* et le *Learning Analytics*. Le *Learning Analytic* permet une nouvelle approche pour identifier les étudiants « à risque », c’est-à-dire ceux dont le pronostic de réussite semble douteux, et pour introduire des interventions visant à améliorer les pratiques d’encadrement, d’enseignement ou d’apprentissage. Dans le présent projet, nous examinerons les spécificités nécessaires pour la transformation des données sur OpenEdX, afin d’en arriver à un modèle prédictif de la persévérance. Nous tenterons de mettre en place des modèles d’analyse, en tentant de reproduire le pouvoir prédictif d’un modèle équivalent à celui réalisé dans les analyses préliminaires. À la lumière de ces résultats, des hypothèses seront présentées sur les moyens à mettre en place pour favoriser l’engagement et la persévérance des apprenants.

Abstract

Perseverance in online courses remains a concern for post-secondary institutions (Dussarp, 2015). The case of MOOCs represents a particular case of online courses, with an even higher dropout rate (Jordan, 2015). The exploitation of computer traces has resulted in innovative approach, such as Academic Analytics, Educational Data Mining, and Learning Analytics. Learning Analytic allows a new approach to identify "at risk" students, that is, those with poor prognosis for success, and to introduce interventions to improve coaching, teaching or learning practices. In this project, we will examine the specificities required for data transformation on OpenEdX, in order to arrive at a predictive model of perseverance. We will propose a predictive model, trying to reproduce the predictive power of a model equivalent to that realized in the multiple preliminary analyzes. In light of these results, hypotheses will be presented on practical implication to encourage learners' commitment and perseverance.

Mots-clés

Analyse des traces ; MOOC ; persévérance ; engagement ; *learning analytic ;* modèle prédictif

Keywords

Trace analysis ; MOOC ; perserverance ; commitment; learning analytic; predictive model

Contexte

Le numérique exerce une influence profonde sur le monde dans lequel nous vivons, et engendre une transformation d'un grand nombre de champs d'activités, notamment celui des sciences sociales et de la formation. Il amène aussi une transformation du regard que la recherche porte sur les contextes d'apprentissage, les paradigmes mobilisés, ainsi que sur les outils et méthodes de recherche. Chacun des comportements des apprenants dans ces différents environnements numériques laisse une trace informatique. Typiquement, le système de log (ou de journaux de traces informatique) va capter chaque action réalisée par chaque utilisateur selon un schéma de type « identifiant, date-heure-minute, action, URL ». L'ensemble de ces traces génère une quantité massive de données à analyser. C'est plus que jamais le cas avec la place prépondérante qu'ont prise les applications d'apprentissage qui roulent dans un environnement technologique client-serveur comme le Web. Cette explosion de la quantité (et parfois de la qualité) des données accessibles potentiellement pertinentes (Diebold, 2000) correspond au domaine du *Big Data*, auquel on commence à s'intéresser en éducation.

L'exploitation des traces informatiques a donné lieu à des approches se rapprochant de l'intelligence d'affaires (*Business Analytics*) en éducation, comme le *Academic Analytics*, le *Educational Data Mining* et le *Learning Analytics (LA)*. Alors que la première approche (*Academic Analytics)* est centrée sur la perspective institutionnelle, le *Educational Data Mining* (EDM) mise plutôt sur les données générées dans des contextes éducationnels et d'apprentissage. La grande majorité des approches déployées actuellement dans le domaine du LA demeurent limitées aux toutes premières étapes (l'extraction et la collecte des traces et données). Les initiatives qui vont jusqu'au déploiement de tableaux de bord permettant des suivis individuels et interventions sont rares.

L'engagement en FAD

Pour de nombreux chercheurs (Fredricks et al., 2004), l'engagement scolaire est constitué de trois composantes : comportementale, cognitive et affective. L'engagement comportemental est lié aux manifestations observables de l'engagement et a trait à la quantité d'efforts. Quant à la qualité de ces efforts, elle fait plutôt référence à l'engagement cognitif (Linnenbrink & Pintrich, 2003). Notre hypothèse sous-jacente à l'analyse des traces est celle qu'elles reflètent un certain continuum d'engagement comportemental. Les traces pourraient peut-être aussi témoigner du niveau d'engagement cognitif, qui correspond à l'effort mental et à l'usage de stratégies, mais de formater les traces en fonction de ce concept est plus complexe. Ainsi, il est possible d'interpréter les traces (usage des ressources et variété des usages) en fonction d'un continuum d'engagement. Selon Molinari *et al.*(2017), « la définition de l'engagement comportemental se fonde sur l'idée de participation et d'indicateurs observables de cette participation ». Dans le contexte de la formation à distance, cela peut se traduire par deux types de traces : les comportements de l'utilisateur à la maison et les comportements sur la plateforme.

Persévérance et abandon dans les MOOC

Le recours de plus en plus grand à la FAD et aux MOOC présente des avantages très importants en termes de démocratisation du savoir, mais les taux de réussite des MOOC, sont alarmants pour plusieurs, variant entre 5 et 10 % (Jordan, 2015). L'autre grand problème des FOAD, qui est probablement aussi un problème des MOOC, est la plus grande variabilité de la qualité (Bernard et al., 2004).

Bourdages et Delmotte (2001) proposent de classer les facteurs liés à l'abandon et à la persévérance en quatre catégories : les variables démographiques (genre, âge, etc.), environnementales (provenant du cadre de vie personnel de l'étudiant), individuelles (émotions, cognitions, motivation) et institutionnelles (design des cours, encadrement, interactions). Ces dernières sont celles sur lesquelles l'établissement peut réellement exercer un contrôle. Bien qu'elle soit encore jeune, la recherche sur les MOOC pointe vers un certain nombre de facteurs liés à l'abandon ou à la persévérance, qui sont essentiellement en lien avec les variables institutionnelles et individuelles; la durée des MOOC, le type d'évaluation ou de rétroaction, le manque de temps disponible (Belanger et Thornton, 2013), les motivations diverses des apprenants (Mackness, Mak et Williams, 2010); et d'autres plus, classiques en FAD, comme le sentiment d'isolement (Erichsen et Bolliger, 2011), la motivation des apprenants et l'absence d'interactions (Waard, Peters, Shelley 2010; Levy & Schrire; 2012).

Ces différents éléments se rattachent à des aspects qui sont bien documentés par les plateformes de cours en ligne. Ainsi, le *Learning Analytic* permet une nouvelle approche pour identifier les étudiants « à risque », c'est-à-dire ceux dont le pronostic de réussite semble douteux, et pour introduire des interventions visant à améliorer les pratiques d'encadrement, d'enseignement ou d'apprentissage. Ces mesures visant à favoriser la réussite des apprenants incluent notamment l'introduction de tableaux de bord pour les enseignants et les apprenants.

Méthodes

Ce projet proposé s'inscrit en continuité avec une programmation de recherche sur les cours en ligne ouverts aux masses (MOOC) qui est financée par Conseil de recherches en sciences humaines (CRSH). Deux publications significatives sont préliminaires à cette proposition, en plus d'un rapport de recherche. Le premier rapport a permis de mettre en place une stratégie d'exportation des données et un continuum d'engagement à partir des traces sur la plateforme SAKAI (portail EDULIB 2012-2013). Ensuite, une première analyse des traces sur les données a été réalisée et a mené à la création des profils typologiques (Poellhuber, Roy et Bouchoucha, sous presse). Finalement, une analyse plus approfondie a permis de nuancer l'existence des profils sur la durée complète d'un cours (Van Den Bossche, 2018). Ces deux publications sont en quelque sorte une preuve de concept du modèle d'analyse des données.

L'objectif de cette proposition est de concevoir un modèle prédictif qui pourrait permettre de définir très tôt dans le parcours étudiant des interventions personnalisées, dans le but de favoriser l'engagement et la persévérance des apprenants sur la **plateforme EdX** (FUN et EDUlib).

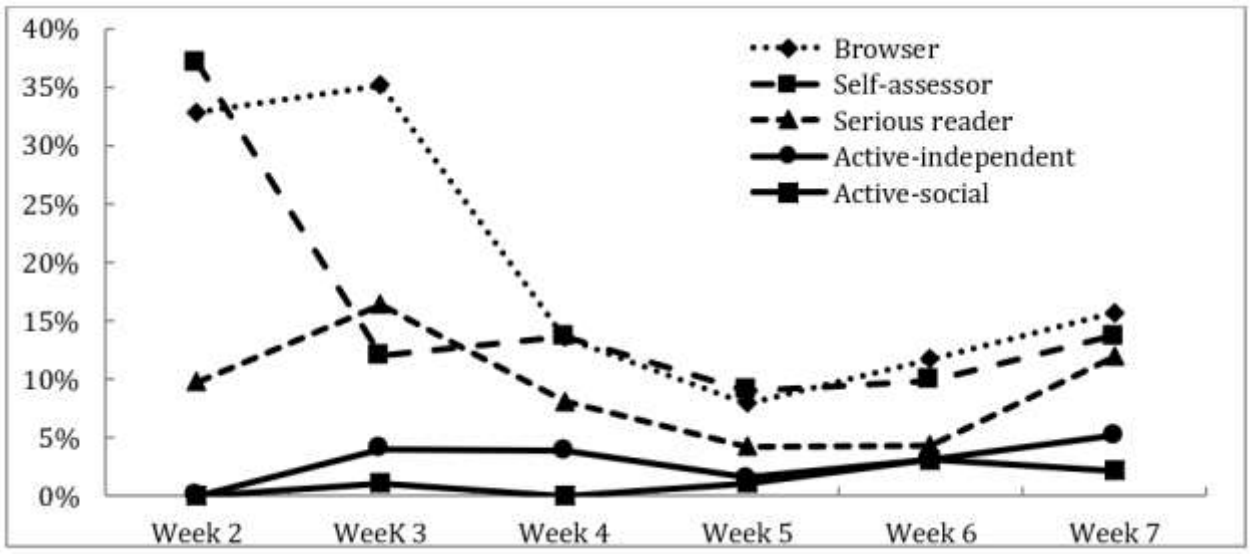
Le système de traces permet de collecter une grande quantité d'interactions avec les ressources : fréquence de connexion, téléchargement des ressources, type d'interaction dans le forum, type de ressources consultées, participation et complétion des tests, etc.

Résultats et discussion

Nous avons d'abord créé une analyse des traces avec les données disponibles sur Sakai (plateforme EDUlib de 2012-2014) (Poellhuber, Roy et Bouchoucha, 2016). L'analyse préliminaire a permis de créer cinq profils distincts : absent, butineur, évaluateur, actif indépendant et actif social. Ces profils se comportent de façon différente sur la persévérance, avec un taux de réussite variant de 26% à 93% respectivement (Poellhuber, Roy et Bouchoucha, sous presse). Ces profils ont été élaborés à partir d'un score composite, créé à partir d'une analyse de correspondances multiples. Le score d'engagement était, entre autres, constitué du visionnement des vidéos, la participation

aux tests (absente, partielle ou complète), le téléchargement des lectures, la consultation des ressources web et la participation aux forums de discussion (lire, répondre, écrire). À la lumière de cette analyse, nous constatons que le risque d'abandon pour les étudiants actifs sont pratiquement nuls en début de parcours qu'il est plutôt élevé pour qui sont présents de façon sporadique (Figure 1). À noter que cette analyse inclut uniquement les utilisateurs actifs à partir de la semaine 2, excluant ainsi ceux qui ne se présentent pas (« no-show »).

Figure 1. Profils en fonction du risque d'abandon par semaine (tiré de l'article Poellhuber, Roy et Bouchoucha, sous presse).



La deuxième analyse permet de voir qu'une large proportion des participants modifient leur comportement en cours de participation, changeant ainsi de profils. Selon Van Den Bosche (2018), « … participants tend to switch from profile as weeks go by, undermining the validity of profiles assigned thanks to data collected at the beginning MOOC. This statement is further reinforced by the fact that individuals do not remain in adjacent profiles, as we could observe sharply oscillating behaviors during the period of analysis. ».

Dans le cadre d'analyses exploratoires, un modèle prédictif de la persistance a été élaboré selon un modèle de régression logistique et a permis de classer avec 89 % d'exactitude les apprenants qui ont persévéré dans le MOOC à partir des traces d'activité et de données autorapportées de la semaine 2. Le profil des participants était la variable la plus importante dans ce modèle prédictif (Poellhuber, Roy et Bouchoucha, sous presse).

Toutefois, l'ensemble des analyses porte sur un échantillon d'un seul cours, sur la plateforme SAKAI. Ce faisant, même si l'on peut considérer ces analyses comme une preuve de concept, il est nécessaire de mettre en application sur des données plus pertinentes (différents cours) pour le contexte européen (FUN) et québécois (EDUlib) des MOOCs. La plateforme OpenEdX est la plus populaire au monde pour les MOOC.

OpenEdx propose une architecture différente pour les traces proposées : Inscription, Navigation, interaction avec les vidéos, interaction avec le manuel, interaction avec les problèmes, examens, favoris, notes, interaction avec les librairies, forum de discussion, tests à réponses ouvertes, apprentissage par les pairs, sondage, complétion des contenus, contenus externes, activités en équipe, certification. Cela implique donc une réflexion autour de la correspondance entre les traces

disponibles et les agrégations nécessaires pour arriver à établir des profils similaires. L'idée est de pouvoir non seulement proposer un continuum d'engagement (ex. : lire le forum est moins engageant que d'écrire sur le forum), mais également de trouver une façon de combiner les différentes activités pour créer des profils d'utilisateurs. Nos résultats préliminaires permettent d'affirmer que ces profils sont les meilleurs prédicteurs de la persévérance. D'ailleurs, nous sommes loin d'être les seuls à parler de ces profils (Kahan, Soffer et Nachmias, 2017; Khalil et Ebner, 2017; Kizilcec, Piech et Schneider, 2013; Milligan, 2012)

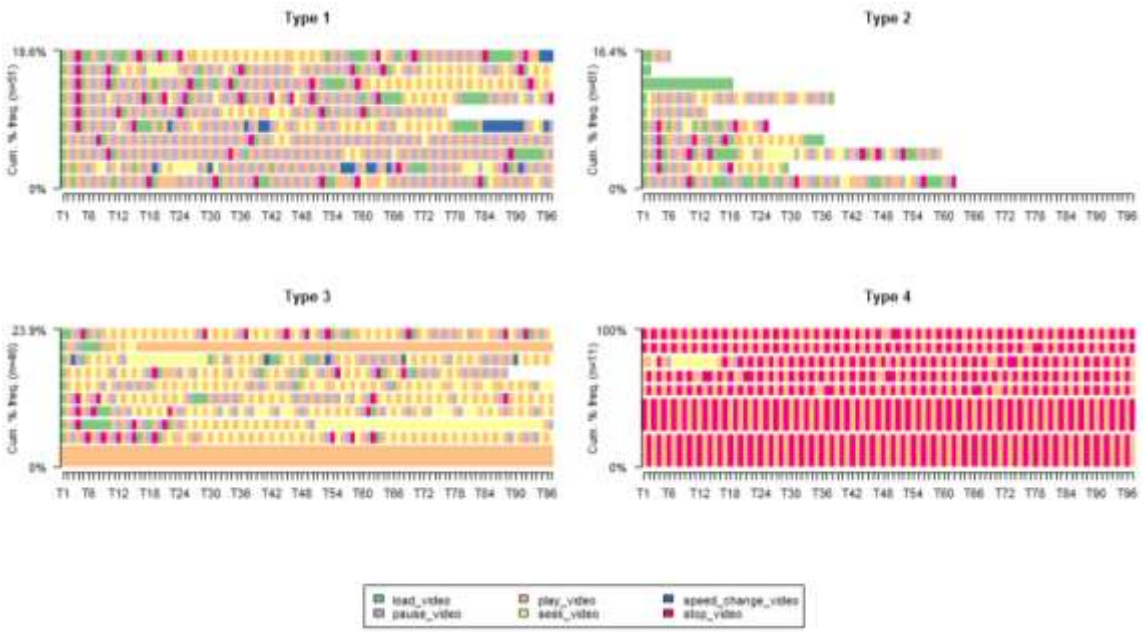
De nombreuses études proposent d'utiliser les traces OpenEdX afin de mieux comprendre l'engagement et les comportements des utilisateurs (Huang, Huang, Lu, Tseng et Yang, 2016 Maldonado-Mahauad et al., 2017; Sunar, White, Abdullah et David, 2017; Veeramachaneni et coll., 2014). Ces études ont regardé une variété de variables. Sunar et ses collaborateurs (2017) proposent d'ailleurs un continuum basé sur une perspective linéaire (un seul axe) des activités. Bien que l'analyse simplifie l'interprétation, cela ne correspond pas nécessaire à une réalité pratique : peut-on mettre au même niveau la participation au forum à celui de la réalisation des activités de lecture?

D'autres auteurs ont plutôt préconisé l'utilisation de données autorapportées des comportements des utilisateurs : proportion des activités complétées, proportion des documents consultés, proportion du cours complété, etc. (Hone et El Said, 2016). Les données autorapportées permettent de nuancer certaines traces numériques (ex. : un clic sur un document ne nous dit pas le temps passé sur ce document).

Veeramachaneni et coll. (2014) ont créé une architecture des données afin de standardiser les variables entre les différents modèles d'analyse. Le modèle organise l'interaction avec les ressources sur OpenEdX sous quatre catégories : soumettre, observer, collaborer et rétroaction. L'implantation préliminaire de ce modèle sur nos données n'a pas permis des résultats concluants. En premier lieu, il faut comprendre que la mise en place de ce modèle exige une grande appropriation de la structure de programmation. L'implantation de MOOCdb implique d'intégrer du langage Python et du langage MySql. De plus, malgré qu'il s'agisse dans les deux cas de OpenEdX, l'installation d'une itération implique l'adaptation sur certaines variables. Finalement, le modèle prédictif s'est avéré moins efficace que notre premier modèle de prédiction, avec un taux se situant entre 70 et 73% selon les MOOC.

En parallèle de cette implantation, nous avons également implanté un lecteur vidéo permettant de documenter les actions des apprenants (charger la vidéo, lecture, pause, etc.). Cette analyse est absente du modèle proposé par Veeramachaneni et coll. (2004). La figure 2 permet de constater différents profils d'utilisateur de la vidéo, principalement basé sur la fréquence des comportements. Des analyses plus approfondies ont également mené à réfléchir à des séquences d'actions sur les vidéos. Nous croyons que les comportements des vidéos, sources primaires des contenus dans les MOOC étudiés, peuvent également améliorer le modèle de prédiction de la persévérance.

Figure 2. Analyse exploratoire des traces de la vidéo



Conclusion

À la lumière de tous ces écrits et des différents modèles d'analyse présentés, dans la présentation actuelle, nous mettrons en lumière les spécificités nécessaires pour la transformation des données sur OpenEdX et une typologie qui propose une approche pratique des modèles théoriques afin de pouvoir intervenir lors des cours en ligne. Nous tenterons de mettre en place des modèles d'analyse, en tentant de reproduire le pouvoir prédictif d'un modèle équivalent à celui réalisé dans les analyses préliminaires. À la lumière de ces résultats, des hypothèses seront présentées sur les moyens à mettre en place pour favoriser l'engagement et la persévérance des apprenants.

Références

Belanger, Y. et Thornton, J. (2013). *Bioelectricity: A quantitative approach. Duke University's first MOOC*. Récupéré de DukeSpace : http://dukespace.lib.duke.edu/dspace

Bernard, R. M et al. (2004). How Does Distance Education Compare With Classroom Instruction? A Meta-Analysis of the Empirical Literature. *Review of Educational Research*, *74*(3), 379–439. https://doi.org/10.3102/00346543074003379

Bourdages, L., et Delmotte, C. (2001). La persistance aux études universitaires à distance. *Journal of Distance Education/Revue de l'enseignement à distance*, *16*(2), Récupéré de : http://www.ijede.ca/index.php/jde/article/view/176/353

Diebold, F. (2000). "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. (Discussion of Reichlin and Watson papers), in M. Dewatripont, L.P. Hansen and S.Turnovsky (Eds.), *Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society*. Cambridge: Cambridge University Press, 115-122.

Dussarps, C. (2015). L'abandon en formation à distance. *Distances et médiations des savoirs*. 10 | 2015. Récupéré le 13 décembre, 2018 de : http://journals.openedition.org/dms/1039

Erichsen, E. A., & Bolliger, D. U. (2011). Towards understanding international graduate student isolation in traditional and online environments. *Educational Technology Research and Development*, *59*(3), 309-326.

Fredricks, J. A., Blumenfeld, P. C. et Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59-109.

Hone, K. S., et El Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study. *Computers & Education*, *98*, 157-168.

Huang, J. C., Huang, A. Y., Lu, O. H., Tseng, H. C., et Yang, S. J. (2016). Learning Dashboard: Visualization of learning behavior in MOOCs. In *The International Workshop on Technology-Enhanced Collaborative Learning (TECL 2016) In conjunction with CRIWG/CollabTech 2016* (Vol. 1, p. 25).

Jordan, K. (2015). Massive open online course completion rates revisited: Assessment, length and attrition. *The International Review Of Research In Open And Distributed Learning*, *16*(3). doi:http://dx.doi.org/10.19173/irrodl.v16i3.2112

Kahan, T., Soffer, T., et Nachmias, R. (2017). Types of Participant Behavior in a Massive Open Online Course. *The International Review of Research in Open and Distributed Learning*, *18*(6). doi: https://doi.org/10.19173/irrodl.v18i6.3087

Khalil, M., et Ebner, M. (2017). Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories. *Journal of Computing in Higher Education*, *29*(1), 114-132. doi: https://doi.org/10.1007/s12528-016-9126-9

Kizilcec, R. F., Piech, C. et Schneider, E. (2013, April). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170-179). ACM. doi: https://doi.org/10.1145/2460296.2460330

Levy, D. et Schrire, S. (2012). Troubleshooting MOOCs: The case of a massive open online course at a college of education. In EdMedia: World Conference on Educational Media and Technology (pp. 761-766). Association for the Advancement of Computing in Education (AACE).

Mackness, J., Mak, S., & Williams, R. (2010). The ideals and reality of participating in a MOOC. In *Proceedings of the 7th international conference on networked learning 2010*. University of Lancaster.

Manyika, J., Chui,. M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Milligan, C. (2012). Change 11 SRL-MOOC study: Initial findings. Blog. Caledonian Academy. Retrieved from http://worklearn.wordpress.com/2012/12/19/change-11-srl-mooc-study-initial-findings/

Molinari *et al*. (2016). L'engagement et la persistance dans les dispositifs de formation en ligne : regards croisés », *Distances et médiations des savoirs.* 13|2016.

Poellhuber, B., Roy, N. & Bouchoucha, I. (2016). Les relations entre attentes, valeur, buts, engagement cognitif et engagement comportemental dans un MOOC. *Revue internationale des technologies en pédagogie universitaire*,*13*(2-3), 111–132.

Sunar, A. S., White, S., Abdullah, N. A., et Davis, H. C. (2017). How learners' interactions sustain engagement: a MOOC case study. *IEEE Transactions on Learning Technologies*, *10*(4), 475-487.

Van Den Bossche, O. (2018, en évaluation). Analysis of behavioral engagement and participants' profiles based on trails of activity. Mémoire de *Master*. Université libre de Bruxelles.

Veeramachaneni, K., Halawa, S., Dernoncourt, F., O'Reilly, U-M., Taylor, C. et Do, C. (2014) Moocdb: Developing standards and systems to support mooc data science. arXiv preprint arXiv:1406.2015.

Ward, M., Peters, G., & Shelley, K. (2010). Student and faculty perceptions of the quality of online learning experiences. *The International Review Of Research In Open And Distributed Learning, 11*(3), 57-77. doi:http://dx.doi.org/10.19173/irrodl.v11i3.867