GTNUM2

ÉTAT DE L'ART

2020

L'analytique des apprentissages avec le numérique

Groupes thématiques de la Direction du numérique pour l'Éducation (DNE – TN2)

MARS 2020

Directrice de publication
Marie-Caroline Missir

Coordination de projet
Jean-Michel Perron

Directeur artistique
Samuel Baluret

Responsable artistique
Isabelle Guicheteau

Conception graphique
DES SIGNES,
le studio Muchir et Desclouds

Mise en pages
Michaël Barbay

L'ANALYTIQUE DES APPRENTISSAGES AVEC LE NUMÉRIQUE

Florence Cherigny
Hassina El Kechai
Sébastien Iksal
Hugues Labarthe
Marie Lefevre
Vanda Luengo

SOMMAIRE

AVANT-PROPOS

Ce rapport a été rédigé par le Groupe thématique numérique n° 2 (GTnum2), « Learning Analytics », à l'initiative de la Direction du numérique pour l'Éducation (DNE-MENJ). Le travail réalisé dans ce groupe de travail comporte cinq axes conçus de façon à pouvoir être conduits de façon indépendante, tout en étant complémentaires les uns des autres :

> axe 1. Méthodes et outils de Learning Analytics ;
> axe 2. Usages et Learning Analytics ;
> axe 3. Entreprises EdTech et Learning Analytics ;
> axe 4. Aspects juridiques, éthiques, déontologiques et Learning Analytics ;
> axe 5. Terminologie des Learning Analytics.

Pour la réalisation de ce travail de prospective, le GTnum2 a organisé son travail selon deux approches distinctes : l'une, basée sur l'état de l'art ; et la seconde, basée sur des enquêtes.

Le présent document présente l'état de l'art sur les Learning Analytics.
Il est composé de quatre parties :

> une première partie présente les communautés scientifiques qui travaillent sur les Learning Analytics, ainsi que les questions de recherche qui sont posées ;
> une seconde partie présente les méthodes et outils des Learning Analytics ;
> une troisième partie traite de l'utilisation et des usages des Learning Analytics dans les contextes scolaires ;
> une dernière partie traite des questions éthiques liées à l'utilisation des Learning Analytics.

Évaluer les capacités d'abstraction des apprenants, détecter leurs pertes d'attention, adopter une pédagogie différenciée, dresser un bilan personnalisé actualisé au fil de l'apprentissage : voici autant de tâches qui reposent sur la capacité d'un enseignant à observer, analyser et réinvestir les traces comportementales et cognitives d'un apprentissage. Bien malgré lui, ce professionnel ne capte qu'une petite partie de ces données, ce qui limite ses possibilités d'interprétation d'un geste, d'un exercice inabouti, d'une erreur de réappropriation. Comme dans d'autres domaines, l'observation humaine non instrumentée est limitée et fragile.

Avec le glissement des activités d'apprentissage vers des dispositifs numériques, ces traces changent de statut : en temps réel ou en différé, à distance ou en présentiel, elles n'ont jamais informé, de façon aussi fine et massive, l'écart entre le dire et le faire. Tableaux numériques, ordinateurs, tablettes, liseuses, smartphones sont susceptibles de capter toujours plus de données sur ce qui est en train de se jouer, sur le plan verbal et comportemental, dans un processus d'apprentissage. Produire, collecter, analyser et réinvestir ces traces numériques permettraient d'aider les acteurs de la communauté éducative – apprenants, parents, personnels d'éducation, enseignants, gestionnaires et administrateurs – dans les enjeux auxquels ils doivent respectivement faire face, dans la perspective du Socle commun de connaissances, de compétences et de culture, qui requiert désormais d'évaluer à parts égales la maîtrise de la langue, les connaissances disciplinaires et les capacités d'autonomie et d'initiative.

Révéler ce qui se joue dans un processus d'apprentissage est l'enjeu des Learning Analytics. À la croisée des sciences sociales et de l'informatique, une communauté scientifique s'emploie depuis une dizaine d'années à développer des technologies et des techniques pour mieux comprendre les ressorts de l'apprentissage, de façon à en améliorer l'accompagnement et l'environnement, notamment par des dispositifs informatiques adaptables et adaptés.

ANALYTICS, DATA SCIENCE & DATA MINING

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. Si de nombreux praticiens francophones utilisent le néologisme « Analytique » pour nommer leur discipline, il reste encore peu usité.

Les méthodes employées sont issues du champ de la science des données (*Data Science*), dont le syntagme apparaît en 1996 sous la plume de Chikio Hayashi [Hayashi, 1998]. En 2003, le *Journal of Data Science* est lancé. Ces méthodes se sont développées à compter des années 1960, à la croisée des statistiques et de l'informatique, comme en témoignent les travaux de Peter Naur [1969] ou de John W. Tukey [1977]. Popularisées en 1994 par le succès du Data Mining en marketing, elles ont toutefois plus profondément innervé le champ scientifique : ainsi, dès sa création en 1970, la revue *Computers in Biology and Medicine* devient pionnière en

Analytics [Baker, Siemens, 2014]. Certains chercheurs prétendent rassembler les sciences formelles sous l'étendard de cette nouvelle science des données [Zhu, Xiong, 2015]. De fait, le principe même du Data Mining est étranger aux théories qui concernent le comportement humain comme la linguistique, la psychologie ou la sociologie : la structure des données serait consubstantielle aux données elles-mêmes. On ne peut pourtant pas réduire les Analytics à une opposition entre statistique probabiliste et algorithmes de modélisation des données *ab nihilo*. L'enjeu des Analytics est de dépasser les simples descriptions et extrapolations, pour leur substituer la modélisation, la recommandation et la prédiction [Davenport *et al.*, 2010]. En visant une prise de décision motivée par les données elles-mêmes (*data driven decision-making*), les Analytics sont considérées pour améliorer la productivité et les résultats des organisations [Brynjolfsson *et al.*, 2011].

Glossaire. Définitions usuelles en science des données

Big Data : désigne un jeu de données massives, dont la taille excède celle de capture, de conservation, de gestion et d'analyse des outils logiciels de bases de données typiques. Par contraste, on désigne par l'expression Thick Data des données qualitatives recueillies et analysées pour trouver le sens d'un phénomène.

Data Mining : en recherche fondamentale, application d'algorithmes, issus de la statistique ou de l'intelligence artificielle, pour l'extraction de l'information (utile et inconnue) de gros volumes de données non structurées.

Business Intelligence : méthodes d'analyse et de modélisation pour anticiper des scénarios liés au fonctionnement d'une organisation.

Knowledge Discovery in Databases (KDD) : processus de préparation, de sélection, d'apprêt et d'interprétation des données par les techniques du Data Mining [Fayyad *et al.*, 1996]. En 1997, l'apparition d'une nouvelle revue, Data Mining and Knowledge Discovery, consacre l'ascendance de l'expression « Data Mining sur KDD ».

Si le syntagme « Learning Analytics » ne s'impose qu'en 2011, les données d'apprentissage sont déjà très intensivement mobilisées par les équipes pluridisciplinaires œuvrant pour la conception d'« environnements informatiques pour l'apprentissage humain » (EIAH), parmi lesquels les tuteurs intelligents. La modélisation a priori de l'apprenant, l'analyse exploratoire de données [Benzecri, 1973], la création du groupe « Intelligence artificielle et didactique » [Balacheff, 1994], puis les conférences internationales AIED (*Artificial Intelligence in Education*) et francophones (EIAO/EIAH) ont posé les fondements de problématiques et de méthodes réinvesties par les Learning Analytics. Quatre facteurs expliquent plus largement leur développement [Baker, Siemens, 2014].

> Le volume des données mis à disposition des chercheurs a très rapidement progressé : soit grâce à des archives publiques comme le *Pittsburgh Science of Learning Center DataShop* [https://pslcdatashop.web.cmu.edu/] ; soit par la multiplication des dispositifs d'apprentissage en ligne (LMS – pour *Learning Management System*, MOOCs – pour *Massive Open Online Courses*) permettant de capter les interactions des utilisateurs ; soit par croisement avec des données académiques.

> Ces données sont davantage structurées et utilisables : les travaux se succèdent en matière d'interopérabilité des données venant de plusieurs environnements d'apprentissage [Walker, 2012, Niemann *et al.*, 2013].

> Les capacités de calcul des smartphones d'aujourd'hui dépassent celles des ordinateurs d'il y a dix ans.

> Enfin, de nouveaux frameworks permettent de gérer des données à la mesure du Web ; et de nombreux outils d'analyse, adaptés de la Business Intelligence à l'éducation, permettent de mener des recherches sans être nécessairement avancé en programmation ou en sciences statistiques. Orange Data Miner [https://orange.biolab.si/], RapidMiner [https://rapidminer.com/] ou Weka [www.cs.waikato.ac.nz/ml/weka/], par exemple, incluent ce type d'algorithmes.

De quelles données parle-t-on ? De données qualitatives et de traces : les données qualitatives correspondent par exemple aux réponses à des formulaires des usagers, tandis que 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) ; la complétion du programme de formation (lectures des pages de cours, tentatives aux quiz) ; le succès (scores obtenus) ; la participation (posts et messages). D'aucuns pourraient juger que la granularité des données désormais engrangées est bien moins fine que celles mobilisées dans les expérimentations antérieures, sur les tuteurs intelligents, lorsqu'elles permettaient d'évaluer la concentration, les émotions et les capacités cognitives de l'apprenant *via* des capteurs physiologiques, l'analyse automatique des expressions faciales ou des instruments d'auto-évaluation [Calvo, D'Mello, 2011 ; D'Mello, Graesser, 2012]. En l'espace de deux-trois années, environnements, problématiques et méthodes ont changé : on est passé de données a priori destinées à évaluer certains aspects émotionnels, comportementaux ou cognitifs, à des données tout-venant mais massives, à la fois par le nombre des utilisateurs, l'intensité des usages d'un utilisateur particulier et l'entrée du monde réel dans le monde numérique (géolocalisation, biométrie). La masse croît et les techniques permettent de structurer a posteriori ce qui est pertinent.

Il faut clairement distinguer les Learning Analytics des Academic Analytics. Proposer des méthodes, visualisations, algorithmes en vue d'améliorer les résultats des apprenants, renforcer leur engagement, optimiser leurs expériences d'apprentissage : tout ceci relève des Learning Analytics. Soutenir la représentation de minorités ethniques, augmenter la productivité de l'organisation, allouer des ressources aux établissements en déficit de résultats : ces objectifs relèvent en revanche de pilotages institutionnels et de stratégies politiques (*vs* Sclater, 2014, p. 4]. Les Learning Analytics ont une visée cognitive. Appliquer les recettes de l'informatique décisionnelle (ou Business Intelligence) à la sphère éducative en vue du pilotage institutionnel relève des Academic Analytics [Campbell *et al.*, 2007 ; Long, Siemens, 2011]. La confusion reste vivace mais elle est abusive.

Tableau 1. Niveaux, objets d'analyse et acteurs en Academic et en Learning Analytics (d'après Long, Siemens, 2011, p. 4)

Academic Analytics			Learning Analytics		
Niveaux	**Analyses**	**Bénéficiaires**	**Niveaux**	**Analyses**	**Bénéficiaires**
Local	Profils d'apprenants, performances académiques, *knowledge flow*, concentration des moyens financiers	Administrateurs finances, marketing	Individuel	Compréhension de la performance personnelle en relation avec des objectifs d'apprentissage et habitudes de travail des camarades	Apprenants, personnels d'éducation et enseignants
Régional	Comparaison entre systèmes éducatifs, qualités et standards, classements	Finances, gestionnaires, administrateurs	Cohortes	Réseaux sociaux, développement conceptuel, analyse de discours, progression intelligente	Enseignants
National et international		Gouvernements, organisations, Unesco, OCDE	Équipe pédagogique	Modélisation prédictive, modèles d'échec/succès	Personnels d'éducation et enseignants

Fondées sur la collecte de données pour chaque apprenant, les Learning Analytics permettent une communication différenciée aux acteurs de l'apprentissage en variant les niveaux d'analyse, le temps écoulé et la granularité des données [Brown, 2012]. Au fil des contributions, les Learning Analytics élargissent leurs méthodes à cinq champs d'action. Quels sont-ils ? Un résumé est présenté dans le tableau suivant.

Tableau 2. Des objectifs communs à EDM et aux Learning Analytics
(d'après Baker, Siemens, 2014, p. 257-262)

	Objectifs du traitement	Méthodes	Quelques travaux de référence
Prédiction entre variables	Développer un modèle permettant de prédire une variable (dépendante) à partir de la combinaison d'autres variables (indépendantes). À partir d'un jeu de données complet mais de taille réduite, on détermine la valeur de la variable à prédire. Le modèle est statistiquement validé de façon à être appliqué à plus grande échelle.	Classification (variable prédite binaire ou catégorielle) : arbre de décision, forêt d'arbres décisionnels, règles de décision, régression séquentielle et logistique. Régresseurs (variable continue). Estimation de connaissance latente (type spécifique de classifieurs).	Travaux de portée générale : Dekker *et al.*, 2009 ; Feng *et al.*, 2009 ; Ming Ming, 2012. Les algorithmes d'estimation de connaissance sont au fondement des tuteurs intelligents (Koedinger, Corbett, 2006).
Découverte de structures	À l'inverse d'une tentative de prédiction sur une variable dépendante, cette méthode vise à établir la structure des données sans idée préconçue sur l'objet recherché, sans intérêt a priori pour aucune variable.	*Clustering*, ou partitionnement des données : trouver les données qui se regroupent naturellement entre elles, en séparant le jeu en plusieurs sous-ensembles.	Grouper des étudiants (Beal *et al.*, 2006). Différencier les actions des apprenants (Amershi, Conati, 2009).
		Analyse factorielle : utilisée dans le même but de créer des sous-ensembles homogènes. Analyse de réseaux sociaux : des modèles sont développés à partir des relations et interactions entre individus.	Quels choix de design sont généralement faits par les concepteurs des tuteurs intelligents ? (Baker, Yacef, 2009). Évaluer l'efficacité de différents groupes de projet (Kay *et al.*, 2006). Positionner des apprenants dans le réseau et mesurer leur perception de la communauté éducative (Dawson, 2008). Évaluer l'engagement d'un apprenant dans sa scolarité (Macfadyen, Dawson, 2010).
		Découverte du modèle du domaine.	Tracer l'apprentissage durant l'utilisation d'un tuteur intelligent (Cen *et al.*, 2006).

Fouille de relations	Découvrir les relations entre variables au sein d'un jeu de données avec un grand nombre de variables, soit pour trouver quelles variables sont étroitement associées avec une tierce variable d'intérêt, soit quels couples de variables sont étroitement reliés.	Fouille de règles d'association.	Trouver des patterns d'étudiants obtenant de bonnes performances pour pouvoir faire de meilleures suggestions aux étudiants qui éprouvent des difficultés (Ben-Naim, 2009).
		Fouille de corrélation : trouver des corrélations linéaires positives ou négatives entre variables.	Corrélations calculées parmi les variables sur la conception des leçons d'un système de tuteur intelligent et la prévalence des étudiants détournant le système (étudiants qui utilisent le logiciel pour avancer sans consulter la documentation du cours). De petits problèmes scénarisés mènent à une plus grande proportion d'étudiants détournant le système que des scénarisations riches ou pas de scénarisation du tout (Baker, Yacef, 2009).
Traitement de données pour évaluation humaine	Représenter les données d'apprentissage de façon efficace permet d'agir sur la pédagogie.	Méthodes de visualisations.	Aide à la prise de décision ; visualisation de la trajectoire des étudiants durant leur scolarité ; identification de schémas parmi les étudiants ayant réussi ou non ; inférence d'étudiants à risques suffisamment tôt pour guider des interventions (Bowers, 2010).
		Carte de chaleur ou nuages de points, courbes d'apprentissage, diagrammes.	Évolution des performances (Koedinger *et al.*, 2010), mesure de la motivation (Hershkovitz, Nachmias, 2008).
Découverte avec les modèles	Les résultats d'une analyse par fouille de données sont utilisés dans une autre analyse par fouille de données. Les modèles utilisés ne sont pas nécessairement obtenus par des méthodes de prédiction mais ils peuvent être obtenus à travers d'autres approches comme le *clustering*, ou l'ingénierie des connaissances (Studer, 1998).	Modèles de prédiction utilisés au sein d'un autre modèle de prédiction.	Les modèles de prédiction d'apprentissage robuste des étudiants (Baker *et al.*, 2011) ont généralement dépendu de modèles sur les comportements métacognitifs des étudiants (Aleven *et al.*, 2006), qui à leur tour dépendent d'évaluations de connaissance latente d'étudiants (Corbett, Anderson, 1995) qui ont dépendu de modèles sur la structure du domaine (cf. Koedinger *et al.*, 2012).
		Un modèle de prédiction est utilisé dans une analyse de fouille de relations : à l'étude, la relation entre les prédictions du modèle initial et des variables additionnelles.	Beal *et al.*, 2008 : développe un modèle de prédiction d'étudiants détournant un système et le corrèle aux différences individuelles entre étudiants pour comprendre quels étudiants sont plus vraisemblablement engagés dans ce type de comportement.

1 QUESTIONS DE RECHERCHE EN SCIENCE DES DONNÉES DE L'APPRENTISSAGE

Hugues Labarthe et Vanda Luengo

Cette section présente un panorama de recherches expérimentales. Cet état de l'art ne peut rendre compte de l'ensemble des travaux entrepris dans le domaine des sciences des données de l'apprentissage mais il prétend éclairer les questionnements les plus actifs et les recherches les plus abouties. À l'hétérogénéité du millier de publications imprimées dans les actes des conférences, nous avons privilégié une revue exhaustive des articles passés par le second tamis du *Journal of Educational Data Mining* (JEDM) et *du Journal of Learning Analytics* (JLA), depuis leur création. Nous présentons ainsi 17 titres récemment parus, répartis dans quatre tableaux thématiques, qui présentent les travaux dont l'objectif, les techniques et les résultats nous semblaient les plus significatifs.

Pour chacun des thèmes retenus, ces recherches peuvent être motivées par des logiques diverses : d'une part, des protocoles de recherche fondamentale ambitionnent de se rapprocher au plus près de la « boîte noire » en captant la moindre interaction de l'apprenant pour agir sur son apprentissage et l'améliorer ; d'autre part, une recherche davantage appliquée se soucie de donner aux acteurs institutionnels les moyens de mesurer les flux d'apprenants et d'agir sur les performances du système, sur les succès et les échecs observés de l'apprentissage, en particulier l'abandon (*drop out*).

Prédire la progression de l'apprenant

La prédiction du parcours (comportemental ou cognitif) d'un apprenant est l'un des plus anciens problèmes des « environnements informatiques pour l'apprentissage humain » (EIAH), des systèmes experts aux MOOCs. Avec l'acquisition en temps réel de données d'apprentissage toujours plus fines et massives, les chercheurs ont développé des techniques d'analyse issues du Data Mining pour classer des profils, les orienter en fonction des capacités estimées, adapter les contenus, déployer des stratégies d'engagement et lutter contre le décrochage. Ces analyses sont menées à différents niveaux de granularité, de la simple interaction en temps réel d'un individu (microgenèse) aux apprentissages d'une cohorte sur une période donnée. Dans les communautés scientifiques, des modèles prédictifs et des méthodes variées sont mises en œuvre, comme l'atteste le tableau suivant (Tableau 3).

Tableau 3. Objectifs, techniques et apports des modèles de prédiction

Référence	Objectif	Données	Techniques	Résultats
Zimmermann *et al.*, 2015	Évaluer la puissance de prédiction des résultats scolaires et leur agrégation, comme indicateurs de performance.	81 variables pour une population de 171 étudiants.	*Post hoc*. Modèle de régression.	Les résultats de licence (undergraduate) peuvent expliquer 54 % de la variance dans les résultats de cycles supérieurs. La moyenne de notation globale de la 3e année est la variable la plus signifiante. Les résultats fournissent une base méthodologique pour dresser des lignes générales pour les comités d'admissions.
Knowles, 2015	Évaluer la probabilité de passage, pour chacun des 225 000 collégiens (grade 6 to 9), de 1 000 écoles du Wisconsin.	Cohorte des 12-13 ans en 2005. Les variables portent sur les résultats, l'assiduité, le comportement, la mobilité entre écoles, etc.	Deux fois par an. Régression logistique.	Le système fournit une probabilité de passage pour chaque apprenant + un classement (bas, modéré, haut). Les étudiants reçoivent une catégorie de risque pour 4 sous-domaines : scolarité, présence, comportement, mobilité. Le système identifie 65 % des échecs et des retards dans l'avancement avant l'entrée au lycée (high school) avec de faibles taux de fausses alarmes.
Ferguson, Clow, 2015	Évaluer la méthode de clustering mise au point par Kizilcek pour déterminer les modèles d'engagement sur un MOOC.	24 000 participants à 4 MOOCs délivrés par The Open University, en sciences physiques, en sciences de la vie, en arts et en économie.	Clustering (partitionnement de données).	Kizilcek *et al.*, 2013 établissent une méthode de clustering permettant d'identifier 4 modèles d'engagement. Appliquée au nouveau jeu de données, ni la méthode ni les modèles ne semblent adaptés à une ingénierie pédagogique sensiblement différente de celle du MOOC étudié par Kizilcek. L'enjeu est alors de proposer une nouvelle méthode d'analyse temporelle.
Aguiar *et al.*, 2014	Prédire l'attrition grâce à la mesure de l'engagement, fondée sur l'usage du portfolio numérique.	Portfolios électroniques de 429 étudiants en 1re année d'ingénierie à Notre Dame (Indiana, États-Unis).	Algorithmes de classification (*Naïve Bayes*, arbre de décision, régression logistique…).	Les auteurs proposent une méthode pour mesurer l'engagement des étudiants à partir de leurs portfolios numériques et montrent en quoi ces nouveaux indicateurs peuvent améliorer la qualité de prédiction.
Martin *et al.*, 2013	Établir les différents chemins dans l'apprentissage des fractions avec le jeu en ligne Refraction.	Codage a posteriori des transitions entre états du jeu.	Fouille de données et visualisations de graphes.	Un algorithme de classification permet de regrouper les étudiants en fonction des transitions entre états du jeu. Il identifie des profils d'apprenants : ceux allant droit à la solution, ceux qui expérimentent.

La problématique de la prédiction sert deux logiques différentes. Au niveau macro, dans une logique de recherche appliquée, les institutions éducatives sont en quête de modèles pour prévenir l'échec et améliorer leurs résultats. Au niveau micro, dans une perspective plus expérimentale, les chercheurs tentent d'inventer de nouveaux modèles pour se rapprocher au plus près de la « boîte noire » de l'apprenant en captant la moindre interaction dans l'espace d'une situation d'apprentissage : une frappe, un clic, un coup d'œil sont autant de traces pour informer les processus de l'apprentissage. Dans un cas comme dans l'autre, des modèles préexistants à la naissance des communautés EDM et LAK sont repris, itérativement affinés et adaptés aux nouveaux environnements d'apprentissage et à leurs enjeux.

Ainsi, de nombreuses solutions commerciales fondées sur différents modèles de prédiction du décrochage s'épanouissent sur les marchés éducatifs les plus concurrentiels. En 2007, John P. Campbell implémente sur le LMS Course Signals un calcul de régression logistique fondé sur les résultats, l'engagement, les indicateurs sociologiques, le nombre de messages postés sur le forum, de messages envoyés et de devoirs complétés [Arnold, Pistilli, 2012]. Essa et Ayad développent en retour, pour le LMS Desire2Learn, le *Student Success System*, un programme alliant une plus grande diversité de modèles à un outil de diagnostic.

Les données relatives aux activités d'apprentissage, aux prérequis, à l'ingénierie pédagogique et à la dynamique du cours, aux modalités d'apprentissage (en ligne, en présentiel ou hybride) sont mobilisées selon une approche comportementale. De la même façon que la présence en ligne d'un apprenant est évaluée selon ce même calcul de régression logistique, sa capacité à achever un parcours, sa participation et sa sociabilité sont évaluées. Le gestionnaire du cours choisit un ensemble de règles pertinentes, et leur poids relatif, pour obtenir une probabilité selon trois niveaux : « à risque », « risque potentiel » et « succès » [Essa, Ayad, 2012]. Ces recherches se poursuivent avec la massification de la formation en ligne et le succès des MOOCs à partir de 2012. Doug Clow quantifie l'usure de la motivation d'un apprenant à travers ses activités sur trois sites en ligne [Clow, 2013]. Kizilcec présente une méthode de classification qui identifie différentes trajectoires d'engagement basées sur les interactions de l'apprenant avec les lectures de vidéos et les exercices [Kizilcec *et al.*, 2013]. Première tentative de déconstruction du désengagement, ces travaux sont destinés à une constante réévaluation [Ferguson, Clow, 2015].

Dans la microgenèse de l'apprentissage, les trois modèles diagnostiques au fondement de la recherche expérimentale, posés dans la décennie 1990, se sont considérablement développés au bénéfice de la multiplication des EIAH, tels les tuteurs intelligents, mais peinent encore à être adaptés au format des LMS et MOOCs. Le modèle de *Knowledge Tracing* [Corbett, Anderson, 1995] dérive de la théorie ACT-R, qui vise à modéliser l'ensemble des processus cognitifs [Anderson, 1983]. Postulant que l'apprentissage de nouvelles connaissances est procédural, ce diagnostic vérifie le transfert de la connaissance de la mémoire déclarative à la mémoire procédurale. À ce modèle s'ajoute un second : le diagnostic *Constraint-based Modeling* [Mitrovic *et al.*, 2007 ; Ohlsson, 1994] qui stipule que l'apprentissage se fait par la confrontation de l'apprenant avec ses erreurs et se base sur la théorie *Performance Errors* [Ohlsson, 1996]. Seuls les états du problème ayant un intérêt pédagogique sont modélisés, permettant de relever une erreur typique du domaine. Enfin, le diagnostic *Control-based* [Minh Chieu *et al.*, 2010] est basé sur le modèle didactique cKc [Balacheff, Gaudin, 2002 ; Balacheff, 1995], implémenté comme un réseau bayésien dynamique.

Des évolutions de ces modèles sont ainsi proposées pour passer d'une démarche centrée Expert vers une démarche centrée Données, comme avec le modèle de régression linéaire *Additive Factor Model* (AFM) [Cen *et al.*, 2008] et les évolutions *Performance Factor Model* (PFM) [Pavlik *et al.*, 2009] ou *Performance Factor Analysis* (PFA). Enfin, ces approches centrées données permettent également la comparaison entre les modèles d'apprenants de façon plus systématique, tout en maintenant la possibilité de la réplicabilité des résultats [Lallé *et al.*, 2013].

Mesurer les interactions sociales

En s'appuyant sur les apports fondamentaux de Vitgovsky et du socio-constructivisme, une communauté scientifique s'attache à évaluer le rôle des interactions sociales au sein des EIAH : la communauté Computer-Supported Collaborative Learning (CSCL), née en 1995, porte cette conviction que la cognition et l'action humaine sont socialement et culturellement médiées. L'objectif est donc de comprendre et valoriser ces processus d'apprentissage collectif à l'encontre d'un enseignement transmissif, dont le Graal demeure la performance individuelle dans le cadre d'évaluations sommatives. L'apprentissage social permet aux individus de clarifier leurs intentions, étayer leur apprentissage, approfondir par l'échange. Son modèle émane des interactions de trois dimensions : cognitive (la situation problème), affective (la motivation) et sociale (les échanges).

L'analyse de l'apprentissage social (Social Learning Analytics) s'appuie notamment sur l'analyse des réseaux sociaux, permettant de détecter des communautés (Clauset *et al.*, 2004 ; Fortunato, 2010) ; d'identifier des sous-ensembles cohésifs dans un réseau (mesurables en termes de proximité, fréquence, affinités) (Reffay, Chanier, 2003) ; d'investiguer sur la densité de ces réseaux (Borgatti *et al.*, 2009) ; et, dans le cas d'un réseau égocentrique, d'identifier les personnes aidantes ou avec lesquelles surgissent des conflits liés à une incompréhension mutuelle (Haythornthwaite, De Laat, 2010). Caractériser les liens entre acteurs ajoute une dimension à l'analyse : les individus se fient généralement à tous types d'interlocuteurs quand il s'agit d'accéder à de nouvelles connaissances ou d'apprendre de façon informelle, mais ils se tournent vers des personnes de confiance pour approfondir des connaissances (Levin, Cross, 2004). Le concept de Social Learning Analytics émerge notamment autour des travaux de Ferguson (Ferguson, 2009). L'analyse de l'apprentissage social tend ainsi à révéler les régimes d'engagement des apprenants, soit dans une activité sociale (envoyer un message, sympathiser, suivre un tiers), soit en créant des traces réutilisables (publier, rechercher, taguer, évaluer).

Ce domaine de recherche prend de l'ampleur, à partir de 2011, avec la diffusion massive d'un nouveau type de formation en ligne : les MOOC (Massive Open Online Courses). Des cohortes de dizaines de milliers d'apprenants se retrouvent sur des plateformes d'apprentissage en ligne pour des formations relativement courtes (de 6 à 12 semaines) nécessitant, dans le cas des MOOC connectivistes (les cMOOC), de travailler de façon collaborative. C'est l'échelle même des expérimentations qui s'en trouve bousculée : on passe dès lors d'échantillons relativement restreints à la mobilisation de gisements de données considérables. Si les bases théoriques de cette analyse des interactions sociales sont largement fondées par les recherches des années 1990-2000, les emprunts plus récents à l'analyse des réseaux sociaux permettent d'élargir les techniques de mesures et de représentations de nouvelles biométriques, dont tout porte à croire qu'elles vont être toujours plus précises : l'usage d'une caméra Kinect permet déjà d'étudier la synchronisation des postures physiques, et pourra demain être corrélé à de nouveaux indicateurs sur la voix ou le rythme cardiaque, pour révéler l'intensité des interactions sociales.

Tableau 4. L'actualité de la recherche en Social Learning Analytics

Référence	Objectif	Données	Techniques	Résultats
Joksimović *et al.*, 2015	Mesurer, dans un cMOOC, l'influence des modèles de discours sur le capital social des apprenants.	17 324 messages sur deux sessions du Mooc *Connectivism and Connective Knowledge*.	Graphes non dirigés hebdomadaires pour chaque média (FB, Twitter, blogs). Analyse linguistique selon la métrique COH.	La qualité des écrits est déterminante pour l'apprentissage. Les apprenants d'un cMOOC se joignent plus volontiers à des individus qui utilisent un style informel, narratif et cependant cohérent. Le langage définit donc la position structurelle dans un réseau social d'apprentissage.
Waters *et al.*, 2014	Identifier, classifier les formes de collaborations (fraude, compétences individuelles).	Données fictives et réelles : examen de 203 vrai/faux passés par 97 étudiants d'informatique.	Approche bayésienne, modèle de Rash, méthode Sparse Factor Analysis.	À partir de deux nouveaux algorithmes, les auteurs ont su, tant sur des données fictives que réelles, détecter et identifier des formes de collaboration. Cette méthode permet à l'instructeur de se focaliser sur un petit nombre d'apprenants.
Skrypnyk *et al.*, 2015	Analyser le réseau social d'un MOOC connectiviste à travers les interactions via Twitter.	2 483 tweets de 800 participants. Données démographiques publiques sur les réseaux et le Web.	Mesures au niveau des nœuds (centralité de degré, de proximité, d'intermédiarité, de vecteur propre, de clique) et du réseau (modularité).	Au cours de l'avancement du MOOC, un groupe de participants a joué un rôle similaire à celui des animateurs. Les nœuds avec les plus hauts degrés de centralité concernent les hashtags, confirmant ainsi l'intérêt des apprenants pour une des thématiques changeantes. Des sous-ensembles ont émergé avec, pour chacun, des contributions spécifiques.
Schneider *et al.*, 2013	Tester une analyse séquentielle des regards et évaluer la robustesse de cette métrique pour prédire l'apprentissage.	22 dyades isolées communiquent via un canal audio seul ou un canal audio + un indicateur de regard du collaborateur.	La position du regard (nœuds) et son mouvement (transition) sur un support pédagogique sont transformés en graphes non dirigés.	Les auteurs démontrent l'importance de considérer le regard comme indice d'attention soutenue mutuelle. En adaptant la théorie des graphes à leur problématique, ils identifient de nouveaux indicateurs (nombre de liens, taille des nœuds, intermédiarité), qui permettent d'évaluer l'intensité d'une collaboration et la compréhension réciproque des apprenants.
Ahn, 2013	Identifier des types de comportement, mesurer la corrélation entre ces comportements et les compétences du XXIe siècle.	Enquête sur 189 participants. Collecte de données agrégées via l'API [1] Facebook de 99 apprenants.	Analyse factorielle, modèles de régression.	Le modèle de régression prend les activités pour variables indépendantes (envoyer un message, partager une information, devenir ami, joindre un groupe), et les compétences pour variables dépendantes (négociation, réseautage, esprit critique, jeu, multitâche, appropriation, navigation transmédia). Il illustre quel type d'activités renforce quelles compétences.

1. API, pour Application Programming Interface, ou interface de programmation applicative, permet de rendre les données ou les fonctionnalités d'une application existante disponibles, afin que d'autres applications les utilisent.

Analyser le discours

L'apprentissage ne résulte pas des seules capacités cognitives ou du comportement d'un apprenant isolé [Mercer, 2004]. Le langage est l'un des premiers vecteurs par lequel les apprenants construisent du sens : son usage est influencé par les buts, les sentiments et les relations, très variables selon les contextes [Wells, Claxton, 2002]. Le langage fait partie de la situation d'apprentissage [Gee, Green, 1998 ; Wertsch, 1991]. Il supporte et conditionne le succès d'activités conjointes, tant pour la combinaison des connaissances, des compétences, l'utilisation d'outils et la capacité à travailler ensemble. Comprendre l'apprentissage de cette façon requiert de faire attention aux processus de construction du savoir dans le groupe. L'analyse du discours tend à rendre ces processus d'apprentissage visibles et ouvre la possibilité de les améliorer à différents niveaux, de l'individu au groupe restreint, puis à la cohorte.

Avec le déploiement de l'apprentissage à distance, le forum, lieu d'une discussion asynchrone, devient un objet de recherche. Les forums aident à trouver et partager l'information et promeuvent la réflexion critique de communautés d'apprenants [Anderson, 1996]. Plusieurs modèles d'apprentissage en ligne ont été développés, dont celui de Communauté d'interrogation [Community of Inquiry, CoI] [Garrison *et al.*, 1999]. Ce modèle part du principe que dans toute collectivité, l'apprentissage se fonde sur l'interaction de trois éléments principaux : la présence de la société signifie la capacité des apprenants à participer à la vie de la collectivité, aux plans social et affectif ; la présence de la cognition indique l'aptitude à fabriquer du sens grâce à une communication soutenue des milieux en ligne ; la présence de l'enseignement désigne des activités visant à faire activement participer les apprenants et à maintenir la communication. Analyser une conversation entre apprenants permet d'appréhender leur degré d'ouverture, leur capacité de décentrement et leur propension à reformuler leur pensée. Dans ce cadre, trois modalités de discours ont été distinguées par Mercer : discours dialectique, cumulatif et exploratoire [Mercer, 2004].

Dans la lignée de ces apports théoriques, ce sont les communautés investies dans la conception de modèles de remédiation au sein d'un tuteur intelligent qui se sont employées à dégager des indicateurs permettant d'identifier automatiquement les modalités d'interventions pour étayer un apprentissage. Avec l'émergence des réseaux sociaux, les chercheurs en Learning Analytics s'efforcent d'identifier les types d'échanges : challenges, explorations, évaluations, raisonnements. Ainsi, la plateforme de délibération Cohere a été étendue par De Liddo et ses collègues pour fournir des Learning Analytics qui identifient : les sujets de conversation et le point de vue des apprenants ; les types de contributions des apprenants et leurs accords/désaccords ; l'organisation du réseau discursif et le rôle des apprenants dans ce réseau ; les liens entre apprenants [De Liddo *et al.*, 2011].

Tableau 5. L'actualité de la recherche en analyse de discours

Référence	Objectif	Données	Technique	Résultats
Snow *et al.*, 2015	Proposer une méthodologie pour évaluer la souplesse stylistique des rédacteurs confirmés.	45 lycéens complètent 2 tests (capacités avant/après) et 16 essais, en 8 sessions.	Traitement automatique du langage naturel (TALN) : narrativité, cohésion, entropie.	Les auteurs proposent une méthode d'évaluation discrète et rapide permettant de détecter les relations entre souplesse stylistique, lecture, vocabulaire, connaissances et écriture. Seule la combinaison entre TALN et entropie permet de saisir la progression des apprenants vers plus de souplesse stylistique.
Rebolledo-Mendez *et al.*, 2013	Dispositif pour rechercher des modèles d'interaction en lien avec l'étayage motivationnel d'un tuteur intelligent.	70 fichiers de logs capturant le comportement de 35 apprenants avec un tuteur intelligent.	Post hoc. Corrélation entre types collaboratifs. Mesure d'association entre les actions de l'apprenant (cf. D'Mello *et al.*, 2010).	Un graphe dirigé permet de visualiser et d'identifier des transitions significatives dérivées de dyades d'actions. Première approche à compléter par des entretiens avec les apprenants.
Ming Ming, Nobuko, 2014	Identifier le lien entre cognition informelle (opinion, anecdotes, description) et formelle (explication théorique).	1 330 messages asynchrones écrits par 17 étudiants, codés par eux-mêmes, sur un cours en ligne.	Méthode de Monte-Carlo par chaînes de Markov (MCMC) itérative.	Évalue comment 3 types cognitifs (opinion informelle, élaboration, preuve) et 3 types de métacognition sociale (demander une explication, demander comment utiliser, opinions différentes) renforcent la vraisemblance de nouvelles informations ou explications théoriques dans les messages postérieurs.
D'Mello *et al.*, 2010	Méthode pour la détection automatique de modèles collaboratifs avec un tuteur intelligent.	50 heures de tutorat filmées entre apprenants et experts et 47 296 interactions codées.	Construction d'un graphe orienté des transitions entre interactions de l'apprenant et test d'hypothèses.	À la différence des modèles de Markov cachés, des règles d'association séquentielle et des modèles log-linéaires, cette étude détecte les transitions fréquentes et les représente sous forme de graphe. L'étude des stratégies, actions et dialogues d'experts humains permet aux auteurs de concevoir la modélisation informatique d'un tuteur intelligent.

Donner à voir l'apprentissage

À l'origine des Learning Analytics, les Visual Data Analytics reposent sur deux principes essentiels : donner aux acteurs de la communauté éducative un pouvoir de décision par l'exploration du modèle qui sous-tend les données d'apprentissage elles-mêmes (Duval, 2011). Les tableaux de bord (*dashboards*) tiennent des systèmes d'aide à la décision (Decision Support System, DSS) apparus dans les années 1970, dans la foulée de la numérisation des processus métiers. Dans le domaine éducatif, les tableaux de bords numériques se sont imposés avec les premiers LMS (Learning Management Systems). Tous les LMS ne proposent pas des interfaces de Visual Data Analytics aussi poussées que Blackboard Analytics, Brightspace, Signals (Purdue University) et ALAS-KA (Khan Academy). Au-delà d'une consultation des données d'apprentissage, ces logiciels prétendent permettre de dépister le décrochage, de personnaliser les contenus. En permettant aux apprenants d'accéder à leurs propres données, ces technologies coïncident avec l'émergence du Quantified Self ou Self Tracking (mesure de soi), des techniques d'évaluation quantitative systématique.

Les tableaux de bord numériques offrent une interprétation visuelle de larges ensembles de données pour découvrir, interroger, comprendre les modèles portés par ces données afin de modifier ses représentations (Verbert *et al.*, 2013). Ces tableaux visent à seconder l'enseignant dans le développement de l'attention, la compréhension et la métacognition

des apprenants. D'un simple clic, ils permettent de passer d'une visualisation de haut niveau à l'interrogation des données de bas niveau. Pour répondre aux aspirations d'une économie de l'attention [Goldhaber, 1997], la conception de l'interface [User Interface Design] devrait être claire, interactive et modulable, de façon à pouvoir répondre aux besoins spécifiques de chaque type d'utilisateurs. Les tableaux de bord sur le marché sont relativement bien documentés sur le plan scientifique [Verbert *et al.*, 2013]. Santos *et al.* présentent les résultats du tableau de bord StepUp! et discutent son apport aux problèmes et besoins des apprenants [Santos *et al.*, 2013]. Les 1 500 étudiants de Purdue qui ont expérimenté le tableau de bord Course Signal ont obtenu des résultats significatifs d'assiduité, en comparaison avec des cohortes similaires [Arnold, Pistilli, 2012].

Au département de sciences de l'éducation, à Utrecht [Pays-Bas], van Leeuwen et ses collègues ont évalué la pertinence du tableau de bord dans la gestion de classe. Un enseignant doit surveiller les activités cognitives de ses apprenants, répartis en îlots [Van Leeuwen, 2015]. Dans une première expérimentation [The Concept Trail], les apprenants doivent mener une tâche collaborative par messagerie synchrone. Au fur et à mesure de la discussion, des concepts essentiels pour résoudre la tâche affleurent. L'enseignant suit en temps réel la progression des groupes par le biais de lignes de temps où s'affichent ces concepts, échangés via la messagerie. Une deuxième expérimentation [Progress Statistics] permet à l'enseignant de suivre le nombre de mots écrits par les groupes d'apprenants dans l'éditeur collaboratif et dans le chat associé. Ces outils confèrent à l'enseignant une vue d'ensemble et une capacité d'intervention à distance qu'il ne peut avoir pendant un cours en regardant par-dessus l'épaule de l'un ou de l'autre. Dans ces deux expériences, la visualisation n'améliore ni ne baisse la détection des problèmes cognitifs. En revanche, elle augmente la fréquence des interventions de l'enseignant et renforcerait sa capacité à poser un diagnostic.

Tableau 6. Évaluer l'utilité des tableaux de bord pour l'apprentissage

Référence	Objectif	Données	Dispositif	Apports et limites
Van Leeuwen, 2015	Assister en temps réel les enseignants dans l'évaluation d'apprenants engagés sur une tâche collaborative.	Évaluation de 5 groupes par 14 enseignants. Fichiers de logs des actions de ces enseignants en temps réel.	Représentation de matrices avec en colonne les groupes et en ligne la *timeline* des actions de l'enseignant.	Cette étude permet de comprendre comment les enseignants interagissent avec des visualisations de données d'apprentissage.
Martinez-Maldonado *et al.*, 2015	Élaborer un processus itératif de conception, validation et déploiement d'outils de visualisation pour favoriser l'attention de l'enseignant sur la collaboration et la progression des groupes.	Manipulations physiques d'objets virtuels sur des tables interactives, temps de parole et progression des tâches de 500 étudiants.	Entretiens et scénarios d'utilisation à partir de prototypes papier ; études contrôlées sur un simulateur ; analyse des interventions en situation réelle.	Cette étude identifie 5 étapes dans le processus de conception d'outils de visualisation pour renforcer l'attention dans une activité de groupes sur tables interactives : identification du problème, validation des prototypes, simulation avec de vrais enseignants, études pilotes, expérimentation en classe. Des travaux complémentaires sont nécessaires pour adapter ce processus à un environnement d'apprentissage hybride ou à distance.
Charleer *et al.*, 2014	Évaluer LARAe, un tableau de bord pour enseignants : visualisation des traces, badges, contenus des étudiants.	26 étudiants en groupes de 3. Matériaux et échanges en ligne. Évaluation de l'utilisation de LARAe par 6 enseignants sur écran 27 pouces.	Twitter API, fils RSS, Badge API. Visualisations : matrice des badges, ligne de temps, liste d'activités, détail des activités, options de filtrage.	LARAe donne les moyens de visualiser l'abondance des traces laissées par les étudiants et les enseignants durant un cours. L'attribution de badges met l'accent sur les activités les plus importantes et les objectifs de formation. LARAe aide les enseignants et les étudiants à tirer parti des données qu'ils génèrent.

Sur un marché de plus de 2 milliards de dollars [Roberge, 2013], les éditeurs de tableaux de bord s'engagent dans une levée de fonds et une bataille juridique intense. En réaction à l'accumulation de brevets, George Siemens porte, dès l'origine de SoLAR, un projet de conception d'une plateforme ouverte qui intégrerait une gamme variée de modules d'analyse dédiés à l'apprentissage [Siemens *et al.*, 2011]. L'enjeu du libre répond alors à trois principaux défis : l'innovation et la généralisation des tableaux de bords exigent que les traitements, les algorithmes et les technologies employés soient ouverts ; la plateforme doit être modulaire et permettre d'intégrer tous types d'outils liés à l'analyse et à l'adaptation de l'apprentissage ; enfin, cette plateforme doit offrir des fonctionnalités appropriées à chacun, chercheurs et producteurs de contenus, assistants d'éducation et apprenants, enseignants et personnels d'encadrement. Par ailleurs, une autre initiative, Apereo OpenDashboard, est désormais accessible en Open Source.

2

LES OUTILS ET MÉTHODES DES LEARNING ANALYTICS

Marie Lefevre et Sébastien Iksal (coord.), Julien Broisin, Olivier Champalle, Valérie Fontanieu, Christine Michel, Laurent Polese, Amel Yessad

Les Learning Analytics, ou l'analyse des traces d'apprentissage, sont définis comme « l'évaluation, l'analyse, la collecte et la communication des données relatives aux apprenants, leur contexte d'apprentissage, dans la perspective d'une compréhension et d'une optimisation de l'apprentissage et de son environnement » [Long *et al.*, 2011]. Les Learning Analytics suivent donc un cycle passant par les étapes de **collecte des traces, d'analyse de ces traces et d'exploitation (souvent de la visualisation)** des traces et des indicateurs produits par l'analyse [Fayyad *et al.*, 1996 ; Clow, 2012 ; Stamper *et al.*, 2011].

La **collecte** des données est une étape primordiale car il s'agit de récupérer toutes les données numériques représentant l'activité des usagers afin de mener un processus d'analyse et d'obtenir un reflet du déroulement des situations d'apprentissage. Ainsi, la collecte des données concerne l'observation de l'apprenant et le **traçage** de ses interactions médiées par les outils, le **stockage** des traces récoltées et l'**import** de traces dans les outils d'analyse. Ces différentes actions peuvent être faites par un unique outil de collecte et d'analyse, ou peuvent être réparties dans différents outils à combiner pour mener à bien les analyses.

L'**analyse** des traces consiste à manipuler les données pour essayer d'extraire des informations. Certaines analyses vont nécessiter des **pré-traitements** sur les données pour les mettre en forme, les nettoyer, vérifier leur fiabilité, etc.

Ces étapes sont parfois complétées par une étape de **partage** des traces, des processus d'analyse, des indicateurs et/ou des visualisations produites.

Tout au long de ce cycle, plusieurs acteurs entrent en jeu [Greller *et al.*, 2012] : **les apprenants, les équipes techniques et pédagogiques, les institutions, les familles et enfin, les chercheurs**. Ces acteurs ont différents rôles lors de l'analyse et selon les contextes et les outils d'analyse. Les apprenants sont tantôt sujets de l'observation pour produire des traces, tantôt ils prennent le rôle d'analystes, en utilisant des outils réflexifs permettant de comprendre leurs traces. Il en est de même pour les enseignants.

Cette section présente une étude des outils et méthodes proposés dans le cadre des Learning Analytics par la communauté de recherche française. Nous avons essayé, à chaque fois, de compléter cette étude en donnant aussi des exemples d'outils professionnels.

Présentation de l'étude

Notre étude a porté sur 34 outils ou méthodes (cf. annexe) issus ou utilisés dans le cadre des projets de recherche en France. Il est possible que certains outils soient absents, probablement en raison de l'absence ou quasi-absence de publications. La première partie de l'étude a consisté à établir une liste de ces outils en précisant différentes caractéristiques connues, pour chacun. L'une des finalités porte sur l'identification de catégories et/ou de regroupements selon des caractéristiques communes. L'autre se focalise sur la mise en correspondance de ces outils avec les différentes étapes classiques d'un cycle d'analyse de données : la collecte, le stockage, l'analyse et la visualisation. En parallèle, nous souhaitons mettre en évidence les aspects liés au partage et à l'interopérabilité de ces outils.

Nous avons ensuite écarté les outils qui ne sont plus disponibles ou qui sont obsolètes aujourd'hui. Nous sommes arrivés à une liste réduite de 18 outils et méthodes que nous avons pu classer selon les différentes étapes du cycle d'observation et selon le public visé par ces outils. Nous avons identifié les rôles des personnes amenées à prendre l'outil en main (analyste, développeur, décideur), ainsi que les rôles des personnes destinataires des résultats d'analyse (apprenant, enseignant, administratif). Nous nous focalisons ici sur la liste restreinte que nous avons catégorisée selon le cycle des Learning Analytics.

> **Outils permettant la collecte de données :** D3K0DE, DDART, kTBS4LA, Laalys, Lab4ce, Limesurvey, SBT-IM, Tatiana, TRAVIS, UnderTracks, UTL

> **Outils permettant le stockage des données :** Abstract, D3K0DE, DDART, DisKit, dmt4sp, Laalys, Lab4ce, LEA4AP, Limesurvey, SBT-IM, Tatiana, TRAVIS, T-store, UTL

> **Outils permettant l'analyse des données :** Abstract, D3K0DE, DDART, DisKit, dmt4sp, kTBS4LA, Laalys, LEA4AP, SBT-IM, Taaabs, Tatiana, Transmute, TRAVIS, T-store, UnderTracks, UTL, EM0DA

> **Outils permettant la visualisation des données :** Abstract, D3K0DE, DDART, DisKit (+Transmute), dmt4sp (textuelle), kTBS4LA (+Taaabs), Laalys, Lab4ce, LEA4AP, Limesurvey, SBT-IM, Tatiana, TRAVIS, UnderTracks, UTL, EM0DA

> **Outils permettant le partage et l'interopérabilité :** D3K0DE, SBT-IM, Tatiana, UTL.

Avec ces outils issus de la recherche, nous avons aussi tenu à présenter des outils professionnels disponibles et permettant de répondre en partie aux attentes et aux besoins des analystes et des enseignants.

Résultats de l'étude présentés selon les étapes des Learning Analytics

L'analyse des traces d'apprentissage suit un cycle composé de plusieurs étapes : collecte, analyse et exploitation/visualisation. Dans cette section, les outils analysés sont classés et détaillés suivant ces différentes étapes.

CONSTRUIRE LE JEU DE TRACES

Les types de traces les plus communément considérés sont les interactions de l'utilisateur avec le dispositif de formation. Ces traces sont enregistrées automatiquement. Peu de systèmes utilisent les traces de logs standards construites par les applications client-serveur car elles manquent de précision. Les systèmes implémentent plutôt des capteurs spécifiques dans les applications de formation sur la base des modèles de traces nécessaires pour réaliser les analyses. Les capteurs sont donc positionnés sur les modalités d'interaction pertinentes pour adapter le système ou construire les indicateurs qui permettent la prise de décision des utilisateurs considérés (apprenant, enseignant ou analyste).

Les informations collectées suite à ces interactions peuvent être relatives à l'interaction elle-même (quelle fonctionnalité est utilisée, à quel moment, pendant combien de temps…) ou aux résultats de l'interaction (contenu d'un message de *chat*, séquence audio ou vidéo relative à une conversation, document produit à plusieurs…). Les traces descriptives de l'interaction sont structurées sur la base d'un modèle, alors que les traces des produits de l'activité ne le sont souvent pas.

D'autres traces, désignées comme « déclarées, ou auto-rapportées », sont aussi considérées dans les systèmes d'analyse. Elles sont renseignées manuellement par l'utilisateur. Elles correspondent à un jugement de l'utilisateur sur son activité ou sur les acteurs partenaires de son activité. Ces traces peuvent être fermées et structurées selon un modèle, ou être ouvertes. Ces traces sont généralement textuelles.

Étape de collecte

Il existe différentes solutions pour la collecte de données selon les possibilités techniques (disponibilité ou non d'une connexion réseau) ou les besoins en matière de calculs :

> la collecte de données d'observation avec stockage dans les applications éducatives. Ce stockage local permet de tracer sur des outils non connectés et d'utiliser les traces avec différents systèmes d'analyse. Par exemple, avec Tactiléo, les élèves sortent de l'établissement avec les tablettes qui tracent leurs activités et, à leur retour, les données peuvent être transférées vers des systèmes d'analyse évolués. Enfin, les traces proviennent des applications elles-mêmes, et nous pouvons ainsi obtenir des informations très précises ;

> la collecte de données d'observation avec stockage dans une base de traces, telle TraceMe. Cette solution délocalisée a l'avantage d'imposer un format unique, un standard que respectent tous les systèmes d'apprentissage qui envoient leurs traces. Cela permet par la suite de capitaliser et de réutiliser les processus d'analyse sur différents systèmes d'apprentissage ;

> la collecte de données d'observation avec stockage dans un outil d'analyse, tel Travis, T-Store, DDART, UTL, D3KODE. Ici, le système d'analyse intègre directement les traces et peut donc réaliser les analyses en temps réel, ce qui permet de fournir des informations rapidement aux enseignants. L'inconvénient repose sur le fait que les traces sont plus difficilement partageables entre les différents systèmes d'analyse ;

> enfin, la collecte de données autres, comme les sondages de LimeSurvey. C'est une technique différente qui repose sur les réponses apportées par les différents acteurs. Si elle relève plus du ressenti que de l'observation réelle, elle apporte toutefois des informations complémentaires et utiles pour améliorer la compréhension de l'observation.

Préparation des données

Les données collectées en vue de la production de Learning Analytics sont souvent des données hétérogènes, tant sur la provenance que sur le format et le contenu. Un travail conséquent est nécessaire afin d'uniformiser ces données et de les préparer pour les outils d'analyse qui possèdent un nombre limité de formats en entrée. C'est ici qu'interviennent les outils comme Talend [2] Open Studio, dont l'objectif est de préparer ces données. Talend est un ETL (*Extract Transform and Load*) et repose donc sur trois étapes : extraire, transformer et charger les données.

Une fois les informations extraites des différentes sources, Talend permet de réaliser plusieurs opérations de transformation comme de la standardisation (mettre des dates sur un format unique), de la suppression de redondance, du tri, de la vérification en supprimant les données inutilisables ou en signalant les anomalies… Toutes ces opérations peuvent être définies et paramétrées sous la forme de routines, de façon à être réutilisables sur d'autres jeux de données. Enfin, la dernière étape consiste à charger les nouvelles données obtenues dans les nouveaux emplacements comme les bases de données exploitées directement par les outils d'analyse de données. L'intégration de données avec Talend peut se pratiquer soit avec l'édition Open Source (Talend Open Studio for Data Integration), soit avec la version professionnelle (Talend Data Integration).

2. https://fr.talend.com/

Il est important de noter que même si cet outil est graphique et puissant, il est toutefois réservé à des utilisateurs avertis qui maîtrisent les modèles de données et qui possèdent les compétences nécessaires à l'élaboration des routines de transformation.

Modèles de traces

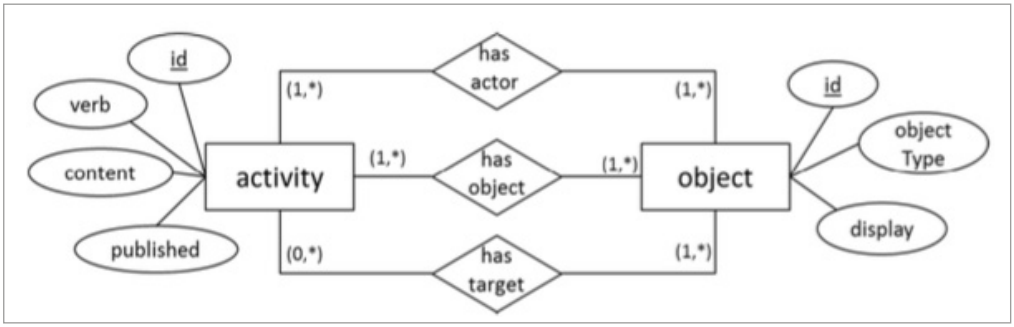
Les outils présentés plus loin dans ce document s'appuient sur des modèles de données propriétaires. Parmi les nombreux modèles de traces développés par la communauté scientifique, nous décrivons dans cette section ceux ayant fait l'objet d'une adoption à grande échelle.

Attention Metadata, ou **Attention.XML**, est une spécification ouverte pour tracer et partager les artefacts sur lesquels l'utilisateur a porté son attention (un document lu, regardé ou écouté par un utilisateur, par exemple). La conception de Attention.XML repose sur trois prémisses : (1) les *flux* Attention correspondent à un utilisateur spécifique, (2) les enregistrements Attention représentent les *objets* sur lesquels l'utilisateur a porté son attention, et (3) les objets peuvent être collectés à partir de différentes *sources* de données.

Toutefois, le manque de détails concernant l'usage des objets par les utilisateurs a été considéré comme un inconvénient majeur de cette initiative. Alors [Wolpers *et al.*, 2007] ont introduit **Contextualized Attention Metadata** (CAM), une extension d'Attention.XML, afin de capturer plus finement les informations comportementales des utilisateurs. Un événement est décrit, entre autres, par une estampille horaire et une description. Il peut être associé à une action d'un certain type et détaillée par des données associées. De plus, un événement se produit dans un certain contexte lors d'une session particulière. CAM a été adoptée dans le cadre du projet européen Open Discovery Space dont l'objectif était de concevoir un portail dédié à la recherche et à la recommandation de ressources, ainsi qu'à la constitution de communautés de pratique.

La spécification **Activity Streams** [Snell *et al.*, 2012a] définit un format pour décrire les activités réalisées par un utilisateur sur un système ou une application ; elle n'a pas été conçue spécifiquement pour l'éducation. Un « Activity Streams » est une collection d'une ou plusieurs activités réalisées par un individu, et généralement exprimée au format JSON.

Figure 1. Représentation simplifiée du schéma Activity Streams



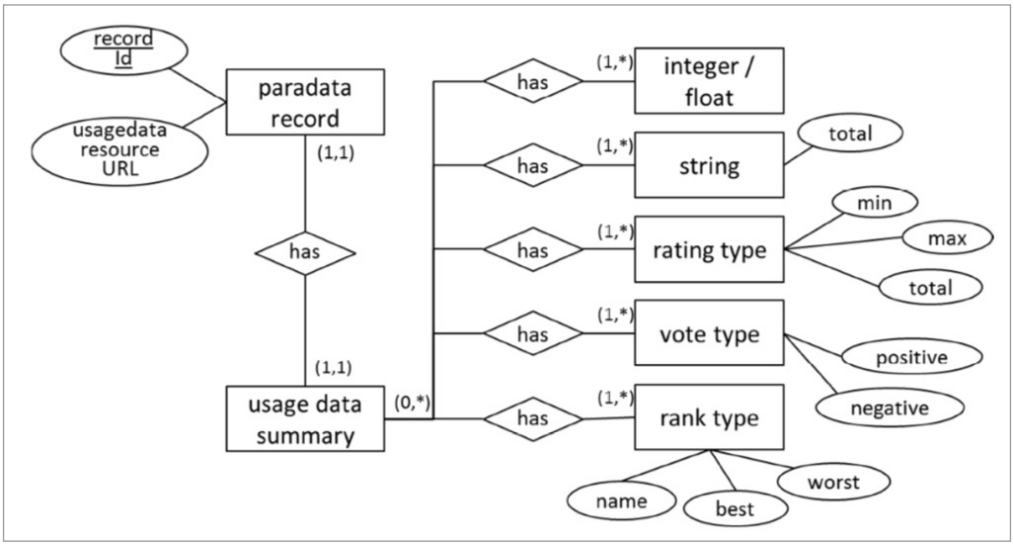
La Figure 1 montre les éléments principaux du schéma « Activity Streams ». Une activité doit au moins contenir une description de l'entité ayant réalisé l'activité (propriété *actor*) ainsi que la date et l'heure auxquelles l'activité a été publiée (propriété *published*). Le groupe de travail Activity Streams recommande qu'une activité contienne également un verbe, un objet et un identifiant. Le verbe identifie l'action décrite par l'activité (« consulte », « évalue », « envoie »...), l'objet décrit l'artefact sur lequel a été réalisée l'activité (la ressource consultée ou le message envoyé, par exemple), et l'identifiant permet de retrouver une activité parmi un ensemble d'activités. La propriété *target* est optionnelle et peut être utilisée lorsqu'elle est indiquée par le verbe ; par exemple, dans l'activité « Pierre a envoyé un message à Paul », « Paul » est la cible de l'activité.

Le schéma « Activity Base » [Snell *et al.*, 2012b] propose une spécification pour décrire les valeurs des propriétés *actor*, *object* et *target,* mais n'importe quel objet Activity Streams peut être étendu par d'autres propriétés qui ne sont pas définies dans cette spécification, afin d'apporter autant de flexibilité que possible.

Le **Learning Registry** [Bienkowsky *et al.*, 2012] est une infrastructure qui permet aux enseignants et apprenants de découvrir et d'utiliser des ressources pédagogiques stockées dans différents systèmes internationaux, américains en particulier. Learning Registry stocke des informations sociales telles que les tags, commentaires ou évaluations réalisées par les utilisateurs, en plus des traditionnelles métadonnées décrivant une ressource pédagogique. Ces données, appelées « paradata », sont ensuite partagées dans une infrastructure commune à des fins d'agrégation et/ou d'analyse. Le Learning Registry est inspiré d'Activity Streams. L'élément racine est une collection d'activités caractérisées par (1) un acteur décrivant l'entité ou la personne ayant réalisé l'action, (2) un verbe traduisant le type d'activité, (3) l'objet sur lequel l'activité a été effectuée, (4) une liste d'objets liée à cet objet, et (5) une description textuelle et l'estampille horaire correspondant à la publication de l'activité.

National Science Digital Library (NSDL[3]) propose des ressources digitales de qualité dans le domaine des sciences et technologies de l'ingénieur et des mathématiques, à destination de la communauté éducative. Les ressources sont décrites à l'aide de métadonnées LOM (domaine d'apprentissage, type d'audience et niveau, par exemple) [Learning Technology Standards Committee, 2002], mais aussi par des données statistiques d'usage et des données renseignées par les utilisateurs du système (paradata telles que des évaluations ou des commentaires, par exemple) [Blomer, 2012].

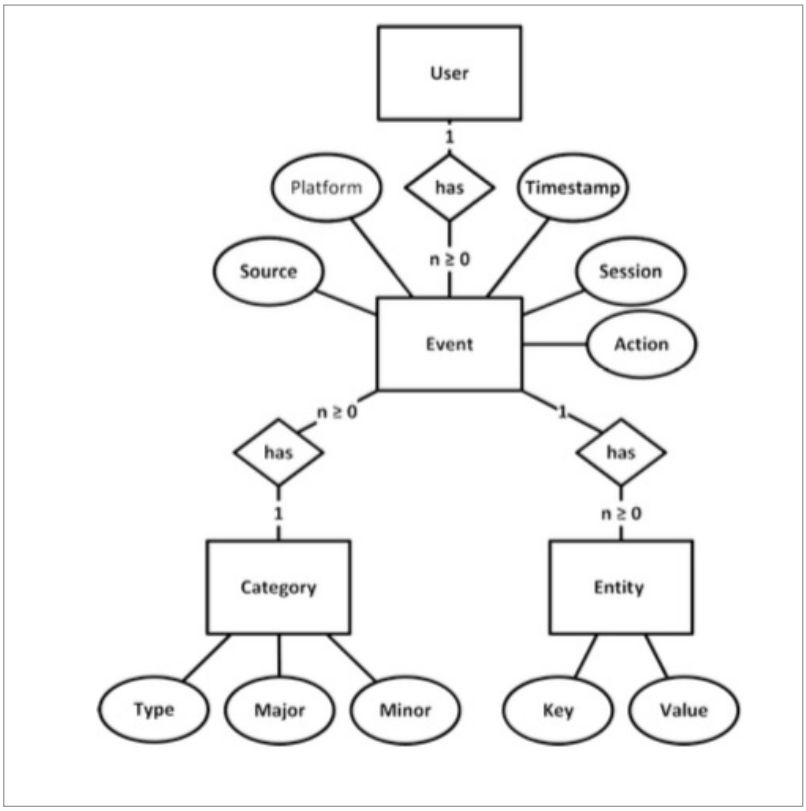
Figure 2. Schéma simplifié des paradata NSDL [Niemann et al., 2012]



Le schéma de données décrivant un enregistrement NSDL est illustré par la Figure 2.
Un enregistrement est identifié de manière unique, et renferme l'URL de la ressource pour lesquelles les paradata s'appliquent. L'élément le plus important décrivant un enregistrement NSDL est *UsageDataSummary*, puisque celui-ci comprend différents types de données représentant différentes informations : le nombre de fois qu'une certaine action a été réalisée sur la ressource (commentée, lue, téléchargée…) ; une valeur textuelle exprimant un commentaire ou un tag ; la valeur de l'évaluation moyenne de la ressource ; le nombre de « *(un)like* » d'une ressource ; le rang de la ressource au sein d'un ensemble de ressources. Cet élément contient également le laps de temps pendant lequel les données statistiques ont été collectées. Notons que d'autres informations peuvent également décrire un enregistrement, puisque n'importe quel élément supplémentaire peut être spécifié.

3. http://www.nsdl.org

Figure 3. Schéma LCDM [Muslim *et al.*, 2016]



Enfin, une autre initiative appelée Learning Context Data Model (LCDM) [Lukarov *et al.*, 2014] a été développée dans le cadre du projet Learning Context [4]. Cette proposition se concentre sur des propriétés d'extensibilité et de mobilité, et adopte une approche centrée utilisateur pour représenter les données d'interaction issues d'environnements mobiles, de plateformes web, etc. Le schéma illustré par la Figure 3 vise à représenter le contexte dans lequel se produisent les interactions, avec une sémantique la plus riche possible. LCDM associe un utilisateur à un ensemble d'événements décrits par, entre autres informations, le nom de l'application à l'origine de l'événement, le type de plateforme utilisée (mobile, Web, ordinateur de bureau), l'estampille horaire, le type d'action réalisée par l'utilisateur, la catégorie de l'événement (privé, professionnel ou académique), et l'entité sur laquelle s'est produit l'événement (par exemple, le titre d'une ressource éducative).

Au-delà de ces initiatives internationales, des standards voient le jour et sont intégrés dans les outils existants ou en cours de développement. Deux principaux standards sont largement utilisés dans les outils de collecte et d'analyse de traces : **xAPI** et **IMS Caliper.**

Experience API, ou xAPI, a vu le jour avec le « Projet Tin Can », en 2013. xAPI, supporté par l'initiative Advanced Distributed Learning (ADL), a été conçu par une communauté de chercheurs, ingénieurs et praticiens pour succéder au standard SCORM, dont l'objectif était de proposer des paquetages capables de structurer des scénarios pédagogiques dans le but de les déployer sur différentes plateformes d'apprentissage en ligne de type Moodle. À l'heure actuelle, la spécification xAPI est utilisée pour capturer les expériences d'apprentissage des utilisateurs sur n'importe quel logiciel de bureau ou fondé sur les technologies du Web : « The Experience API (or xAPI) is a new specification for learning technology that makes it possible to collect data about the wide range of experiences a person has (online and offline) [5]. » xAPI s'appuie sur le paradigme sujet, verbe, complément (« *I did this* » ou « *je fais ceci* », en français, par exemple) pour décrire les expériences d'apprentissage, ou statements, et propose un modèle de représentation des données qui doivent être capturées ; xAPI est une spécification, ce n'est pas un outil ou un logiciel.

4. http://learning-context.de/
5. https://xapi.com/overview/

En vue de faciliter l'échange et l'interprétation des *statements* xAPI, la communauté xAPI définit des référentiels de vocabulaire extensibles pour décrire les sujets, verbes et compléments spécifiques à un domaine particulier ou à une communauté de pratique [6]. Ce principe permet de s'assurer que toute application utilise les mêmes termes pour désigner les mêmes artefacts.

Une autre initiative de l'IMS Global Consortium, nommée Caliper Analytics, a été proposée en 2015 pour permettre aux établissements de recueillir des données d'apprentissage dans le but de mieux comprendre et visualiser les données relatant les activités d'apprentissage, et de présenter ces informations aux étudiants et enseignants de manière intelligible, pour améliorer les différents processus d'apprentissage, de conception, de tutorat, etc.

IMS Caliper définit un certain nombre de profils de métriques, chacun d'eux modélisant une activité d'apprentissage ou une activité supportant les processus d'apprentissage. Chaque profil fournit un vocabulaire commun, c'est-à-dire un ensemble de termes et de concepts propres à un domaine particulier, sur lesquels les concepteurs et les développeurs d'applications peuvent s'appuyer pour décrire de façon uniforme les interactions des utilisateurs. L'annotation d'un document, la lecture d'une vidéo, la réalisation d'un questionnaire ou la notation d'un devoir sont quelques exemples des nombreuses activités ou événements que les profils d'IMS Caliper tentent de décrire ; la liste exhaustive des profils disponibles peut être consultée en ligne [7].

Les données xAPI ou IMS Caliper collectées sont ensuite enregistrées dans des systèmes appelés Learning Record Store, qui font l'objet de la section suivante.

Comme on peut le constater, la préparation du jeu de données est une étape incontournable des Learning Analytics. Il existe un certain nombre de propositions concernant le type d'informations à modéliser (modèle de l'utilisateur, modèle d'activité…). Ces propositions servent souvent de guide dans les prototypes construits, afin de faire une sélection des informations importantes. En parallèle avec ces propositions, les outils informatiques nécessitent de prendre en compte un format de stockage des données. Pour cela, plusieurs options sont possibles mais le plus souvent, l'utilisateur doit convertir ses données dans un format un minimum standardisé. Des outils comme kTBS ou UTL permettent l'importation de traces dans différents formats (CSV, XML, JSON) ; d'autres, comme UnderTracks, se basent sur des systèmes de gestion de bases de données classiques (SQL). Les logiciels de statistiques plus généralistes fonctionnent aussi avec l'importation de formats standards.

Enfin, des initiatives autour de la norme xAPI, exprimée en JSON, ont été proposées (comme CSVtoXAPI), qui facilitent la conversion vers xAPI. Toutefois, chacun de ces outils nécessite l'identification des informations et la mise en forme dans les formats standards (CSV, XML, JSON). Pour réaliser la préparation des données, le logiciel professionnel Talend aide par exemple à l'extraction et la conversion des informations, avec des processus plus ou moins automatisés.

STOCKER LES DONNÉES

Les données collectées, pour être analysées par la suite, peuvent être stockées au sein des systèmes d'apprentissage ou dans des dépôts de données externes. Les travaux actuels de la communauté, francophone et internationale, vise à proposer des dépôts de données spécifiques aux données d'apprentissage.

Un **LRS (Learning Record Store)** est un composant essentiel de l'interopérabilité des activités d'apprentissage. C'est le dépôt dans lequel sont stockées les traces d'apprentissage, ici appelées *statements*, et il est le garant du respect du standard choisi, comme xAPI ou Caliper. Ces *statements* peuvent provenir de différents systèmes s'ils sont autorisés sur le LRS.

6. http://xapi.vocab.pub/
7. www.imsglobal.org/caliper-analytics-v11-profiles-summaries

Ces standards étant ouverts à l'extension, il est nécessaire de convenir des vocabulaires à respecter entre tous les systèmes nourrissant et interrogeant le LRS, pour obtenir un minimum d'interopérabilité entre ces systèmes.

Les *statements* stockés dans un LRS ont vocation à être accessibles et interrogeables via des API, grâce à des outils tiers autorisés tels que des LMS (Learning Management System) ou des outils d'analyse et de fouille de données. Du fait de ce mode de fonctionnement, le LRS est souvent utilisé de façon asynchrone, pour produire des analyses, bilans, statistiques, voire de la recommandation.

Mais certains exploitent aussi le LRS pour déclencher directement des actions en fonction des *statements* observés, ce qui permet de mettre le LRS et les *statements* au cœur des fonctions de tous les outils gravitant autour.

Un LRS peut également avoir la charge d'agréger et d'analyser directement les *statements* pour les rendre plus simplement exploitables dans des tableaux de bord ou des outils d'analyse de traces de plus haut niveau.

Aujourd'hui, il existe des solutions de LRS en mode SaaS [8] qui permettent une mise en place très rapide, avec peu d'expertise informatique. Il convient toutefois de garder à l'esprit que les *statements* intègrent souvent dans leur contenu l'identité de l'utilisateur, ce qui pose d'évidentes questions de confidentialité lorsque le LRS est hébergé chez un partenaire. Mais ceci reste vrai pour des versions On-Premise [9] : en cas d'intrusion, des données à caractère personnel peuvent être exposées et exploitées. Différentes solutions permettent de s'en prémunir, comme l'anonymisation systématique des *statements* émis.

ANALYSER LES JEUX DE TRACES

De nombreux travaux se concentrent sur l'étape d'analyse à proprement parler, en suivant des approches différentes et en visant des publics différents.

Outils généralistes de statistique

Une première approche consiste à utiliser des outils généralistes de statistique (R, SAS, Stata, SPSS, SPAD, etc.) pour la mise en œuvre de méthodes couramment utilisées dans le champs des sciences humaines, autrement dit des outils non spécifiques de l'analyse de données de type traces. Ces outils permettent principalement d'analyser des données tabulaires, mais certains d'entre eux donnent également la possibilité d'analyser des données séquentielles, ce qui requiert alors l'utilisation de méthodes spécifiques de ce format. Ces outils se distinguent d'outils plus orientés, tels les logiciels conçus prioritairement pour les données issues de sondages (ergonomie facilitant la lecture des variables à choix multiples, utilisation de poids de sondage, de sous-populations, etc.), ou encore des logiciels dédiés à l'analyse statistique de corpus textuels (basés sur la fréquence de cooccurrence de mots ou sur la fréquence de champs sémantiques), etc.

Avec les outils généralistes, pour les analyses statistiques classiques, les traces recueillies doivent se présenter au format « standard » de ces outils, à savoir sous forme de tables de données croisant en ligne des individus statistiques et en colonne des variables (aussi appelées « descripteurs, caractéristiques »). Dès lors, que ce soit directement, si le format convient, ou après transformation (prétraitements avant l'importation des données), ces outils permettent d'appliquer un ensemble de méthodes de traitement (recodages, création de nouvelles variables, etc.) et de méthodes d'analyse.

8. Le logiciel en tant que service, ou Software as a Service (SaaS), est un modèle d'exploitation commerciale des logiciels dans lequel ceux-ci sont installés sur des serveurs distants plutôt que sur la machine de l'utilisateur. Les clients ne paient pas de licence d'utilisation pour une version, mais utilisent librement le service en ligne ou, plus généralement, payent un abonnement.
9. À l'inverse des solution SaaS, les logiciels en version On-Premise sont installés directement sur la machine de l'utilisateur.

Parmi les méthodes d'analyse, il s'agit principalement :

> de **décrire** les caractéristiques de l'ensemble des individus statistiques ou de sous-populations en résumant chaque variable isolément (tris à plat, moyenne, médiane…), deux variables simultanément (corrélation, tri-croisés) ou encore trois variables et plus, simultanément. Dans ce dernier cas, les analyses factorielles de type analyse en composantes principales (variables quantitatives) ou analyse factorielle des correspondances (variables qualitatives) visent à restituer le meilleur résumé, notamment visuel, des proximités entre variables d'une part, et entre individus d'autre part ;

> de **comparer** des sous-populations (tests de comparaison de moyennes, de comparaison de distributions, etc.) ;

> de **classer** les individus au regard d'un ensemble de variables par classification supervisée (analyse factorielle discriminante, arbres de décision…), le principe étant de définir, à partir d'un échantillon d'apprentissage, les règles discriminant au mieux les individus de sorte de pouvoir classer de nouveaux individus, ou bien de regrouper les individus par classification non supervisée (classification ascendante hiérarchique, méthode des centres mobiles…), afin de constituer des groupes homogènes, appelés « profils ou classes », à l'aide d'une distance calculée entre individus (distance tenant compte des différentes variables de l'analyse). Une fois les profils obtenus, l'interprétation du résultat passe par une synthèse, produite par l'analyste, des caractéristiques qui confèrent une homogénéité interne aux profils et, inversement, une hétérogénéité entre profils ;

> de **modéliser** une variable au regard d'un ensemble d'autres variables. Les modèles de régression se déclinent en différentes familles de modèles répondant à différents contextes d'analyse (selon le type et la distribution des variables). Parmi les modèles courants, on peut citer l'analyse de la variance (ANOVA), qui vise à déterminer si une ou plusieurs variables qualitatives (on parle souvent de « facteurs » ou « conditions » dans le cas d'expérimentations) sont influentes sur une variable quantitative.

Parmi les méthodes visant à modéliser une caractéristique, on peut distinguer deux objectifs, non exclusifs l'un de l'autre.

> Expliquer : les modèles de régression linéaire multiple, par exemple, permettent de déterminer, pour une caractéristique cible (à expliquer), quelles sont les caractéristiques explicatives, ou autrement dit influentes, sur les valeurs de la caractéristique cible, et celles qui ne le sont pas ; d'évaluer l'effet d'une caractéristique, toutes autres caractéristiques étant égales par ailleurs, et de définir le poids de chacune des caractéristiques influentes du modèle. Pour aller plus loin, afin de tenir compte du contexte dans lequel se trouvent les individus (par exemple la classe, l'établissement ou un autre contexte d'apprentissage), les modèles multiniveaux permettent d'intégrer à la fois des variables descriptives du contexte (niveau 2) et des variables descriptives des individus (niveau 1), afin de modéliser une variable mesurée sur les individus (niveau 1).

> Prédire : les arbres de décision, parmi les méthodes d'apprentissage supervisé, donnent hiérarchiquement les descripteurs les plus discriminants d'une variable cible, le résultat étant produit sous forme schématique d'arborescence facilitant la lecture du résultat. Ces méthodes permettent notamment de prévoir la valeur de la caractéristique cible pour un nouvel individu dont on connaît les valeurs des différents descripteurs.

Lorsque les données recueillies par les outils d'apprentissage retracent l'activité **séquentiellement,** c'est-à-dire sous la forme d'une **suite ordonnée d'actions**, certains outils généralistes de statistique permettent de déployer des méthodes, en particulier les méthodes d'appariement optimal [Lesnard *et al.*, 2006], visant à classer les séquences, et par extension les individus, en fonction de la suite d'éléments qui composent les séquences. Il s'agit d'une méthode de classification non supervisée utilisant une distance entre séquences deux à deux. Cette distance est établie en paramétrant le coût d'une insertion, d'une suppression et d'une substitution, afin de calculer pour chaque paire de séquences le coût total des modifications nécessaires pour rendre les deux séquences identiques. Une fois les profils obtenus, l'homogénéité des séquences regroupées dans un

profil peut ensuite être visualisée à l'aide de chronogrammes (ou « Time Lines ») et décrite au regard des actions les plus fréquentes, des patterns d'actions, etc.

Le format séquentiel des données n'exclut cependant pas le passage vers des données tabulaires calculées à partir des séquences (on parlera alors de « données agrégées ») afin d'obtenir des descripteurs de celles-ci (le nombre d'actions d'un type donné, la durée des séquences…), puis l'application des méthodes statistiques classiques énoncées ci-dessus.

Concernant la démarche globale de l'analyste, elle peut avoir une visée exploratoire et/ou confirmatoire, à savoir une découverte préliminaire de la distribution des données pour procéder ensuite à leur description multidimensionnelle (par exemple, la recherche de profils sans modèle a priori), et/ou une mise à l'épreuve de modèles théoriques ou d'hypothèses de recherche. Pour une partie des méthodes statistiques, elles requièrent une mise en œuvre par des spécialistes.

Pragmatiquement, même si un ensemble de calculs peut être opéré à l'aide de scripts, cette approche se distingue de celle utilisant des algorithmes automatiques car l'utilisation des outils généralistes implique une progression pas à pas dans l'analyse. En effet, les résultats obtenus au fil de l'analyse ont une incidence sur les choix suivants de méthodes, choix de paramétrages et de spécification des modèles. Les deux approches répondent à des objectifs différents.

Algorithmes automatiques

Une autre approche consiste donc à utiliser des algorithmes automatiques d'analyse de traces, afin d'obtenir des informations qui sont soit exploitées automatiquement par des systèmes, soit affichées aux différents acteurs.

Ainsi, Laalys [Muratet *et al.*, 2016] est un outil d'analyse reposant sur un réseau de Pétri, qui permet d'associer des étiquettes pédagogiques aux actions des élèves et calcule un score à partir de ces étiquettes. Ces dernières renseignent les enseignants sur le comportement d'un élève. Le principe de cet algorithme d'étiquetage est de fournir des informations sémantiques en caractérisant les écarts détectés entre la résolution de l'élève et celles préconisées par les enseignants. Le résultat de l'analyse est alors affiché aux enseignants via une interface graphique.

Autre exemple, T-Store [Zarka *et al.*, 2013] est un système de gestion des traces (SBT) qui gère le stockage, la transformation et l'exploitation des traces collectées par des applications externes. Pour exploiter les traces, des transformations sont utilisées. T-Store fournit des fonctions de transformation prédéfinies, ainsi qu'une transformation personnalisée basée sur les automates à états finis. Ces transformations permettent d'extraire des connaissances qui sont ensuite exploitées pour fournir de l'assistance aux utilisateurs finaux, et notamment des recommandations.

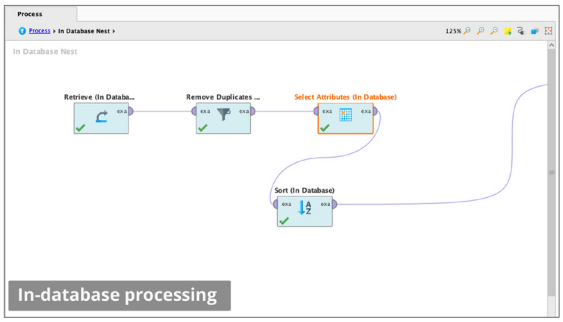
D'autres travaux utilisent des algorithmes automatiques, mais en laissant l'analyste intervenir dans le paramétrage de l'algorithme. Ainsi, DisKit [Fuchs, 2018] permet d'extraire des connaissances présentes dans les traces via l'utilisation de l'algorithme dmt4sp [Mannila *et al.*, 1997]. Cet outil permet d'extraire des motifs décrivant des séries d'événements selon des contraintes, pouvant être ajustées par l'analyste, concernant le nombre minimum d'occurrences des motifs, sur leur taille…

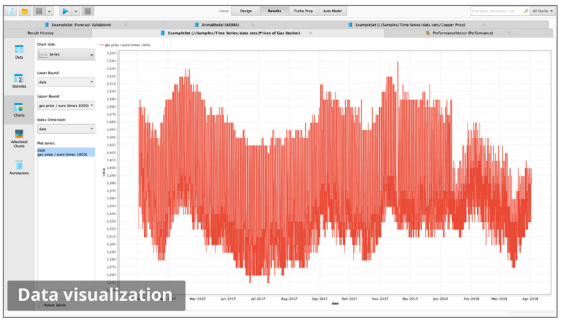
Les outils d'analyse des données pédagogiques proposés par les professionnels de l'éducation sont souvent intégrés dans les solutions d'apprentissage proposées. Nous présentons ici deux outils dédiés à l'analyse des données, outils externes aux plateformes d'apprentissage.

> R [10] est un logiciel de statistique libre. Plus précisément, il s'agit d'un langage et d'un environnement informatiques de calcul statistique. Il est couramment utilisé dans le domaine de la recherche, et plus globalement par les statisticiens. Comme tout logiciel d'analyse statistique, il permet de traiter des fichiers de données (importer, exporter, fusionner des fichiers de données, etc.), de traiter des données (création, recodage, fusion de variables, etc.), d'appliquer des méthodes d'analyses (statistique descriptive, inférentielle, modélisation, text-mining, etc.) et de produire des graphiques (diagrammes, histogrammes, nuages de points, plans factoriels...). L'outil étant enrichi par un grand nombre de contributeurs, il permet de mettre en œuvre un très grand nombre de méthodes d'analyses. Ces méthodes sont disponibles au travers de librairies : huit sont disponibles au moment de l'installation, et de très nombreuses autres sont téléchargeables (plus de 13 000 packages), donnant accès à des développements récents d'analyses. Son utilisation passe par la succession d'opérations et d'appels de fonctions en lignes de commande, et donne la possibilité de créer des scripts pour l'automatisation de calculs. Ainsi les fonctions, les méthodes et les graphiques sont largement paramétrables. En ce sens, R se différencie des outils basés sur le « clic-boutons ». Pour aller plus loin dans l'interopérabilité, l'environnement permet l'appel de code écrit dans les langages C, C++ et Fortran, ainsi que l'appel d'objet R à l'aide du langage C.

> RapidMiner [11] est à la fois un logiciel Open Source gratuit, et un produit commercial, destiné à la recherche d'informations dans des données de type textes ou images. Contrairement à R, tous les outils sont proposés via des interfaces graphiques. RapidMiner propose des fonctionnalités de Data Mining et de Machine Learning, dont la préparation de données (chargement, transformation, prétraitement), la visualisation des données (cf. Fig. 4), leur modélisation, l'apprentissage automatique, l'exploration de texte et l'analyse prédictive. Les processus d'analyse des données peuvent être construits à partir d'opérateurs décrits dans des fichiers XML qui se combinent les uns aux autres, au sein d'une interface graphique (cf. Fig. 4).

Figure 4. RapidMiner





Langage dédié : du langage informatique à la langue naturelle

Une autre approche consiste à utiliser un langage dédié à la manipulation des traces. Ainsi, l'environnement UTL (Usage Tracking Language) [Iksal, 2011 ; 2012] a été conçu pour la conception et l'opérationnalisation d'indicateurs prescrits. Il est nécessaire dans ce contexte de savoir ce que l'on souhaite observer et pour quelle raison. Ensuite, cet environnement permet la description des données d'analyse ainsi que de leurs règles de calcul dans un format indépendant de toute plateforme d'apprentissage. Cette solution facilite la réutilisation des descriptions dès lors que les éléments sur lesquels sont basées les analyses existent dans les traces importées. UTL est composé d'un éditeur web pour les données, d'un calculateur pour l'opérationnalisation ainsi que de connecteurs permettant l'élaboration de tableaux de bord.

kTBS (*kernel for Trace-Based Systems*) [Champin *et al.*, 2013] est une implémentation de référence Open Source d'un système à base de traces modélisées (SBTm). La notion centrale des SBTm est celle de trace modélisée, définie comme une liste d'éléments observés, appelés « obsels ». Chaque obsel est décrit par un type, un ensemble d'attributs, et deux estampilles

10. De nombreux ouvrages sur R ont été publiés et autant de tutoriels en accès libre permettent de découvrir et/ou approfondir le langage R. Site du projet : www.r-project.org/
11. https://rapidminer.com/

temporelles début et fin, délimitant l'intervalle durant lequel cet obsel a pu être observé. Chaque trace est associée à un modèle de traces, qui spécifie les types d'obsels que la trace peut contenir, ainsi que les attributs de chaque type d'obsels. Ainsi, le modèle de traces permet d'expliciter la structure et la sémantique sous-jacente d'une trace. Cette connaissance est capitalisable, puisque plusieurs traces décrivant des activités similaires peuvent faire référence au même modèle. Le kTBS utilise le modèle de données RDF [12], qui offre la flexibilité nécessaire pour représenter les traces modélisées selon divers modèles de traces. kTBS fournit un ensemble d'opérateurs de transformation, depuis les simples filtres jusqu'à des réécritures complexes spécifiées en SPARQL [13]. Il permet également la définition d'opérateurs personnalisés. Tous ces opérateurs sont à coder en utilisant le langage RDF.

Une autre approche consiste à exploiter la langue naturelle pour interroger les données et ainsi permettre à des analyses non-informaticiens d'interroger le système à base de traces. Ainsi, SPARE-LNC (pour SPArql REquest en langage naturel contrôlé) [Kong Win Chang *et al.*, 2015] est un langage dont l'objectif est de proposer une alternative au SPARQL pour interroger les traces stockées dans le système à base de traces kTBS. Ce langage est guidé par une grammaire algébrique ayant comme base soit la langue française, soit la langue anglaise. L'entrée de l'utilisateur est analysée dans son intégralité via l'ensemble de règles composant la grammaire. L'ordre des règles n'est pas absolu, permettant certaines libertés. On peut différencier dans la définition de la grammaire deux groupes de règles. Le premier groupe permet la création de phrases exprimant des conditions sur des éléments à récupérer dans les traces. Le deuxième permet la définition de phrases gérant les éléments récupérés, par exemple en opérant des calculs ou des sélections sur ce qui est récupéré. Le langage utilisé pour requêter la base est alors composé d'un ensemble de ces phrases qui satisfont la grammaire proposée. Chacune de ces phrases correspond ainsi à une sous-requête du langage, formant un texte qui décrit les données à récupérer en énonçant un ensemble de contraintes et un ensemble d'actions à réaliser. Cet ensemble de phrases est ensuite traduit automatiquement en SPARQL pour interroger la base de traces.

Outils graphiques

Une dernière approche consiste à fournir des **outils de manipulation graphique** des traces. Ainsi, DDART [Michel *et al.*, 2017] permet aux étudiants et aux enseignants utilisant Moodle de combiner des traces hétérogènes puis, via une interface graphique, de concevoir des indicateurs en choisissant les entités (les éléments sur lesquels se font les calculs) ; les types de données (fréquence, intervalle de temps, contenu, description) liées aux entités ; les types de calculs ; et enfin, les types de visualisations. La spécification de tout nouvel indicateur provoque un affichage direct du visuel de l'indicateur. Ce calcul dynamique permet à l'utilisateur d'adapter facilement la conception de l'indicateur pour atteindre la forme voulue. D'autres travaux se concentrent sur la proposition de tableaux de bord dynamiques.

SBT-IM (système à base de traces pour le calcul d'indicateurs sur la plateforme Moodle) [Djouad *et al.*, 2011] est un système à base de traces (SBT) spécifique dédié à la définition d'indicateurs d'activité collaboratifs et individuels dans les activités de la plateforme collaborative Moodle [14]. SBT-IM permet de définir un indicateur et de choisir une visualisation en présentant à l'auteur un système de renseignement d'informations progressif, des informations générales aux informations détaillées, sur le calcul de l'indicateur en cours de définition. Le parcours des traces se fait via une suite de tableaux affichant les données filtrées par opérations successives.

Abstract [Georgeon *et al.*, 2012] est une application web proposant des outils graphiques de manipulation des traces pour l'analyse de l'activité humaine en temps réel. L'objectif est de pouvoir mener une analyse de traces d'activité pour la modélisation cognitive de l'utilisateur. L'outil propose différentes fonctionnalités : un éditeur d'ontologie pour spécifier les modèles

12. www.w3.org/RDF/
13. www.w3.org/TR/rdf-sparql-query/
14. https://moodle.org/?lang=fr_fr

des différentes traces, un éditeur de transformation pour spécifier différentes règles de transformations applicables aux traces, un moteur de transformation, un système de visualisation des traces et du résultat des transformations et, enfin, un outil de requêtes pour rechercher des occurrences de schéma dans les traces.

kTBS4LA (kTBS for Learning Analytics) est une surcouche de kTBS permettant l'interprétation des traces et la manipulation des concepts manipulés dans le kTBS, sans nécessité de connaissance en langage de programmation. Cette application web capitalise les travaux proposés par Samotraces [15] et SamotraceMe [Derbel *et al.*, 2015] sur la visualisation de traces multi-vues, multi-échelles et multi-sources, réexploite la collection de composants web de Taaabs [16] pour manipuler et visualiser graphiquement les traces contenues dans un kTBS, et intègre le langage SPARE-LNC [Kong Win Chang *et al.*, 2015]. Pour analyser les traces d'interaction issues d'un EIAH, un utilisateur de kTBS4LA procédera en plusieurs étapes. Il doit tout d'abord importer les traces issues de cet EIAH dans kTBS4LA, ce qui lui permet également de définir le modèle des traces pour une situation correspondante à l'usage de cet EIAH. Il peut ensuite explorer les traces importées à l'aide de différents outils de visualisation, et créer de nouvelles traces modélisées permettant de mieux comprendre l'activité des apprenants.

D3KODE [Champalle *et al.*, 2016] (*Define, Discover and Disseminate Knowledge from Observation to Develop Expertise*) est une plateforme web adossée à kTBS dans un but de réutilisation et de partage de connaissances d'analyse de traces numériques. L'interface de D3KODE s'adresse à des utilisateurs non informaticiens et a donc été conçue de manière à faciliter l'analyse de traces numériques en dégageant l'utilisateur des notions techniques du kTBS. Les fonctionnalités du prototype permettent aux utilisateurs d'importer des données de bas niveaux et de les transformer en informations de plus hauts niveaux. Le résultat est présenté sous la forme d'une synthèse visuelle sur plusieurs niveaux d'abstraction, où chaque « observation » d'un niveau N est reliée à ses origines dans le niveau N-1. Les niveaux sont construits via des règles créées au travers d'une interface dédiée. Les règles créées sont réutilisables, et donc partageables, entre utilisateurs. La synthèse visuelle est interactive et permet à un analyste d'accéder aux informations de chaque niveau, règle et observation en cliquant sur le point qui l'intéresse.

UnderTracks [Bouhineau *et al.*, 2013] est un outil d'assistance à la création de processus d'analyse qui permet de guider les choix d'opérateurs en fonction de vues sur les données. Il permet l'import et le stockage des traces, la gestion des opérateurs d'analyse ainsi que la construction visuelle des processus d'analyse et la représentation graphique des résultats. Toutefois, les données produites ne sont pas persistantes, et nécessitent donc d'être recalculées chaque fois que l'on souhaite les réexploiter.

VISUALISER LES RÉSULTATS DE L'ANALYSE

Présenté en introduction de ce rapport, le cycle d'analyse de traces passe par plusieurs étapes : collecte, analyse, visualisation des traces et des résultats produits par l'analyse [Fayyad *et al.*, 1996 ; Clow, 2012 ; Stamper *et al.*, 2011]. La majeure partie des outils cités dans cet état de l'art permet de parcourir une grande partie de ce cycle. Tous ne possèdent cependant pas une « visualisation graphique » qui se prête à l'analyse des traces et/ou à la compréhension des comportements des apprenants dans les EIAHs.

Les outils retenus dans cette section, possèdent une visualisation graphique des traces « intégrée » (et non externe) : Abstract, D3kode, DDART, kTBS4LA, Travis, SBT-IM, Tatiana, UnderTracks, Lab4CE, Emoda et Transmute. Nous avons donc fait le choix d'exclure de cette section tous les outils qui ne possèdent pas d'aspect visualisation graphique autre que textuelle, tels Laalys ou UTL par exemple.

15. Le projet Samotraces : http://sourceforge.net/projects/samotraces/
16. Le projet TAAABS : https://projet.liris.cnrs.fr/sbt-dev/tbs/doku.php/tools:taaabs

En matière de visualisation graphique, parmi les outils cités dans cette section, il est possible de distinguer deux catégories d'applications :

> les outils d'exploration « généralistes » de traces numériques non exclusivement dédiés aux Learning Analytics, mais dont l'utilisation est aussi possible et avérée dans d'autres contextes ;

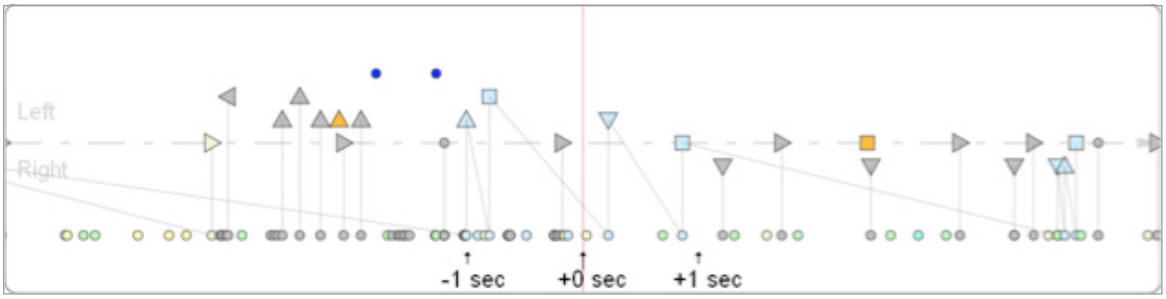
> les outil de type « tableau de bord », conçus spécifiquement pour les traces numériques provenant d'EIAHs. Ils possèdent un panel d'indicateurs graphiques élaborés pour « observer » l'activité suivant plusieurs orientations. Ils ciblent les apprenants et/ou les enseignants/tuteurs.

Outils d'exploration généralistes

Parmi les formes de visualisation graphique les plus représentatives et les plus courantes des outils « généralistes », la « Time Line [17] » est la plus récurrente. Il en existe plusieurs variantes, parfois associées avec d'autres visualisations graphiques plus classiques, tels des histogrammes ou des camemberts.

Abstract [Georgeon *et al.*, 2012] propose une visualisation unique d'une seule trace numérique dans une optique d'analyse et de découverte de l'activité. La trace représentée est soit une trace première (avant transformation), soit une trace abstraite (transformée) avec des liens « origines » entre ses observés. Abstract est tourné en direction de l'analyste et n'est pas dédiée exclusivement aux traces d'EIAHs.

Figure 5. Visualisation d'Abstract



Dans la lignée d'Abstract, D3KODE possède une Time Line plus riche [Champalle *et al.*, 2016]. La représentation est interactive et plusieurs niveaux d'abstraction peuvent être représentés. D3KODE est tourné sur la capitalisation et le partage des connaissances d'analyses ; la traçabilité et la réutilisation des connaissances de transformations (abstractions) mobilisées sont plus développées. Cette traçabilité permet aux utilisateurs de comprendre comment les niveaux d'abstractions et leurs observés sont construits.

Transmute [Barazzutti *et al.*, 2016] est une interface graphique interactive et personnalisable conçue pour assister l'interprétation de traces et la découverte de connaissances. La visualisation des traces se présente sous la forme d'une Time Line assez classique (dans la lignée d'Abstract et de D3KODE), qui permet de visualiser très rapidement la trace « en cours » et la trace « transformée » par l'analyste. La figure suivante présente l'interface de Transmute couplée à Diskit, une application orientée fouille de données.

17. Une Time Line est une représentation symbolique des événements sur un axe temporel horizontal qui permet d'explorer la dimension temporelle des données sur une période de temps qui peut être ou non paramétrée (min, heure, jour, mois…).

Figure 6. Interface de Transmute

kTBS4LA est une plateforme web d'analyse de traces dont l'objectif est de faciliter la manipulation graphique des données collectées ainsi que leur visualisation rapide. La vue principale des données est une Time Line dont l'organisation graphique des observés en « colonne » permet une vue « en profondeur » de l'activité, tout en conservant un aperçu global sur une période de temps importante. Pour compléter l'analyse, kTBS4LA propose aussi l'emploi d'autres indicateurs, tels des histogrammes et des camemberts [Casado *et al.*, 2017].

Tatiana [Dyke *et al.*, 2010], outil d'analyse graphique de traces conçu pour des chercheurs en sociologie, exploite aussi le principe de Time Line pour présenter les interactions médiatisées et assister leur analyse.

Figure 7. Chronotime dans l'outil Tatiana

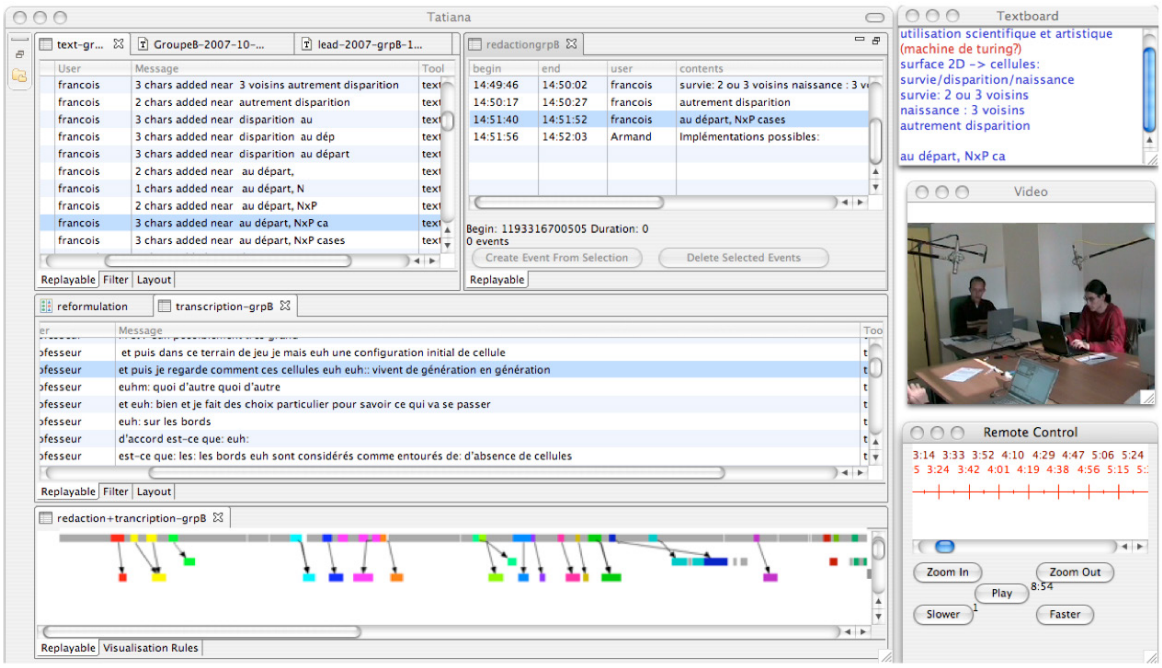
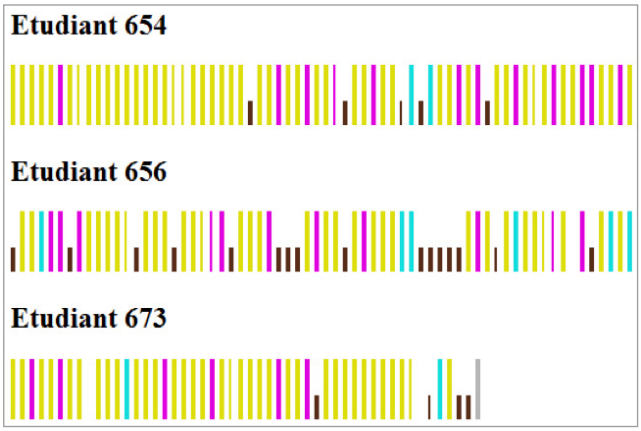


Figure 8. Comparaison de plusieurs traces dans UnderTracks



UnderTracks [Bouhineau *et al.*, 2013] est un outil d'analyse « généraliste » basé sur la plateforme Orange. Dédié à l'origine à la capitalisation de processus d'analyse de traces d'apprenants, il peut être utilisé pour d'autres contextes. UnderTracks propose un panel extensible de visualisations graphiques, dont un pattern de type Time Line offrant notamment la possibilité de comparer plusieurs traces d'activités.

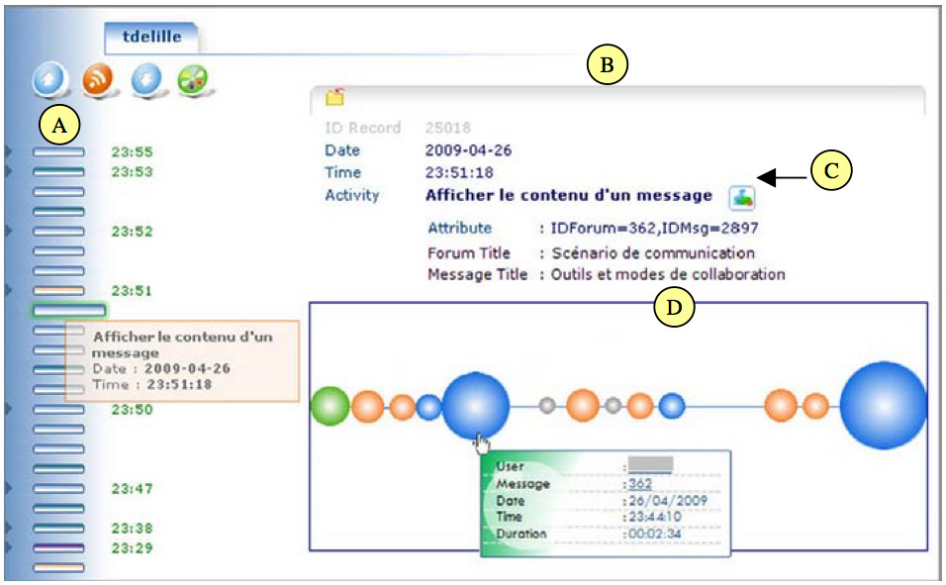
Outils type « tableau de bord » dédiés aux EIAHs

Les outils de types « tableaux de bord » sont dédiés à l'analyse de traces numériques provenant de plateformes d'apprentissage spécifiques. Ils possèdent un panel d'indicateurs graphiques élaborés pour « observer » l'activité suivant plusieurs facettes. Les utilisateurs cibles peuvent être les apprenants, dans un but réflexif, et/ou les tuteurs dans un but de compréhension et/ou de régulation de l'activité d'apprentissage.

Travis [May *et al.*, 2011] est un outil d'analyse de forums dédié à la plateforme Moodle. Il utilise les traces numériques de Moodle afin de proposer à l'enseignant une vue d'ensemble interactive des échanges et consultations du forum par les élèves. Travis propose plusieurs indicateurs :

> une Time Line conçue pour faciliter l'exploration des échanges du forum, avec des bulles plus ou moins importantes selon le nombre de messages échangés dans un fil de discussion ;
> plusieurs indicateurs statistiques pour visualiser l'activité individuelle des étudiants sur le forum.

Figure 9. Interface de Travis et la Time Line des échanges dans un forum



SBT-IM [Djouad *et al.*, 2011] permet de collecter des traces à partir des plateformes d'apprentissage Moodle dans le but de créer des indicateurs pour analyser les traces d'activités des étudiants sur la plateforme. SBT-IM permet la création et la réutilisation d'indicateurs ainsi que plusieurs visualisations graphiques interactives de l'activité. L'activité des étudiants peut par exemple être visualisée chronologiquement, mais aussi sous forme d'indicateurs plus classiques, de type histogramme ou camembert.

Lab4ce [Broisin *et al.*, 2017a ; Broisin *et al.*, 2017b] est une plateforme web munie d'un ensemble de fonctionnalités supports à l'apprentissage et d'un outil de visualisation en direction du tuteur et des apprenants. Lab4ce propose plusieurs types de visualisations graphiques, principalement réflexives :

> **comparaison sociale** : un ensemble de barres de progression qui reflète le niveau de performance des apprenants par un code couleur ;
> **réflexion a posteriori de l'activité d'apprentissage** : une Time Line permettant aux apprenants d'analyser en détail leurs propres actions ainsi que celles de leurs pairs. Pour chaque ressource sélectionnée, une chronologie des instructions exécutées est représentée. Chaque nœud de la chronologie représente une instruction ; celui-ci est coloré selon sa justesse technique, tandis que le détail de l'instruction apparaît dans un simili-terminal lorsque le curseur est positionné sur le nœud correspondant ;
> **analyse des stratégies mises en œuvre :** un indicateur de type Time Line permet aux apprenants de visualiser l'évolution des stratégies qu'ils mettent en œuvre tout au long de la réalisation des activités d'apprentissage.

EMODA [Ez-Zaouia, Lavoué, 2017] est un tableau de bord conçu pour aider les enseignants engagés dans des formations en ligne pour l'apprentissage des langues. La particularité de l'approche EMODA réside dans son ambition de détecter les émotions des apprenants pour faciliter le lien entre les enseignants et les apprenants à distance. EMODA adopte une approche multimodale et considère quatre sources de données : audio, vidéo, self-report et traces d'interaction. Les données audio et vidéo correspondent aux échanges de communication enregistrés lors des séances. Les données de self-report sont renseignées par l'apprenant avant et après la séance. Les traces d'interactions sont utilisées pour qualifier les actions liées à des émotions particulières au cours de l'activité. EMODA propose des résultats sous forme d'indicateurs discrets, bi ou multidimensionnels. La visualisation des émotions est transmise à l'enseignant au travers d'un tableau de bord avec différents formats de visualisation : histogrammes, émoticônes, courbes d'évolution temporelle, images significatives de la séance.

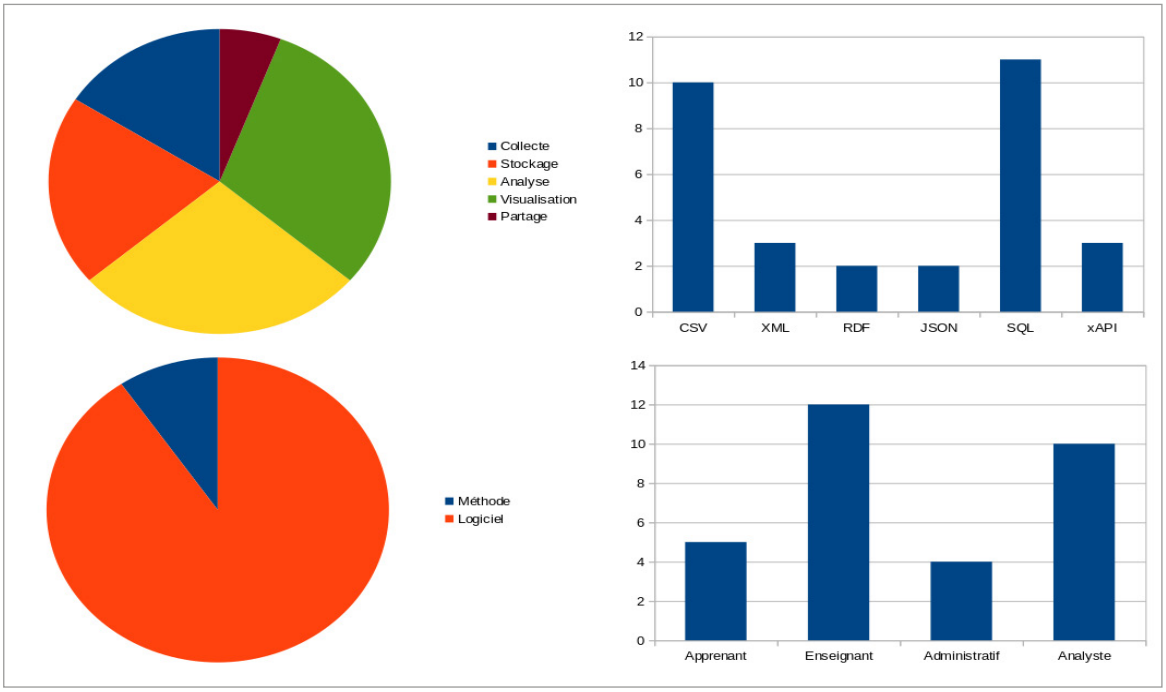
DDART [Michel *et al.*, 2017], pour *Dynamic Dashboard Based on Activity and Reporting Traces*, est à la fois un outil de reporting et de monitoring. L'approche de DDART est centrée apprenant, dans un contexte *Project-Based Learning* (PBL). L'objectif est d'aider les apprenants à collecter, analyser et visualiser les traces significatives de leurs activités par eux-mêmes. DDART peut s'intégrer à un ENT conçu avec Moodle. Son intégration offre, en complément des outils de gestion du travail collaboratif (forum, chat, wiki…) de Moodle, deux outils de planification et de suivi de l'activité des élèves et du groupe :

> le *reporting tool* permet de spécifier différents objectifs à atteindre, de décrire comment l'activité se réalise ou de juger la qualité de la réalisation intrinsèque. Les élèves peuvent ainsi collecter des éléments d'information sur leur activité, gérer leurs rapports d'activité, les modifier a posteriori ou faire des commentaires généraux sur les contextes des activités. Ils peuvent également lire et commenter les rapports des autres élèves du groupe, ce qui est très utile pour que l'élève se situe par rapport au groupe et puisse aussi prendre du recul sur son activité ;
> le *dashboard* offre une vue globale sur des indicateurs de suivi des activités rapportées et des activités réalisées avec les outils de l'ENT. Il permet ainsi de suivre qui travaille sur quoi, combien de temps, quel est le jugement des uns et des autres… Ces indicateurs sont créés dynamiquement par les élèves ou l'enseignant, en fonction des besoins du projet pédagogique.

Synthèse des outils étudiés

Une analyse détaillée (cf. Fig. 10) des outils que nous avons étudiés montre que les outils sont essentiellement des logiciels et que l'analyse des données et leur visualisation représentent plus de 50 % des outils disponibles. Le public visé est en priorité les enseignants et les analystes, ce qui s'explique par le fait que les analyses sont souvent réalisées en fonction du besoin des enseignants d'observer le déroulement de leur situation d'apprentissage. Pour les analystes, cela est dû principalement au fait que de nombreux outils ont été conçus dans le cadre de la recherche et de travaux de chercheurs en analyse de données. Ces analyses souvent complexes, nécessitent au préalable un travail de la part de spécialistes des données. Enfin, concernant les formats de données pris en charge, nous retrouvons les formats classiques des plateformes de formation, c'est-à-dire le stockage en interne dans les bases de données (SQL) et les formats d'import/export, avec notamment le CSV. Ce bilan démontre qu'il reste un travail conséquent à mener sur l'interopérabilité et l'utilisation d'environnements standardisés comme xAPI, le partage de résultats et leur réutilisabilité, ainsi que la prise en compte des autres usagers, comme les apprenants.

Figure 10. Analyse des 21 outils étudiés (outils issus de la recherche française)



Perspectives en matière de modèles et d'outils pour les Learning Analytics

Les travaux menés dans le cadre des Learning Analytics sont assez variés et couvrent toute la chaîne de traitement des données. Toutefois de nombreuses étapes restent encore à mettre en œuvre afin notamment de diffuser, d'exploiter et d'étendre ces résultats. En effet, les travaux menés dans le cadre du projet ANR HUBBLE [Luengo *et al.*, 2019] ont cherché à aider la capitalisation et le partage des processus d'analyse. Au travers de leur outil, il est possible de décrire les processus et aussi d'obtenir toutes les informations afin de réutiliser ceux déjà saisis. Un autre enjeu lié à ce projet consiste à travailler sur l'adaptation des processus d'analyse en fonction des différentes plateformes et des formats de données utilisés. Les étapes de prétraitement et d'intégration étant coûteuses, il est nécessaire de trouver des solutions facilitant cette réutilisation.

La production de Learning Analytics soulève des questions liées à l'exploitation de ces informations. Le premier usage se porte sur des tableaux de bord, et donc des retours graphiques et visuels, mais l'intégration dans les outils pédagogiques sous la forme de tuteurs intelligents ou d'adaptations des plateformes est aussi primordiale. Le retour en force de l'intelligence artificielle permet l'exploitation de données massives en couplant les Learning Analytics et le domaine du Data Mining (Educational Data Mining), ce qui amène les chercheurs à travailler sur l'exploitation automatique mais aussi l'explicabilité des processus et des données obtenues, afin d'améliorer l'appropriation par les usagers (enseignants, apprenants…).

Enfin, se pose aussi la question des entrepôts de données dédiés à l'éducation dans lesquels les chercheurs pourraient éprouver leurs théories ainsi que les outils développés. Ces entrepôts seraient aussi importants pour les usagers eux-mêmes, qui pourraient y déposer leurs données et bénéficier des outils de la recherche adaptés et directement disponibles dans ces entrepôts.

3

LEARNING ANALYTICS : UTILISATIONS ET USAGES À DES FINS SCOLAIRES

Hassina El Kechai

Dans ses travaux, Michel Serres explique que la révolution numérique fait suite aux grandes inventions et révolutions telles que l'écriture et l'imprimerie. Il précise donc que les impacts sur la société, sur les individus, sont tout aussi importants[18]. L'accent est mis sur le rapport au savoir, l'accès à la connaissance, la temporalité, les mutations générées par de nouvelles compétences acquises grâce à l'avènement de nouveaux « outils » : des manuscrits aux livres édités en grande quantité, à l'accès à Internet et tous ses contenus. Dans tous les cas, les individus accèdent désormais à un nouveau type de savoir et bénéficient d'une démocratisation de ce savoir ; il n'y a plus un dépositaire unique des connaissances. Les usagers ayant accès au numérique partagent et échangent des idées et des expériences. Ainsi ces grandes révolutions, chacune à son niveau et à sa manière, ont redéfini la perception du monde des individus, modifié les modes de communication et transformé les rapports sociaux – et Internet plus que n'importe quel autre outil, peut-être.

Si l'on recentre la question du numérique sur le monde de l'éducation, le numérique a redéfini l'accès et le rapport au savoir, et modifié le rôle et la posture de l'enseignant. Parce que le numérique fait partie intégrante de la vie quotidienne, des usages et pratiques des apprenants (en particulier la génération dite « Y » qui fait tant parler d'elle)[19], il représente un enjeu à l'École et fait s'interroger sur la manière d'intégrer ces technologies de la communication et de l'information tant dans les murs de l'institution, avec les infrastructures et les services numériques proposés par l'administration (WiFi gratuit, environnement numérique de travail, plateformes d'enseignement numérique, etc.), que dans les pratiques pédagogiques (et au sein même des formations).

Il est donc entendu que le numérique fait partie intégrante de la vie de la grande majorité des individus : il s'agit d'un fait de société, le taux d'équipement augmente régulièrement, les gens sont de plus en plus connectés. Pour citer quelques chiffres tirés du « Ba romètre du numérique 2019 » [Crédoc, 2019], il ressort que 88 % des foyers français sont connectés à Internet et 95 % des Français possèdent un téléphone mobile. Les objets eux-mêmes sont de plus en plus connectés, le nombre d'applications téléchargées sur smartphone, ordinateur, tablette, est en augmentation constante, les usages s'intensifient chaque année.

18. Michel Serres, « L'innovation et le numérique », conférence, Université Panthéon 1-Sorbonne, 29 janvier 2013.
19. Plusieurs travaux, tels que ceux du GTnum 4 : « Usages et pratiques numériques des jeunes » animé par Pascal Plantard, insistent sur la nécessaire prise en compte du contexte socio-économique et socioculturel des pratiques numériques des jeunes.

Lorsque l'on examine les chiffres du Crédoc sur les habitudes des internautes français, on voit bien la diversité des usages : 62 % des Français ont effectué au moins un achat en ligne en 2019, 60 % sont membres d'au moins un réseau social, 62 % utilisent des applications pour échanger des messages, 51 % pour téléphoner. La navigation web sur mobile repart à la hausse, après une année de stagnation, pour atteindre 68 %… et la liste n'est pas exhaustive. Multipliées par le nombre d'usagers, le nombre de sites visités, de pages consultées, de recherches en ligne, de commentaires laissés sur les sites, d'informations laissées lors de transactions en ligne, etc., les traces volontaires et involontaires représentent une masse de données de l'ordre de plus de 2 trillions d'octets par jour.

La récolte, l'analyse et l'exploitation de ces données massives représentent ainsi désormais l'un des grands enjeux du numérique et sont devenues une véritable économie. Elles ont intéressé, en premier lieu, les entreprises et les grands groupes commerciaux. Ces derniers cherchent ainsi à mieux comprendre les habitudes et les pratiques des usagers, en tant que consommateurs, afin de mieux cibler et optimiser leurs offres commerciales et de prédire des tendances. Mais la récolte et l'analyse des données concernent en réalité tous les secteurs d'activité, de la santé à l'agriculture, en passant par les services des assureurs, etc., et le secteur de l'éducation ne fait pas exception.

En effet, selon une étude menée par le Center for Digital Education auprès d'enseignants du supérieur, les principaux bénéfices de l'analyse des données dans l'éducation seraient [20] :
> le suivi et la prédiction des performances d'un élève (69 %),
> l'augmentation du taux de diplômés (61 %),
> l'ajustement en temps réel des programmes scolaires (47 %),
> la mesure de la performance institutionnelle de l'*établissemen*t (44 %),
> la prévention d'éventuelles failles dans l'administration grâce à l'analyse (22 %).

C'est dans les années 1980 que les ordinateurs ont commencé à faire leur apparition dans les écoles primaires, les collèges et les lycées. L'environnement scolaire a permis à de nombreux élèves de s'initier à l'informatique.

Aujourd'hui, les ordinateurs portables et les tablettes remplacent de plus en plus les feuilles blanches et les stylos dans les salles de classe. Cette numérisation de l'éducation génère un très grand volume de données relatives à l'apprentissage et à l'enseignement. Les entreprises technologiques et les établissements scolaires peuvent dorénavant s'associer pour convertir ces données en pistes à suivre pour développer de meilleures méthodes d'enseignement, de nouveaux programmes scolaires, et pour remédier aux problèmes des élèves en difficulté. C'est ainsi que les Learning Analytics prennent de plus en plus d'essor et présentent un grand intérêt dans les contextes scolaires.

Cette section a pour objectif d'explorer les pratiques enseignantes en matière d'analytique de données d'apprentissage et des outils utilisés, dans le contexte français. Elle est constituée d'une enquête préliminaire ciblant une trentaine d'enseignants et d'une exploration de quelques-uns des outils qu'ils utilisent, dans une optique d'analytique des apprentissages. Elle montre une certaine volonté d'utilisation des Learning Analytics de leur part, en faisant avec « les moyens du bord ».

Pour ce faire, nous avons procédé en deux temps.

Nous avons interrogé un échantillon d'enseignants pour comprendre le rapport qu'ils entretiennent avec le numérique, et pour prendre connaissance des pratiques mises en place pour suivre l'activité et optimiser l'apprentissage de leurs élèves – qui sont la vocation première des Learning Analytics. Nous les avons également interrogés pour savoir s'ils avaient recours à des outils ou des instruments qui leur permettent de pratiquer l'analytique de données d'apprentissage, même si ces outils n'ont pas comme vocation première de faire des Learning Analytics.

20. Source : article sur lebigdata.fr, mai 2016.

Un questionnaire a été élaboré, basé pour une grande partie sur des questions ouvertes. La vocation de ce questionnaire est d'explorer les pratiques enseignantes en matière d'analytique des données d'apprentissage. Une trentaine de réponses nous ont permis de dresser une première liste de pratiques, d'outils et d'instruments.

QUESTIONNAIRE ÉLABORÉ

Le questionnaire est structuré comme suit.

1) Une page d'accueil informe de l'objectif du questionnaire, de l'anonymat et de la confidentialité des réponses données.

2) Les questions sont ensuite réparties en plusieurs parties avec, pour chacune, des objectifs précis.

Pratiques numériques des enseignants : la diffusion de solutions de Learning Analytics dans les établissements scolaires nécessite impérativement la mobilisation de compétences numériques indispensables, ainsi que des pratiques numériques suffisamment élaborées de la part des enseignants. Il s'agit ici, pour nous, d'explorer quelques-unes de ces pratiques numériques.
Pratiques pédagogiques pour la compréhension et le suivi de l'activité des élèves : recueillir et explorer les pratiques habituelles des enseignants interrogés qui leur permettent de suivre l'activité de leurs élèves et leur progression, s'ils utilisent le numérique pour le faire et, si c'est le cas, quels sont les outils mobilisés.
Freins, obstacles ou difficultés rencontrés pour comprendre l'activité des élèves et pour assurer leur suivi : recueillir l'avis des enseignants concernant les éléments qui constituent des obstacles pour avoir une perception suffisante de l'activité des élèves. L'idée est de savoir si ces éléments peuvent être pris en charge par des solutions de Learning Analytics existantes ou à envisager.
Préconisations pour améliorer la compréhension de l'activité et le suivi des élèves : demander aux enseignants ce qu'ils préconisent comme moyens pour assurer un meilleur suivi de leurs élèves et quels sont les besoins qu'ils expriment, compte tenu de leurs pratiques et des difficultés qu'ils rencontrent pour comprendre l'activité des élèves.
Perception des apports des Learning Analytics pour comprendre l'activité et pour le suivi des élèves : avoir un retour sur la représentation que se font les enseignants interrogés des Learning Analytics, sur leurs apports dans leurs pratiques pédagogiques, et cerner l'intérêt pour eux de disposer de données d'apprentissage au travers de projections d'usages.
Perception des limites de l'utilisation des traces et fiabilité qui leur est accordée : évidemment, nous ne pouvons passer outre les questions liées aux limites et à la fiabilité accordée par les enseignants à l'analyse des données d'apprentissage pour appréhender l'activité de leurs élèves, et à sa pertinence, dans le cadre de leur pratique pédagogique.

RÉSULTATS OBTENUS

Nous présentons ci-après une synthèse des résultats obtenus.

Les réponses obtenues émanent d'enseignants exerçant dans des collèges situés dans les villes suivantes : Château-Landon, Fontenay-sous-Bois, Drancy, Créteil, Villeneuve-Saint-Georges, Neuilly-sur-Marne, Cesson, Brioude, Châteaubriant, Saint-Nazaire, Le Perreux.

B.1. Pratiques numériques des enseignants interrogés

Nous avons classé les pratiques numériques des enseignants selon sept catégories illustrées ci-dessous :

Figure 11. Catégories de pratiques numériques

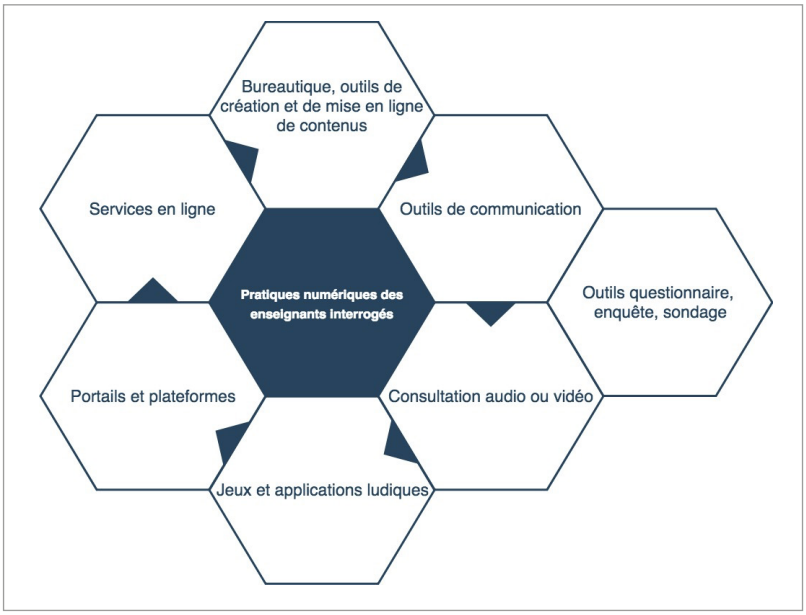


Tableau 7. Catégories de pratiques numériques

Catégorie	Outil utilisé	Finalité
Bureautique, outils de création et de mise en ligne de contenus	Padlet, Google Docs, Framapad	Produire, mettre à disposition et partager des cours.
	Slides.com	Créer des présentations en ligne.
	LibreOffice Writer	Créer des activités.
	Impress	Créer des exposés.
	LearningsApps.org	Créer des applications numériques.
	Piktochart, Emaze, Genial.ly	Créer des fiches de travail, posters, infographies ou diaporamas.
	OpenOffice, LibreOffice	Accompagner les élèves dans la création d'écrits.
Outils de communication	Chats, Forums	Échanger avec les élèves pour les devoirs. Échanger avec le professeur principal.
Consultation audio ou vidéo	Vidéo en ligne Logiciel de montage vidéo et audio Enregistreur audio	Visionner des documentaires. Créer des vidéos. Visionner des vidéos ludiques. Enregistrer des expressions orales en contrôle continu. Visionner des extraits de pièces de théâtre. Visionner des films. Écouter des textes lus. Montages audio et vidéo.
Jeux et applications ludiques	Serious Games	Activités d'apprentissage ludiques.
Portails et plateformes	ENT	Utilisation de l'ENT pour la mise en ligne de documents.
	Moodle	Élaboration de parcours Moodle.
Services en ligne	Google Maps	Essentiellement en cours de géographie.
Outils questionnaire, enquête, sondage	Webquest	Réalisation de sondages et de questionnaires.
	Kahoot!	Création de quiz.

B.2. Pratiques pédagogiques pour la compréhension et le suivi de l'activité des élèves

Les réponses obtenues nous ont permis de collecter essentiellement des informations sur les pratiques d'évaluation, leurs formats et les instruments et outils utilisés pour les réaliser.

Tableau 8. Description des pratiques observées

Pratiques de suivi et d'évaluation	Description de la pratique.
Pratiques d'évaluation	En début de séquence, pour faire le point sur les acquis et les objectifs ; en cours de séquence, pour faire le bilan intermédiaire ; en fin de séquence, pour valider le niveau des compétences acquises → évaluations diagnostique, formative, sommative.
Formats des évaluations	Évaluations selon critères, compétences, connaissances. Évaluation de retours papier. Interrogation papier. Évaluation orale. Évaluation régulière en début et en fin de séquence sous forme d'activités pratiques, d'exposé oral, de compte rendu et de devoir sur table.
Outils et instruments méthodologiques utilisés	Cahier de texte en ligne. Notes et appréciations en ligne. Site Canvas Instructure pour suivre les apprentissages. Pronote. Framapad, Google Docs. Padlet. Cartable en ligne (LMS). Moodle. Grilles définies à l'avance et communiquées aux élèves. ENT. Edpuzzle. Les statistiques de Claroline pour le suivi des élèves sur un cours précis en vue d'accompagnement ou de remédiation. Cahier de textes manuscrit. Scolinfo. Statistiques et résultats d'exercices en ligne. Notes et évaluation des compétences. Un suivi des compétences en lien avec le référentiel suite à différents modes d'évaluation. Suivi des élèves dans un fichier Excel, avec mise en place d'indicateurs colorés pour visualiser les tendances.

B.3. Freins, obstacles ou difficultés rencontrées pour comprendre l'activité des élèves et pour assurer leur suivi

Nous avons classé les difficultés, freins ou obstacles déclarés selon plusieurs catégories, illustrées ci-dessous (Fig. 12) :

Figure 12. Catégories de difficultés, freins, obstacles

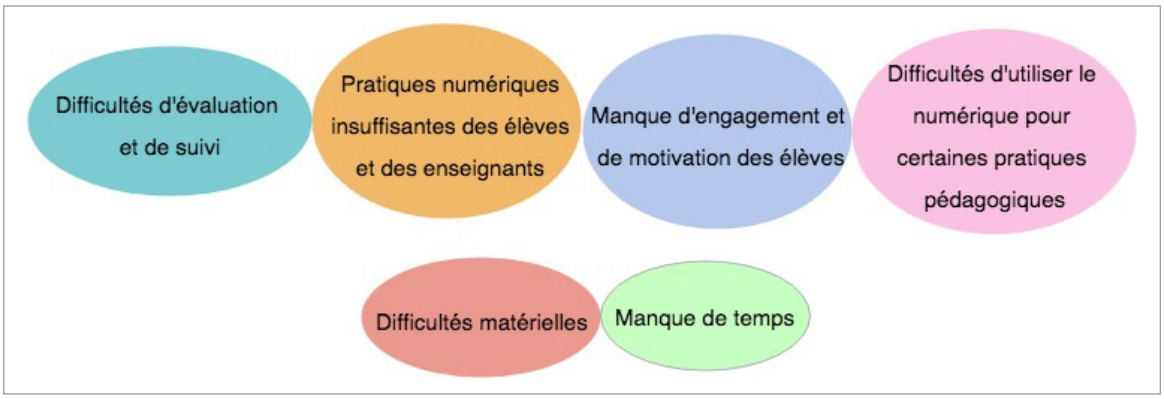


Tableau 9. Description des difficultés

Nature de la difficulté signalée	Description de la difficulté
Difficultés liées à l'évaluation et au suivi	Le côté chronophage des évaluations. Difficulté de cibler les compétences. Le peu de travail réalisé en dehors des cours (difficulté de suivi hors contexte scolaire). Logistique, gestion de classe rendues difficiles par le nombre d'élèves et leurs niveaux de maturité relatifs. La relation entre les notes et la compréhension.
Difficultés matérielles	Pas suffisamment d'ordinateurs en état de fonctionnement. Pas de fibre internet. Difficultés liées à la fracture numérique : pas d'équipements à la maison.
Difficultés liées aux pratiques numériques des collégiens	Manque de compétences numériques chez les collégiens. Temps très lent pour se saisir des outils. Temps lent d'explication et de formation, même pour des tâches simples (allumer/éteindre un ordinateur sous Windows). Difficultés cognitives : compétences de lecture faiblement développées, difficultés pour comprendre le fonctionnement d'un ordinateur…
Difficultés liées aux pratiques numériques des enseignants	Stratégies d'évitement propre au numérique et manque de recul sur leurs propres pratiques en classe avec le numérique. Temps de construction de la différenciation pédagogique avec le numérique. Pas d'usages pour se former au numérique, donc difficultés pour former les élèves.
Difficultés liées au temps	Le manque de temps pour effectuer des reprises/corrections, de suivre tous les élèves. Le manque de temps en raison du découpage du temps par séquences d'une heure. Le nombre élevé d'élèves par classe (manque de temps pour un suivi plus efficient et plus individuel).
Difficultés liées au manque d'engagement et de motivation chez les élèves	Le manque de coopération de certains élèves. Manque de travail des élèves (pas de prise en compte que la note n'est qu'un indicateur). Travail donc uniquement sur les gros coefficients. L'implication personnelle des élèves en dehors de cours.
Difficultés liées à certaines pratiques pédagogiques installées	L'expression orale individuelle permet de mieux cerner l'élève, et permet d'instaurer un dialogue, de la confiance, que le numérique ne peut offrir. Pour les évaluations sur papier, difficile de tout vérifier, on ne peut pas interroger tout le monde. Gérer l'hétérogénéité des classes et proposer un enseignement plus adapté à chacun. Manque de coordination des équipes pédagogiques. Suivi chronophage.

B.4. Préconisations et moyens à mettre en œuvre proposés par les enseignants interrogés pour assurer un meilleur suivi de leurs élèves

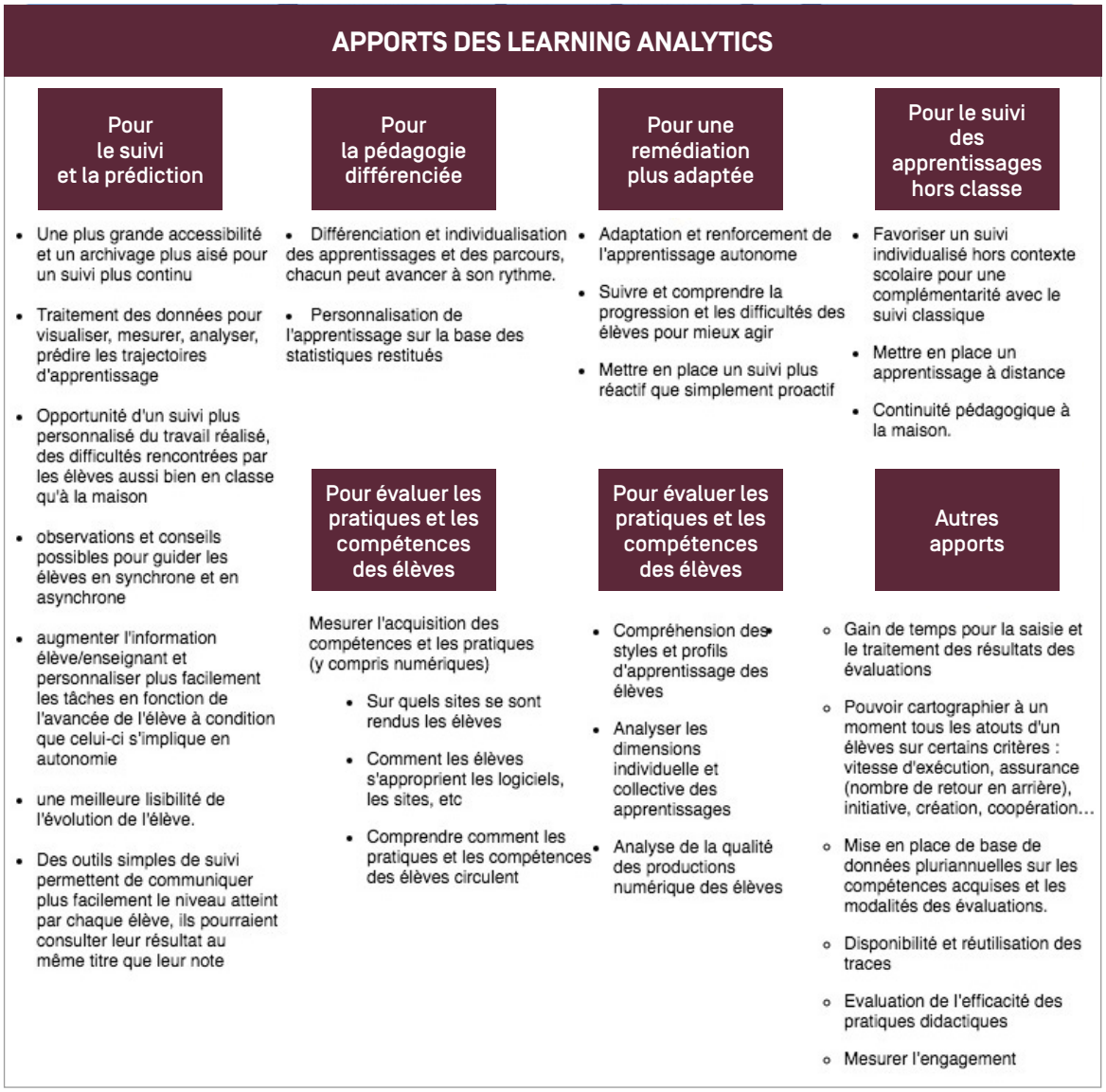
Nous avons classé les préconisations proposées par les enseignants selon deux critères : le type de retour que souhaite avoir l'enseignant pour un meilleur suivi et les actions concrètes ainsi que les moyens nécessaires pour le mettre en œuvre.

Tableau 10. Description des besoins exprimés

Besoins exprimés	Description
Sur les retours pertinents vers l'enseignant	Un suivi sur le long terme, plusieurs années, afin de pouvoir s'adapter aux besoins individuels et collectifs. Des retours précis des attendus exacts en termes de compétences. Des outils pour la différenciation pédagogique, pour mesurer précisément les progrès par compétence des élèves. Un outil qui permettrait une meilleure lisibilité de l'évolution de l'élève. Stats + personnalisation de l'apprentissage.
Sur les actions souhaitées	Des temps d'échanges, de la formation, des démonstrations, des outils conçus pour les usagers. Du temps pour tout mettre en œuvre. Du matériel adapté (ce qui est proposé par les conseils départementaux n'est pas toujours adapté aux usages pédagogiques). Plus de formations numériques pour l'ensemble des professeurs, pour que les équipes pédagogiques aillent dans le même sens. Du temps pour l'analyse. Mise en place de plus de moyens pour les dispositifs hybrides. Avoir des comptes illimités de stockage, avoir un matériel professionnel complet (ordinateurs fixes, portables et tablettes). Disposer d'un Moodle au sein de mon établissement. Outils numériques (questionnaire sur tablette…). Emploi du temps plus souple. Gestion de petits groupes d'élèves. Les grilles d'évaluation sont toujours fournies en PDF, pourquoi ne sont-elles pas disponibles en ligne, à compléter directement sur un espace « privé » d'une classe ? Pouvoir créer un suivi dans Pronote, en concertation avec l'équipe pédagogique de la classe. Base de données élèves, courbe de tendance, évaluation type… Pouvoir disposer de temps de formation des élèves.

B.5. Perception des enseignants des apports des Learning Analytics pour comprendre l'activité et pour le suivi des élèves

Figure 13. Apports des Learning Analytics selon les enseignants



B.6. Perception des limites des Learning Analytics et de la fiabilité des données pour restituer l'activité des élèves

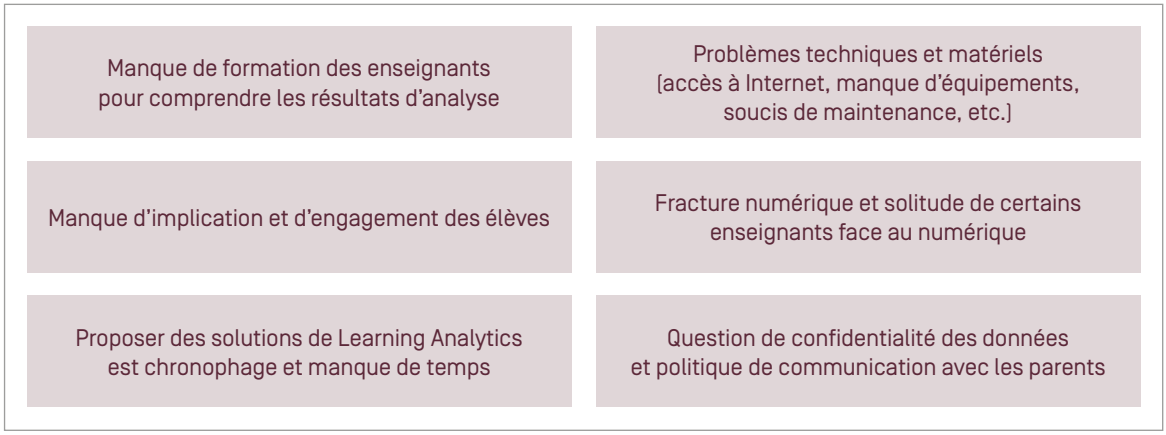
Figure 14. Perception des limites des Learning Analytics

Tableau 11. Description des limites des Learning Analytics

Limites/Fiabilité	Description
Analyse parcellaire de l'activité	Les Learning Analytics ne peuvent fournir une analyse complète de l'activité des élèves, mais plutôt parcellaire (données non représentatives car non complètes). Limite pour prendre en compte ce qui se fait en face à face pour appréhender l'activité dans sa globalité. Non prise en compte de la réflexion hors numérique (le coup de main du copain, le manque d'implication…). Tous les travaux des élèves ne sont pas analysables…
Insuffisance des compétences numériques des élèves et dépendance des Learning Analytics à l'utilisation du numérique	Tout ne peut pas être numérique, les élèves, pourtant « accros » des outils, se lassent de cours trop « numériques ». Les Learning Analytics sont aussi sujettes à la maîtrise du numérique par les élèves (difficulté de tracer quand les compétences numériques de l'élève sont faibles, les Learning Analytics sont conditionnées par l'utilisation du numérique lui-même, en lien avec les compétences numériques).
Questions éthiques et déontologiques	Droit à l'image et respect de la vie privée. Les données ne doivent pas être stockées et ne doivent pas à servir à autre chose que d'évaluer les élèves dans leur classe (utilisées à d'autres fins que pédagogiques… commerciales, par exemple). Les élèves ne sont pas nécessairement au courant des conditions d'utilisation de leurs données, ni de leur conservation. Risque de rupture du contrat moral et de confiance avec l'élève (potentiellement ressenti comme un effet de « flicage »). La confidentialité, même s'il y a déjà des restrictions au sein des établissements. La question de la sécurité des données. Tentation de comparer des élèves entre eux et d'établir des profils…
Confiance accordée à l'analyse des traces	Manque de formation et de recul nécessaire pour pouvoir analyser et interpréter les résultats objectivement. Risque d'erreurs énormes avec les élèves qui ne sont pas motivés et qui veulent juste se « libérer » de la tâche. Les données peuvent éventuellement cacher une tentative de tricherie. En distanciel asynchrone, on ne sait pas qui est devant l'ordinateur, et quel est l'investissement de l'élève dans l'activité : est-il concentré sur le travail ou fait-il un jeu vidéo en même temps ? Opacité de l'analyse pour l'enseignant : c'est à l'enseignant d'interpréter les données traitées par la machine. Les traces ne donnent qu'un état du passage. Elles ne traduisent pas si la compétence a été réellement acquise. Les traces ne doivent être considérées que comme des indicateurs de l'activité des élèves. Elles permettent de donner une tendance seulement.

B.7. Perception des difficultés et freins pour le déploiement de solutions de Learning Analytics dans les établissements scolaires

Figure 15. Perception des difficultés et freins au déploiement des Learning Analytics



Conclusion

Notons que quand l'enseignant n'a pas recours aux outils numériques, il utilise différents instruments qui lui permettent de collecter des données informelles et non structurées pour assurer le suivi de l'activité de l'élève, pour assurer une pédagogie différenciée, pour le suivi continu, y compris hors classe, etc. En effet, le suivi des élèves étant important d'un point de vue pédagogique, les enseignants ont recours à des instruments, des outils pour assurer ce suivi, au-delà de la traditionnelle démarche consistant à évaluer les connaissances. Toutes ces pratiques et ces données sont autant d'éléments pertinents pédagogiquement, qui pourraient être instrumentés par des outils de Learning Analytics, d'où l'importance d'analyser ces pratiques de façon fine pour les prendre en compte dans les nouvelles solutions logicielles de Learning Analytics.

Toutefois, mettre des outils de Learning Analytics à disposition des enseignants ne suffit pas pour qu'ils les intègrent dans leurs pratiques pédagogiques, même si elles facilitent souvent le travail d'évaluation, du suivi, de la personnalisation des apprentissages réalisé grâce au numérique.

Ce que nous retenons de ce qui a été déclaré par les enseignants dans notre enquête, c'est que, bien que les apports et l'utilité des Learning Analytics soient plutôt bien perçus par les enseignants, leur avis concernant leur déploiement dans les établissements scolaires reste très mitigé du fait du manque de formation au numérique, du manque de compétences numériques aussi bien chez les enseignants que chez les élèves. Le manque de moyens matériels au niveau des établissements est également souligné. Les enseignants sont aussi conscients des limites de l'utilisation des données d'apprentissage pour appréhender l'activité de l'élève de façon globale, car les apprentissages ne se cantonnent pas uniquement au numérique. Les questions liées à l'utilisation des données personnelles sont également au cœur des préoccupations des enseignants quand la question de la collecte des données se pose.

4 LES ENJEUX ÉTHIQUES DES LEARNING ANALYTICS

Florence Cherigny

À la croisée des sciences sociales et de l'informatique, l'émergence des Learning Analytics conduit effectivement à reposer des questions classiques dans le contexte du Big Data : l'articulation des libertés individuelles et du progrès scientifique et technique, les impacts personnels, sociaux et culturels de l'analytique de l'apprentissage, les problèmes du respect des droits et libertés fondamentaux (protection de la vie privée, lutte contre les discriminations…). Toutefois, l'éthique et la déontologie pouvant se définir à travers un ensemble de valeurs que les individus se fixent en dehors des normes juridiques, les préoccupations juridiques ne seront pas, ici, au centre de l'analyse. Les références normatives contribuent seulement à fournir un outil de réflexion pour ouvrir une discussion dépassant le cadre du droit positif, pour envisager des dilemmes éthiques et déontologiques.

L'éthique interroge le sens du recours aux Learning Analytics dans la perspective d'une réflexion morale reposant sur le respect de soi, des autres et de ce qui nous entoure. L'éthique implique également la recherche d'un ensemble de principes et de valeurs qui sont à la base d'une sagesse de l'action pour pratiquer les Learning Analytics. La réflexion éthique doit donc mobiliser à la fois, une éthique de conviction fondée sur des principes intangibles, sur l'affirmation de valeurs qui donnent un sens à l'action, et une éthique de responsabilité qui s'interroge sur les fins, les moyens et les conséquences des décisions et des actes. Ces deux éthiques s'interpellent pour s'enrichir mutuellement, dans le cadre d'une démarche méthodologique qui concerne tant le caractère éthique de la production des données et des analyses que le caractère éthique de leur utilisation.

Dans tous les cas, réfléchir sur l'éthique en matière de Learning Analytics, c'est d'abord être en proie au doute, celui qui interroge une discipline qui n'est, en soi, ni bonne ni mauvaise, puisqu'elle ne constitue qu'un outil de la pédagogie et de la recherche didactique. Ce doute ne doit pas être perçu comme une manifestation de défiance à l'égard de ceux qui pratiquent cette discipline. Il est un outil de questionnement méthodologique propre à la réflexion éthique. Cette réflexion engage l'esprit critique parce qu'elle renvoie dans son exigence à une universalité d'impératif moral. Le but n'est pas de diaboliser une discipline mais de profiter de ses bénéfices, en prévenant ses possibles dérives.

Dans le cadre de cette démarche critique, la réflexion sur les enjeux déontologiques s'attache, de manière plus circonscrite, aux difficultés suscitées par les pratiques de Learning Analytics pour les professionnels de la communauté éducative (personnels d'éducation, enseignants, chercheurs, gestionnaires et administrateurs, etc.), y compris dans les liens qu'elles impliquent avec d'autres acteurs (les apprenants, les parents d'élèves, etc.). Elle interroge des conceptions didactiques, pédagogiques, éducatives, en soulevant des problèmes portant à la fois sur les conditions de la recherche en Learning Analytics (conditions de soutenabilité de la recherche, de la protection des fruits de la recherche, de la diffusion de la connaissance, etc.) et sur les conditions de sa pratique (information des parties prenantes, recueil des consentements, protection des données personnelles des participants, etc.).

Dans cette section, la réflexion sera centrée, dans une perspective générale, sur les enjeux portant sur l'éthique des Learning Analytics.

Les préoccupations éthiques doivent d'abord conduire à s'interroger, en matière de santé publique, sur les risques physiques susceptibles d'être encourus par les parties prenantes à des expériences de Learning Analytics. Elles invitent également à mesurer les éventuels risques psychologiques découlant du recours aux Learning Analytics pour ces parties prenantes. Enfin, elles doivent interroger les possibles risques sociaux liés au développement d'une discipline qui se situe précisément à la croisée des sciences sociales et de la science des données.

Learning Analytics et enjeux liés au respect de la vie privée

Les risques psychologiques liés à l'utilisation des Learning Analytics concernent à la fois le respect de la vie privée dans un contexte éducatif de surveillance, et le respect de la personnalité dans le cadre du phénomène de quantification de soi.

SURVEILLANCE ET RESPECT DE LA VIE PRIVÉE

Les Learning Analytics soulèvent le problème du droit au secret de la vie privée et du droit à être laissé tranquille.

Le droit au secret de la vie privée

L'analytique des apprentissages implique la production, la collecte et le traitement de données, qui soulèvent inévitablement la question de la protection des données personnelles, elle-même indissociablement liée au respect de la vie privée. Ainsi que l'observe la Commission nationale consultative des droits de l'Homme, dans son avis du 22 mai 2018 sur la protection de la vie privée à l'ère du numérique, « à l'heure du "Big Data", toutes les données sont potentiellement personnelles et font peser un risque sur la vie privée. Car, si le prélèvement de la donnée pris individuellement est théoriquement encadré par une finalité, la somme des données que l'on renseigne en ligne, par le biais de processus d'agrégations et de recoupements automatisés, peut produire des informations nouvelles, constituant parfois des renseignements très détaillés sur les caractéristiques de la vie privée d'une personne ». Ainsi, « d'une part, toute donnée, même a priori anodine, est potentiellement un enjeu de vie privée et, d'autre part, consentir à verser des données à des services en ligne, au cas par cas et pour des finalités spécifiques et distinctes, ne permet pas forcément de contrôler les informations qui seront potentiellement produites par agrégation automatisée, via notamment des applications tierces » (Avis du 22 mai 2018 sur la protection de la vie privée à l'ère du numérique, *JORF* n° 0126 du 3 juin 2018, texte n° 63).

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. En effet, la réussite des élèves est un phénomène complexe et multidimensionnel, ce qui peut inviter à se demander s'il y a intellectuellement des limites fonctionnelles à la définition des données pertinentes à collecter. Certes, des limites

fonctionnelles sont assignées par les prescriptions du RGPD, qui impliquent que les données personnelles soient collectées pour des finalités déterminées, explicites et légitimes et que seules les données strictement nécessaires pour atteindre ces finalités puissent être collectées et traitées. Mais, en matière de Learning Analytics, il est difficile de déterminer à l'avance quelles informations sur l'environnement et les résultats de l'apprentissage seront susceptibles de se révéler probantes ou, au contraire, anecdotiques. L'analyse des activités d'apprentissage s'inscrit dans le cadre d'une approche exploratoire qui propose une modélisation a posteriori. Elle ne correspond pas – et c'est ce qui nourrit précisément les critiques à son égard – aux approches classiques des sciences humaines et sociales, qui cherchent à valider des hypothèses fondées sur un modèle théorique a priori. De ce point de vue, la logique de recherche qui sous-tend l'analytique des apprentissages peut sembler difficilement compatible avec la logique de minimisation de la collecte qui prévaut en matière de réglementation des données personnelles.

Quelles sont les données pertinentes à analyser ? S'agit-il uniquement des données enregistrées dans le cadre d'une activité d'apprentissage ? S'agit-il de données plus larges ? Des données scolaires, au sens large du terme ? Des données sociales (étudiant boursier, salarié…) ? Des données sensibles telles que les données physiologiques ? À cet égard, il est fondamental qu'au-delà de l'exigence du respect de la réglementation en matière de données personnelles, une réflexion sur les conditions de l'éthique de la récolte des données s'instaure. Cette exigence éthique s'avère d'autant plus nécessaire qu'« [il] n'existe pas de définition "officielle" des données scolaires. Le terme mérite d'être précisé. On les considérera dans le présent rapport comme toutes données recueillies dans le cadre scolaire. Le périmètre est donc très large : informations administratives sur les élèves, les enseignants, les personnels administratifs, les intervenants extérieurs, les parents…, les productions d'élèves ou de professeurs réalisées lors d'activités pédagogiques, des traces d'apprentissage » [IGEN-IGAENR, 2018, p. 10]. Cette réflexion paraît d'autant plus importante lorsque les données ne sont pas des traces d'apprentissage à proprement parler, mais plutôt des indicateurs de l'administration qui permettent de quantifier les conditions de vie d'un apprenant, au collège ou au lycée, ou d'un étudiant sur le campus, c'est-à-dire des facteurs extérieurs à l'acte d'apprendre.

La collecte doit toujours être orientée vers une finalité qui permette de mesurer la pertinence des données recueillies en fonction de l'objectif éducatif recherché. Par ailleurs, il y a encore loin du juridique à l'éthique. Ainsi, les nouvelles technologies permettent d'obtenir des données sur les apprenants de plus en plus liées à leur intimité corporelle (utilisation des capteurs biométriques, des empreintes digitales, de la reconnaissance faciale…) qui invitent, au-delà des aspects juridiques (protection de données dites « sensibles »), à se préoccuper, d'un point de vue éthique, du caractère pertinent d'une surveillance généralisée des corps. La collecte des données personnelles s'avère encore plus problématique, sur le plan de l'éthique, lorsqu'elle concerne des personnes mineures. D'une part, parce qu'elle peut alors trouver à s'appliquer à des personnes juridiquement incapables de consentir à cette collecte (mineurs de moins de 15 ans). D'autre part, parce qu'elle peut impliquer des données, parfois fort intimes (données physiologiques), recueillies de manière précoce sur des personnes qui n'en conserveront pas nécessairement le souvenir, et dont il convient pourtant de s'assurer qu'elles seront effectivement, plus tard, en mesure de disposer d'un droit à l'oubli.

La production et la collecte des données de Learning Analytics doit viser, autant que faire se peut, à préserver l'anonymat et/ou assurer la confidentialité des données et protéger la vie privée de toutes les parties prenantes (y compris, dans certains cas, les parents et proches des apprenants). Cette vigilance doit d'autant plus s'imposer que, aujourd'hui, des possibilités illimitées de croisements de données anonymes et de métadonnées permettent très facilement de « ré-identifier » les personnes, quand bien même toutes les données auraient été, en amont, anonymisées. Ne serait-ce que pour ces raisons, il serait judicieux que le responsable d'un traitement de Learning Analytics s'appuie sur l'expertise des délégués aux données personnelles (DPD). Par ailleurs, le traitement de données des Learning Analytics devra s'accompagner d'une explicitation des finalités d'utilisation des données, en particulier pour les mineurs puisque, depuis la mise en vigueur du RGPD, pour la première fois dans la législation européenne, il est demandé que l'information sur le traitement des données pour

les mineurs soit rédigée de manière à ce que les enjeux puissent être compréhensibles par des enfants. Ainsi, les problèmes du statut et du sort des données des Learning Analytics devront être soulevés. Qui pourra avoir accès à ces données et à l'analyse des résultats ? Quel degré de connaissance les apprenants devront-ils ou pourront-ils avoir des données recueillies et analysées ? Au-delà des problèmes juridiques, ces questions sont assurément au cœur d'enjeux éthiques et déontologiques forts.

Le droit d'être laissé tranquille

Les méthodes des Learning Analytics permettent de capter les comportements toujours plus facilement et régulièrement, ce qui soulève le problème du « droit d'être laissé tranquille ». Il est notamment 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. Les caméras à reconnaissance faciale et les bracelets électroniques, qui permettent le suivi des données physiologiques, peuvent également révéler les émotions ou le niveau d'attention des apprenants.

Il semble évident que les mineurs éprouvent des difficultés à prendre conscience des enjeux liés à la protection de leurs données personnelles si l'exemple qui leur est donné « d'en haut », par les adultes (les enseignants, les parents, etc.), ne les incite pas à questionner le sens de la collecte et du traitement de leurs données. Dans cette mesure, même lorsqu'il n'est pas envisagé de recueillir leur consentement, il n'est sûrement pas absurde que l'analytique de l'apprentissage conduise la communauté éducative à aborder avec des jeunes apprenants des enjeux qui relèvent de l'éducation aux médias et à la citoyenneté. Certains estiment qu'il « *paraît difficile de recueillir le consentement des mineurs, quel que soit leur âge, dans le cadre scolaire. En effet, le 11) de l'article 4 du RGPD précise que le consentement consiste en une "manifestation de volonté libre, spécifique, éclairée et univoque, par laquelle la personne concernée accepte, par une déclaration ou par un acte positif clair, que des données à caractère personnel la concernant fassent l'objet d'un traitement". Or, il est permis de s'interroger sur la question de savoir si, dans le cadre scolaire,* l'élève peut être regardé comme donnant valablement son consentement compte tenu de l'autorité *qu'exerce sur lui l'enseignant qui propose l'utilisation d'une application numérique en classe. En tout état de cause, qu'il soit mis en œuvre sur le fondement du consentement de la personne concernée ou de l'exercice d'une mission d'intérêt public* » [Réseau Canopé, 2018]. Cette sensibilisation des apprenants à une réflexion sur le contrôle de leurs données apparaît particulièrement importante lorsque l'analytique de l'apprentissage s'inscrit dans le cadre d'une « mission d'intérêt public » qui permet de soustraire le traitement de données personnelles à l'exigence de consentement préalable. En effet, selon l'article 6 du RGPD, les traitements effectués dans le cadre scolaire, à partir du moment où ils sont nécessaires à l'exécution d'une mission d'intérêt public ne nécessitent pas de consentement préalable. Le RGPD du 27 avril 2016 ne définit pas la notion de « mission d'intérêt public » mentionnée à l'article 6. La gestion de la vie scolaire entre dans ce périmètre. Mais il sera nécessaire de mieux préciser la notion. Tous les services numériques éducatifs relèvent-ils d'une mission d'intérêt public ? La question peut être discutée. En principe, les traitements ayant pour objet de permettre aux élèves ou aux enseignants d'effectuer des formations en ligne (e-learning) et l'utilisation d'un logiciel ou d'un service numérique à des fins pédagogiques entrent dans le champ du service public du numérique éducatif défini à l'article L. 131-2 du Code de l'éducation. Cependant, si le responsable de traitement n'était pas en mesure de justifier que le traitement qu'il souhaite mettre en œuvre rentre bien dans le champ de la mission d'intérêt public dont il est investi, il serait évidemment tenu d'obtenir le consentement des personnes concernées par le traitement. Apprendre aux élèves à réfléchir à la construction et à la protection de leur identité peut s'avérer à terme un puissant antidote aux relations de pouvoir asymétriques de la surveillance favorisées par l'émergence du Big Data.

D'une manière générale, le fait que le consentement de l'apprenant ne soit pas exigé par des prescriptions légales ne doit pas aboutir à exclure l'apprenant d'une réflexion sur des enjeux qui le concernent directement, puisqu'ils impliquent la collecte de données qui lui sont précisément personnelles. Cela paraît d'autant plus important que les Learning Analytics peuvent être au cœur de questions éducatives qui touchent justement à la part de respect

de la vie privée reconnue au mineur, relativement aux prérogatives des parents et/ou des enseignants. À cet égard, le rapport « Données numériques à caractère personnel au sein de l'Éducation nationale » [IGEN-IGAENR, 2018] évoque une expérimentation, menée dans l'Oise, d'usages autorisés par des parents qui soulèvent des interrogations quant à leur caractère éthique, alors même qu'ils respectent le cadre juridique : des élèves ont été équipés de bracelets connectés entre leur domicile et l'école puis, à l'heure du déjeuner, entre l'école et la cantine. À la montée et à la descente du bus, chaque bracelet est détecté par un smartphone qui envoie automatiquement un SMS et/ou un courriel aux parents, afin de les informer que leur enfant est bien arrivé. L'expérience est relatée dans un développement concernant des questions sur l'intérêt public de l'utilisation des données « scolaires », ce qui est déjà de nature à interpeller. Même si l'expérience dépasse largement le cadre des Learning Analytics, elle est intéressante car elle soulève la question des relations parents/enfants sous-jacentes à la collecte des données personnelles des mineurs [21]. Il n'est pas absurde de penser que, au cœur d'enjeux éducatifs forts, les outils des Learning Analytics puissent être accueillis par certains parents comme l'occasion rêvée d'avoir accès à de véritables « chaperons virtuels », ce qui pourrait en favoriser leur acceptation.

Le risque pourrait alors être que, les titulaires de l'autorité parentale ne consentant pas à la collecte des données personnelles, l'enfant ne puisse pas suivre l'enseignement ou interagir dans les mêmes conditions que les autres apprenants, ce qui représenterait un risque de rupture d'égalité entre les élèves. Les difficultés liées à ces situations de traitement différencié des élèves invitent bien souvent à considérer que l'absence de consentement des parents doit être systématiquement palliée par le recours à l'argument de l'exécution d'une mission d'intérêt public. Pourtant, s'il venait à être utilisé de manière systématique, l'argument de la mission d'intérêt public pourrait apparaître aux yeux des parents comme purement invocatoire. L'entrée en vigueur du RGPD a eu, entre autres, le mérite de développer une culture commune autour de la protection des données personnelles. Il est probable que, si l'esprit de l'article 6 du RGPD n'était pas respecté, certaines communautés de parents d'élèves auraient à cœur de démontrer la violation des prescriptions de ce règlement.

Si les Learning Analytics permettent de donner du sens à l'activité de l'apprenant (à travers ses clics, la mesure de son attention, l'analyse de réseaux sociaux…), ils permettent également d'évaluer et de donner du sens aux activités de l'enseignant, la production de données des Learning Analytics s'opérant très souvent dans un va-et-vient entre le formateur et l'apprenant. Si l'enseignant consent à surveiller l'élève, consentira-t-il aussi, en retour, à être surveillé ? La recherche-action en matière de Learning Analytics ne pourra se développer qu'avec l'adhésion des enseignants et des chercheurs, ce qui implique que tous soient rassurés quant aux conditions de la collecte, de la protection et du traitement des données personnelles. Car les résultats des recherches de l'analyse de l'apprentissage sont susceptibles de donner naissance à des applications opportunistes, permettant l'exploitation des données d'une manière qui n'avait pas été prévue au départ. Il est important, à cet égard, de garantir aux parties prenantes que les Learning Analytics visent à favoriser la compréhension et l'amélioration des apprentissages et de s'assurer au maximum que les analyses ne puissent être détournées dans une visée de contrôle académique non consenti par les parties prenantes (contrôle académique des formations, amélioration du rendement des enseignants, évaluation des enseignants au mérite…). Il est également essentiel d'assurer la sécurité des données afin d'éviter la fuite de ces dernières au profit de tiers qui n'ont aucun intérêt légitime à y avoir accès et/ou pourraient en faire un usage inapproprié, ce qui soulève notamment la question de la sécurité et des conditions d'hébergement des données.

21. On observe qu'en Chine, dans la cafétéria de certaines institutions, des caméras identifient chaque élève qui fait la queue et enregistrent le contenu de son plateau repas pour envoyer le détail nutritionnel à ses parents (Turretini E., « En Chine, des "uniformes intelligents" pour les élèves », Le Temps, 06/01/2019), [En ligne].

QUANTIFICATION ET RESPECT DE LA PERSONNALITÉ

La quantification de soi permise par le développement des Learning Analytics soulève des questions relatives à la mesure de la personnalité et à la recherche de la performance.

Quantification et mesure de la personnalité

Les bénéfices attendus des Learning Analytics en matière de respect de la personnalité de l'apprenant sont multiples. En associant les plateformes d'apprentissage en ligne à différents capteurs, les enseignants sont en capacité d'analyser toujours plus finement les réactions de leurs élèves, d'évaluer leur degré d'implication et d'engagement (analyse des fréquences de connexion, des types de lectures, des interactions…) et de proposer un accompagnement proche des besoins de chaque apprenant. En retour, les Learning Analytics permettent à l'élève de disposer d'un environnement d'apprentissage personnalisé, de faire le point sur ses propres forces et faiblesses, d'accéder à des parcours pédagogiques spécifiques, de bénéficier d'outils adaptés à son niveau et de suggestions d'activités correspondant à ses marges de progression. Les Learning Analytics conduisent donc à espérer qu'un véritable enseignement « sur mesure », au service de l'apprenant, ne soit pas incompatible avec un enseignement de masse. Encore convient-il de mesurer cet enthousiasme en rappelant une autre évidence : le cadre des Learning Analytics est en lien étroit avec le phénomène de la quantification de soi, dont l'utilisation peut être instrumentalisée à des fins très diverses.

Les risques liés à l'engouement pour le phénomène du « *quantified self* » (le moi quantifié), très bien décrits par la CNIL dans son étude sur « Le corps, nouvel objet connecté » [CNIL/LINC, 2013], interrogent naturellement l'activité des Learning Analytics, d'autant plus lorsqu'ils concernent des personnes mineures. Si le recours aux Learning Analytics vise à permettre le développement de l'éveil de l'apprenant (développement des capacités cognitives, psychomotrices, etc.), la quantification des performances de l'enfant est susceptible de faire naître une forme de « normopathie » anxiogène. L'élève est-il conforme aux paramétrages ? Est-il suffisamment habile à la lecture, actif physiquement… Il est donc important de souligner que les dispositifs de Learning Analytics doivent être utilisés avec mesure, ce qui implique un questionnement sur l'adéquation de la collecte de données mise en place par rapport à la finalité recherchée. La recherche de ce rapport d'adéquation doit être systématisée, de manière pragmatique, en conformité avec les objectifs de la réglementation en matière de données personnelles, au cas par cas.

Quantification et recherche de la performance

Certains dispositifs conçus pour maximiser le potentiel de l'apprenant dans le cadre de son apprentissage risquent de soulever des questions liées à l'obsession de la recherche de la performance. Dans la mesure où les Learning Analytics permettent de détecter les signes d'ennui ou de perte de l'attention, voire d'alerter l'apprenant sur ce déficit, ils favorisent une évaluation régulière des performances des élèves. Il est important de rappeler que la « réussite scolaire » n'est pas nécessairement le résultat d'un temps linéaire et de prendre conscience des effets anxiogènes d'une sur-sollicitation de l'apprenant. Car les Learning Analytics pourraient contribuer à renforcer la propension à favoriser la production de « résultats » sous forme de notes. Sans oublier qu'une systématisation des outils d'alerte, de rappel à l'ordre ou de recommandation risque de lasser et de déresponsabiliser l'apprenant, tenté de s'en remettre au système sans s'interroger sur sa propre posture, dans le cadre de son apprentissage.

Dans tous les cas, il convient d'être très vigilant sur le fait que l'apprenant ne doit pas être transformé en un « objet » d'apprentissage. À cet égard, les Learning Analytics mériteraient toujours d'être conçus comme un outil au service de l'apprenant et non comme l'objet d'une expérimentation au service d'une institution. L'analyse de données massives et l'objectivation de l'apprentissage ne doivent pas contribuer à transformer les Learning Analytics en une discipline qui se résumerait à déduire la qualité ou l'absence de qualité d'un apprenant de l'observation d'un amas de data, ou à réduire l'apprenant, pour des raisons de confort ou d'économies, à un simple calcul de probabilités. A fortiori, elle ne doit pas transformer l'apprenant en un producteur passif de données, ni lui assigner de se conformer passivement

aux recommandations ou prescriptions d'un algorithme. Il importe de rappeler que l'analyse de l'apprentissage doit viser à fournir des indications pertinentes pour décider ce qui est approprié et moralement nécessaire à l'apprenant. L'analyse de l'apprentissage devrait fonctionner comme une pratique morale aboutissant à la compréhension plutôt qu'à la mesure [Slade, Prisloo, 2013].

Learning Analytics et enjeux de santé publique

D'un point de vue éthique, ne pas mettre les participants dans des situations où ils risquent de subir des préjudices physiques du fait de leur participation à la recherche en Learning Analytics doit être le premier objectif à assurer. À cet égard, c'est principalement le recours aux objets connectés qui suscite question. D'une part, l'exposition aux champs électromagnétique peut inquiéter, spécialement s'agissant des apprenants mineurs qui forment une population particulièrement sensible à ces champs. D'autre part, l'utilisation de certains équipements peut également contribuer à alimenter des problèmes de surexposition aux écrans, problèmes également très prégnants s'agissant des enfants.

L'EXPOSITION AUX CHAMPS ÉLECTROMAGNÉTIQUES.

Ainsi qu'il a été observé dans le rapport « Le numérique au service de l'École de la confiance », « le déploiement progressif des objets connectés dans tous les domaines de la vie sociale incite à mettre ces différents produits interactifs et communicants au service des apprentissages. Demain, les écrans ne seront très probablement plus l'interface dominante entre les individus et les machines. Qu'il s'agisse de suivre ses progressions (avec des bracelets dédiés à l'éducation physique et sportive), de récupérer des données sur des capteurs de toute nature (lunettes, drones, objets domotiques, etc.) ou encore d'apprendre à programmer (des robots par exemple), les objets connectés vont enrichir et renouveler considérablement les modalités d'apprentissage » [MEN/MESRI, 2018]. Ainsi pour les Learning Analytics, au même titre que pour d'autres utilisations (familiales, ludiques, etc.), les risques liés à l'exposition aux champs électromagnétiques ne méritent pas moins d'être soulignés. Ils intéressent particulièrement la santé des mineurs en bas âge et représentent des risques d'autant plus difficiles à appréhender qu'on ne dispose pas toujours d'études cliniques significatives sur cette catégorie de population. Mais il est aussi très vite apparu qu'il existe, pour l'instant, peu d'études portant sur les enfants de moins de 6 ans. La plupart des articles répertoriés, considérant que l'âge de la première utilisation du téléphone mobile se situe rarement avant 7 ans, portent sur des enfants plus âgés. Or, selon ce rapport, la surexposition des jeunes enfants aux radiofréquences issues des objets connectés du quotidien (téléphone portable, tablette, ordinateur, etc.) aurait des impacts sur leurs fonctions cognitives et serait responsable de certains troubles identifiés, comme les symptômes dépressifs ou la perte de mémoire [22]. Les études rapportent des niveaux d'exposition plus élevés chez les enfants que chez les adultes et démontrent que pour toute personne de taille inférieure à 1,30 m, les valeurs limites d'exposition réglementaires sont moins adaptées [23]. Dans l'attente d'études complémentaires, il convient donc d'appliquer un principe de précaution qui doit conduire à ne pas négliger les risques d'exposition aux champs électromagnétiques dans le cadre d'expérimentations sur de très jeunes apprenants, en particulier pour les outils se situant près de la tête (serre-têtes pour mesurer les signes d'activité cérébrale, par exemple) [24].

22. Cf. Anses, « Exposition des enfants aux radiofréquences : pour un usage modéré et encadré des technologies sans-fil », actualité publiée le 08/07/2016, [En ligne].
23. En Europe, des études ont déjà amené le Conseil de santé des Pays-Bas à considérer, en 2011, que les niveaux de référence définis par l'ICNIRP (International Commission on Non Ionizing Radiation Protection) et adoptés par la Recommandation européenne n° 1999/519/CE, autour de 2 GHz, devaient être ajustés.
24. En Chine, l'utilisation de bandeaux « détecteurs d'attention », reposant sur une technologie inspirée de l'électroencéphalographie, semble déjà utilisée dans des écoles primaires. Œuvre d'une start-up de Boston, elle permet de mesurer le niveau d'attention d'un sujet et, éventuellement, le rappeler à l'ordre (cf. Le Gohlisse N., « La Chine testerait des bandeaux "détecteur d'attention" dans les écoles », SiècleDigital, 05/04/2019, [En ligne].

LA SUREXPOSITION AUX ÉCRANS

La surexposition aux écrans constitue aujourd'hui une question importante de santé publique concernant les très jeunes enfants. Les écrans émettent de la lumière bleue, ou lumière HEV artificielle, et ce type de rayonnement peut s'avérer nocif [25]. Des recherches scientifiques montrent qu'une exposition prolongée peut provoquer des lésions photochimiques du cristallin ou de la rétine. Or, le cristallin des enfants qui n'ont pas encore 14 ans ne filtre pas aussi bien la lumière bleue que celui d'un adulte.

Par ailleurs, certains professionnels s'inquiètent de la forte augmentation des troubles intellectuels et cognitifs chez l'enfant en très bas âge, et alertent sur la nécessité de lutter contre la surexposition précoce aux écrans de télévision, d'ordinateur, de tablette et de téléphone. À partir d'éléments cliniques, ils décrivent un trouble nouveau, un syndrome neuro-développemental appelé Epeé, pour « exposition précoce et excessive aux écrans » [Marcelli, Bossière, Ducanda, 2018]. Ce trouble serait lié à la présence de l'écran venant perturber l'environnement de l'enfant, en interférant dans les besoins essentiels à son développement. Certes, le temps d'exposition aux écrans pris en considération par les spécialistes vise en tout premier lieu l'utilisation dans un cadre familial et concerne une population d'enfants en très bas âge (moins de 4 ans). Néanmoins, les interrogations liées à l'apparition de ce nouveau trouble doivent inviter à un principe de vigilance s'agissant de longs temps d'exposition des enfants aux écrans.

Les techniques de suivi oculaire (*eye-tracking*), qui permettent de déterminer si certains passages ont été lus plus ou moins lentement – pour attirer l'attention sur d'apparentes difficultés de compréhension ou sur des signes d'une perte de concentration – doivent aussi s'exercer dans un cadre sanitaire satisfaisant. Les méthodes de Learning Analytics promettent d'analyser les apprenants toujours plus finement, grâce à des dispositifs tels que des oculomètres. Avec des capteurs, le mouvement oculaire sur une page peut être analysé, afin de développer des manuels qui prendront en compte les habitudes de lecture et auront plus d'efficacité. Comme pour toute technologie nouvelle, ces dispositifs au contact corporel de l'apprenant suscitent des craintes quant à leur possibles effets sur la santé, craintes d'autant plus difficiles à appréhender que l'on ne sait pas quand le système visuel achève son développement, et que l'on ne dispose pas, là encore, d'études cliniques significatives sur le sujet.

En tout état de cause, il convient donc de réaliser que, plus les Learning Analytics impliquent des expériences sur un public jeune, plus ils s'adressent à un public fragile ; plus ils impliquent le recours à des technologies nouvelles, plus ils risquent de produire des effets physiques qui n'ont pas encore été pleinement déterminés. La prudence (utilisation d'un matériel adapté à l'enfant) et la modération (limitation du temps d'exposition) sont donc recommandées [26].

Learning Analytics et enjeux en matière d'impacts sociaux

À la croisée des sciences sociales et de l'informatique, l'émergence des Learning Analytics conduit à reposer, dans le contexte du Big Data, les traditionnelles questions du déterminisme et des discriminations.

LEARNING ANALYTICS ET DÉTERMINISME ÉDUCATIF

La production, l'utilisation et la réutilisation des données de Learning Analytics interrogent la question du déterminisme de manière plus ou moins intense, selon les objectifs poursuivis, et selon la perception que chaque communauté scientifique se fait du produit de la recherche

25. Cf. Anses, « Effets sur la santé humaine et sur l'environnement (faune et flore) des diodes électroluminescentes (LED), Expertise Anses 2019 », dossier téléchargeable, [En ligne].
26. Plusieurs travaux, tels que ceux du GTnum 4 : « Usages et pratiques numériques des jeunes » animé par Pascal Plantard, insistent sur la nécessaire prise en compte du contexte socio-économique et socioculturel des pratiques numériques des jeunes.

(notamment suivant qu'il s'agit d'alimenter une machine ou d'amplifier le rôle décisionnel des acteurs de l'apprentissage). Néanmoins, il semble y avoir une hypothèse implicite liée à l'analyse de l'apprentissage, selon laquelle la connaissance de l'apprenant et la modélisation prédictive permettraient d'identifier les apprenants à risque et de personnaliser les interventions afin d'augmenter les chances de réussite. A priori, l'objectif des Learning Analytics réside donc dans la croyance en une capacité de réécrire un avenir non déterministe. De fait, dresser le profil d'un apprenant ou d'une catégorie d'apprenants peut permettre de repérer plus précocement des troubles du langage (dysorthographie et dyslexie), des troubles arithmétiques (dyscalculie) ou encore des troubles de coordination motrice (dyspraxie), et ainsi favoriser une remédiation rapide qui serait de nature à lutter contre l'échec. Mais, ne serait-ce pas également, peut-être, prendre le risque de « déterminer » un peu trop tôt l'avenir déjà « pré-dit » d'un apprenant ou d'une catégorie d'apprenants ? La question d'un futur prédit par le passé est un dilemme éthique classique dans la réflexion sur le Big Data. À cet égard, il est légitime de s'interroger à la fois sur la nature des indicateurs qui permettent, en matière de Learning Analytics, de favoriser l'analyse de situations de prédispositions favorables ou de situations à risques (Predictive Analytics), et sur l'utilisation pouvant être faite des profils ainsi obtenus.

S'agissant de la pertinence des indicateurs, comme dans d'autres domaines, la question des biais potentiels, des simplifications excessives et des éventuelles « bulles de filtre » devra être soulevée. Cette question des biais devra l'être, même lorsque les Learning Analytics sont utilisés de façon agrégée et anonymisée, puisque ces biais peuvent avoir pour effet de perpétuer ou renforcer des discriminations et favoriser des « bulles d'échec ». Car, l'agrégation ayant pour effet de désincorporer les données de leur contexte original, elle augmente les risques de mauvaise utilisation et/ou de mauvaise interprétation. Par ailleurs, même poussée très en avant, la prédiction et la personnalisation conduisent à des risques d'enfermement de l'individu dans une sphère limitée de possibilités et de ségrégation des expériences. Le danger est notamment que, par construction, les Learning Analytics réalisés à partir d'expériences passées et/ou actuelles présentent un tropisme à la répétition et que les catégorisations réalisées en fonction de données historiques soient incomplètes et/ou erronées. Lorsque l'exploitation des données numériques produites par les apprenants s'appuie sur une modélisation permettant de découvrir des informations et des connexions sociales afin d'anticiper, prédire et conseiller l'apprentissage de manière personnalisée, il est particulièrement important que la méthodologie adoptée soit transparente. À cet égard, la proposition d'instrumenter la démarche éthique elle-même, grâce au développement de dispositifs qui prennent en compte une « dimension éthique par construction » (*Ethics by Design*), en étendant la démarche existante de « vie privée par construction », apparaît un des moyens les plus prometteur pour concilier exigences éthiques et déontologiques [Mille, Pérès-Labourdette Lembé].

S'agissant de l'utilisation des données de Learning Analytics, le risque est que l'analytique de l'apprentissage permette, à partir de la prédiction d'un parcours comportemental ou cognitif, de classer les élèves dès leur plus jeune âge pour leur assigner des parcours scolaires adaptés à leur profil, en fonction de leurs chances de réussite ou de risques d'abandon, sans tenir compte de leurs préférences personnelles. Il convient de rappeler que, d'un point de vue éthique, l'activité d'analyse d'apprentissage doit avoir pour finalité d'identifier les moyens d'aider efficacement les apprenants à atteindre leurs propres objectifs, en leur permettant d'exprimer des choix libres.

Un autre danger serait également de désigner à l'enseignant, de manière faussement prémonitoire, les forces ou les faiblesses d'un modèle d'apprenant. De ce point de vue, il y a lieu de s'interroger sur les effets induits qu'il peut y avoir à indiquer à un enseignant quels apprenants sont susceptibles de réussir ou d'échouer. Le risque est d'exposer les apprenants, de manière systématique et répétée, à des méthodes qui conduisent à confirmer des attentes stéréotypées (effet Pygmalion ou effet Rosenthal & Jacobson, [Rosenthal, Jacobson, 1968]). Il y a aussi lieu de s'interroger pour savoir s'il est toujours opportun de communiquer à l'apprenant un pronostic de bon ou de mauvais résultat, considérations qui devraient relever du libre arbitre des enseignants. Mais, à cet égard, on doit s'inquiéter des menaces liées à l'émergence d'un contentieux provoqué par des apprenants qui pourraient considérer comme fautive

l'absence de déclenchement d'une « procédure d'avertissement d'échec imminent », lorsque les tableaux de bord ont révélé un risque d'échec.

Il convient donc encore de se demander comment communiquer sur un pronostic de réussite ou d'échec : en raisonnant de manière neutre, à travers une présentation sèche de calculs de probabilités, ou dans le cadre d'un discours plus ou moins personnalisé de soutien permettant de maintenir la motivation… À cet égard, Régis Chatellier indique que les communications avec les apprenants sont peut-être plus efficaces si elles sont rédigées en termes généraux de soutien (« nous ne faisons que vérifier auprès de vous pour voir comment vont vos études », par exemple), plutôt qu'en termes de probabilités (« nous pensons que vous avez une probabilité de 10 % de terminer votre module actuel ») [cf. Chatellier, 2018].

Mais la plus grande crainte, parfois fantasmée, parfois exprimée, reste sans doute que, combinées avec les méthodes algorithmiques, les Learning Analytics ne deviennent l'antichambre d'un « *Minority Report* éducatif », conduisant les individus à rester prisonniers de leur passé dans l'avenir. À cet égard, il est bien entendu nécessaire de se défier d'une pseudo-objectivité machinique qui permettrait de s'en remettre à la « gouvernementalité algorithmique » mise en évidence par les travaux d'Antoinette Rouvroy, en s'inquiétant du risque de « dictature des algorithmes » dénoncé dès 2012 par la CNIL [CNIL/LINC, 2012]. Il convient de rappeler que le Règlement général européen pour la protection des données personnelles (RGPD), entré en application en mai 2018, interdit les décisions prises sur le seul fondement d'un algorithme et prévoit le droit à une explication en cas de décision prise par un algorithme. Les algorithmes utilisés dans le cadre éducatif ont déjà alimenté de nombreuses polémiques, ainsi que l'ont déjà démontré les « cas d'école » des algorithmes d'orientation et d'affectation Affelnet (Affectation des élèves par le Net) et APB (Admission post-bac). Même lorsque leur code source a été produit conjointement par une administration et un établissement public à caractère scientifique, culturel et professionnel, dans le cadre d'une mission de service public, ce qui rend la publication de ce code source possible, leur cadre juridique reste peu lisible et suscite de nombreuses difficultés. Ainsi, le code source d'APB est clairement un document administratif, aucun tiers extérieur à l'administration détenant des droits de propriété intellectuelle n'ayant été identifié. Mais la mission Etalab, qui coordonne notamment la conception et la mise en œuvre de la stratégie de l'État dans le domaine de la donnée, a elle-même reconnu que « le cadre juridique délimitant APB est assez peu lisible ». Le code informatique d'APB a permis de mettre en œuvre des dispositifs qui n'ont pas été prévus par le législateur, à l'instar du dispositif de tirage au sort pour les filières non sélectives sous tension [27]. Par ailleurs, les règles du RGPD ne couvrent que les effets individuels, et non les conséquences collectives des algorithmes, puisqu'elles n'encadrent a priori que des données personnelles identifiées ou identifiables, et négligent les difficultés suscitées par le recours à des données anonymisées et agrégées. Ne serait-ce qu'en raison de ces difficultés, il conviendrait, en amont, que la mise au point des algorithmes éducatifs s'effectue dans un cadre éthique.

Il est aussi essentiel de travailler très concrètement sur le « droit à l'oubli » du mineur. Certains observent que « les données pédagogiques sont bien plus sensibles que les données médicales, car elles concernent des enfants, qui construisent leur personnalité, et il serait grave de leur assigner des mauvais résultats ou une mauvaise conduite qui les suivront le reste de leurs vies. Pour cette raison, il est impératif que ces données restent confidentielles et qu'elles ne puissent pas être communiquées vers l'extérieur [28] ». En droit, les données personnelles recueillies ne doivent être conservées que pour la durée nécessaire à la finalité du traitement. Ainsi, les traces d'apprentissage collectées pendant la prime jeunesse ne devraient pas être conservées au-delà d'un délai raisonnable, ni risquer de conditionner une vie d'adulte. Pour la CNIL, par exemple, les résultats scolaires des élèves sont des données privées qui doivent être protégées [29]. La Commission s'inscrit d'ailleurs contre toute possibilité

27. Rapport de la mission Etalab sur les conditions d'ouverture du système Admission Post-Bac, 2017, p. 21, [En ligne].
28. La formule est de Gilles Dowek, président du conseil scientifique de la Société informatique de France et Professeur en informatique à l'ENS Paris-Saclay, qui estime que le prochain enjeu sera de faire inscrire les aspects sensibles de ces données pédagogiques dans la loi française, comme cela a été fait pour les données médicales, afin qu'elles soient encadrées et protégées de la même manière (source : Bourdet, 2019).
29. Avis du 22 mai 2018 sur la protection de la vie privée à l'ère du numérique, JORF, n° 0126 du 03/06/2018, texte n° 63, n° 57.

d'accès public aux résultats scolaires. Dans la même logique, elle s'inscrit contre tout tri des élèves/candidats à un emploi, indépendamment des décisions d'orientation des conseils d'orientation (pour les élèves) ou des diplômes.

Il paraît judicieux que le profil d'un élève ne le suive pas de la maternelle jusqu'à l'université, ni a fortiori jusqu'au marché du travail, et certaines évidences méritent parfois d'être rappelées car la question de l'instrumentalisation des Learning Analytics dans une perspective de déterminisme social, notamment du fait d'une « marchandisation des Learning Analytics », ne doit pas être sous-estimée.

LEARNING ANALYTICS ET DÉTERMINISME SOCIAL

Tout d'abord, les administrations et les entreprises de la filière Ed'tech peuvent être très intéressées par les données scolaires, dans le but de réfléchir à une allocation plus efficace des ressources, permettant d'optimiser économiquement les formations. Même lorsque les données de Learning Analytics n'ont pas un caractère personnel, elles présentent une nature stratégique et sont susceptibles de procurer des avantages concurrentiels dans le développement futur de services pour l'éducation [30]. La question mérite donc d'être posée : « Dans quelle mesure l'analytique des activités d'apprentissage instrumentées ne concourent-elles pas aussi à l'accélération de l'industrialisation et marchandisation de la formation ? » [Peraya, 2019]. Ici encore, c'est l'aspect déterministe de la culture du chiffre qui doit être interrogé, notamment au regard du modèle économique des marchés bifaces, qui permet de considérer l'apprenant à la fois comme un consommateur et comme un fournisseur de données. Dans la mesure où les données de l'apprenant – réponses, commentaires, créations diverses, etc. – vont pouvoir nourrir la formation, mais également permettre d'évaluer la pertinence économique du modèle de la formation, les entreprises pourraient être tentées de marchandiser les données des apprenants, soit en subordonnant l'inscription à certains cours à la collecte de certaines données, soit en concédant une remise sur le prix de la formation lorsque les apprenants consentent à la collecte de certaines de leurs données. Par exemple, en 2014, en Angleterre, l'UCAS, l'organisme chargé de l'admission dans les universités britanniques, a vendu à de grandes compagnies l'accès aux données personnelles des étudiants, un commerce qui lui a rapporté cette année-là 14 millions d'euros. Certes, les étudiants pouvaient, lors de leur inscription, demander à ne pas recevoir de communications de la part de l'organisme. Mais il ne leur était pas possible de dissocier publicité et orientation. S'ils refusaient de recevoir des e-mails et courriers de la part de l'UCAS, ils n'avaient tout simplement plus accès aux *newsletters* en lien avec leur parcours académique [31].

C'est alors le problème plus général du droit à l'autodétermination en matière de données personnelles (et de l'apparition d'une forme de « *digital labor* » en découlant) qui pourrait être soulevé. Indirectement, ce sont également les questions du droit d'accès à l'éducation et du risque de discrimination économique entre les apprenants qui pourraient aussi se trouver posées.

LEARNING ANALYTICS ET DISCRIMINATIONS

Les Learning Analytics soulèvent le problème de la distinction entre personnalisation de l'apprentissage et discriminations, et celui du risque de discriminations lié à l'instrumentalisation des données de Learning Analytics.

Personnalisation de l'apprentissage et discriminations

Dans la mesure où les Learning Analytics tendent à proposer des parcours pédagogiques mieux adaptés à l'apprenant, à son stade d'acquisition de connaissances et de compétences, ils forment a priori un formidable outil d'intégration. En favorisant un apprentissage adaptatif, fondé sur la personnalisation, ils ouvrent la voie à une pédagogie différenciée, facteur d'une

30. Le marché mondial de l'analyse de l'éducation et de l'apprentissage est un marché dont on prévoit une augmentation de la valeur estimée de 2,8 milliards USD en 2018, à 13,30 milliards USD en 2026 (source : Data Bridge Market Research, 2019).
31. Source : De Queroiz J.-A. , « Ce service universitaire qui revend les données personnelles des étudiants anglais », LeFigaro.fr, 14/03/2014, [En ligne].

bonne intégration éducative et sociale. Ainsi, le référentiel CARMO (Cadre de référence pour l'accès aux ressources pédagogiques via un équipement mobile – version 3.0, déc. 2018) permet de personnaliser les parcours et les apprentissages, en suggérant aux enseignants différents usages. Ce référentiel prévoit de « proposer aux élèves des ressources pédagogiques différenciées (exercices, cours…) selon leur niveau et leur avancement et/ou leur situation face à un handicap, avec disponibilité individuelle du terminal ; visionner en aval les séquences d'apprentissage afin d'identifier les points de blocage ; visualiser en temps réel la production de l'élève » [MENJ]. Il est certain que les Learning Analytics peuvent jouer un rôle tout à fait favorable pour améliorer l'intégration des élèves en difficulté ou en situation de handicap (détection des problèmes, mise à disposition de ressources adaptées aux élèves porteurs de handicap de type DYS, aux utilisateurs daltoniens…). Dans cette mesure, les Learning Analytics peuvent fortement contribuer aux enjeux d'une école inclusive.

Si la personnalisation de l'apprentissage constitue une des perspectives les plus pertinentes des Learning Analytics, elle implique, en retour, de vives inquiétudes concernant le risque de discriminations. Comment concilier l'objectif des Learning Analytics – collecter et exploiter les traces numériques laissées par les apprenants afin d'assurer des pratiques éducatives individualisées et réflexives – avec la lutte contre les discriminations ? Comment anticiper des démarches favorables à la différenciation pédagogique, notamment en permettant de diagnostiquer les risques de décrochage, tout en luttant contre l'usage d'indicateurs qui peuvent exclure ? S'agissant en particulier des apprenants mineurs, comment garantir aux parents – dont la sensibilité à la culture numérique peut être très variable – que les Learning Analytics seront utilisés à des fins éthiques ? La question peut paraître d'autant plus sensible que, comme l'observe la Commission nationale consultative des droits de l'Homme (CNCDH), « s'il fallait contraindre les services scolaires à une consultation de chaque parent afin de recueillir leur consentement au traitement des données de leur enfant, l'*É*ducation nationale se verrait contrainte d'offrir des services d'enseignement différenciés selon la sensibilité particulière de chaque parent [32] ». La Commission estime qu'il est urgent d'intégrer les données collectées dans le contexte scolaire au champ des traitements relevant d'une « mission d'intérêt public » et d'appliquer les principes de contournement du consentement, de consultation préalable de la CNIL et de conduite obligatoire d'une étude d'impact, lors de l'élaboration de tout nouveau traitement. Selon elle, cette double exception aux dispositions générale du Règlement – mais permise par celui-ci – permettrait de garantir le traitement à parité des enfants dans le cadre scolaire.

Ces préconisations de la CNCDH ont au moins le mérite d'ouvrir la discussion sur ce possible risque de discrimination.

Par ailleurs, les différents capteurs pouvant être utilisés dans le cadre des Learning Analytics – notamment les dispositifs permettant la reconnaissance faciale et l'analyse des émotions, ou les dispositifs de type montres connectées permettant une analyse du rythme cardiaque, de la température, etc. – peuvent conduire à collecter des données personnelles sensibles et/ou des données personnelles dont le croisement permettrait d'établir des profilages pouvant être utilisés à des fins discriminatoires. Ne serait-ce que de ce point de vue, le problème de la confidentialité de certaines données collectées dans le cadre des Learning Analytics s'avère aussi délicat que celui de la collecte de données médicales. Il est important que les institutions gardent confidentiels les détails des handicaps des apprenants qui auront pu être révélés par l'analyse des apprentissages, puisque celle-ci permet d'identifier, parfois même involontairement, certaines déficiences.

Une autre source d'inquiétudes provient de la facilité qu'il pourrait y avoir à déduire, dans la masse des données de Learning Analytics et à partir de données apparemment insignifiantes, des informations « sensibles » telles que les prétendues appartenances ethniques, préférences religieuses ou orientations sexuelles de l'apprenant. Il convient d'insister sur le fait que le croisement des données collectées dans le cadre des Learning Analytics ne devrait pas conduire à la discrimination à l'égard d'une personne ou d'un groupe de personnes à raison

32. Avis du 22 mai 2018 sur la protection de la vie privée à l'ère du numérique, *JORF*, n° 0126 du 03/06/2018, texte n° 63, n° 53.

de leur origine, ou de leur appartenance ou non-appartenance à une ethnie, une nation, une race ou une religion déterminée, ou à raison de leur sexe, de leur orientation sexuelle ou de leur handicap, toutes formes de discriminations condamnables, aussi bien d'un point de vue éthique que d'un point de vue juridique. Plus délicate encore, la question de l'introduction de la discrimination positive dans les Learning Analytics pourrait aussi être posée. Une étude du centre de recherche ECAR (Educause Center for Analysis and Research) fait état d'une expérience de l'université de Washington dans laquelle les Learning Analytics sont utilisés, sur un site regroupant plus de 20 000 étudiants issus des communautés de réserves indiennes et immigrées, pour accompagner les étudiants en difficulté scolaire ou en décrochage. Les représentants de l'université précisant d'ailleurs qu'ils tenaient à maintenir cette approche, même si ces résultats pouvaient impacter leur classement au niveau national et global[33]. Comme dans tout système de discrimination positive, des considérations éthiques conduisent alors à s'interroger sur les biais idéologiques involontairement introduits. Enfin, au-delà du risque d'une instrumentalisation volontairement discriminatoire des Learning Analytics, le danger réside surtout dans le fait que les opérations de typification des apprenants intègrent des biais ou préjugés qui contribuent à la reproduction ou au renforcement des discriminations. Car si, d'un côté l'exploration de données peut aider à reconnaître et modéliser la diversité entre les apprenants, de l'autre, elle contribue à créer des profils de groupes qui traitent les apprenants comme des objets dotés d'un ensemble d'attributs, plutôt que de les considérer comme des sujets à la fois complexes et singuliers. Et, comme dans d'autres domaines, les profilages réalisés dans le cadre des Learning Analytics peuvent intrinsèquement refléter et perpétuer les stéréotypes et préjugés dans les domaines culturel, géopolitique, économique et social. Certains observent ainsi que, de manière assez paradoxale, la collecte de données sur les apprenants pourrait avoir pour effet d'accroître leur vulnérabilité, plutôt que de la diminuer [Prinsloo, Slade, 2016]. Pour éviter les risques liés à une discrimination implicite ou explicite, il importe donc de s'assurer que les analyses sont effectuées sur des ensembles de données représentatifs. Il importe aussi de ne pas céder à la tentation de croire qu'il serait facile, grâce aux Learning Analytics, de supplanter la rationalité humaine afin de rendre des décisions justes et non dépendantes des partis pris subjectifs. Il importe, surtout, de permettre à l'apprenant (ou celui qui le représente juridiquement) de prouver l'existence de biais et de stéréotypes dans le profilage et de démontrer que des analyses prédictives sont fausses ou incomplètes, lorsqu'une décision lui fait grief.

Instrumentalisation des Learning Analytics et discriminations

On sait que les critiques le plus vives formulées à l'encontre du Big Data mettent en avant le fait que la valeur créée repose sur la réutilisation des données à des fins qui n'avaient pas été prévues lors de leur collecte et qui ne servent pas uniquement à améliorer les services. Certaines entreprises ont déjà essayé d'établir des profils de risques déduits de profils psychologiques établis sur la base des données personnelles collectées sur les réseaux sociaux (risques permettant d'adapter la police d'assurance des jeunes conducteurs ou de mesurer la solvabilité prévisible d'un consommateur, compte tenu de la fréquentation de ses « amis »). Il convient que les données de Learning Analytics ne fassent pas l'objet d'une utilisation mercantile qui dévoierait la visée pédagogique ou didactique initialement assignée à la collecte. Les établissements d'enseignement devraient garantir que les données issues des Learning Analytics ne pourront être utilisées à des fins qui ne sont pas strictement éducatives (exploitation de ces données à des fins publicitaires, établissement de profils de risques destinés à des assureurs ou des établissements de crédits, informations permettant le recrutement par des employeurs…). Car le secteur de l'éducation est fortement sollicité par des acteurs économiques, très intéressés par les données scolaires.

La question de la sécurité des données scolaires s'avère polémique. Le 5 décembre 2018, le ministère de l'Éducation nationale et de la Jeunesse et la Commission nationale de l'informatique et des libertés (CNIL) ont signé une « convention relative à la protection des

33. Source : « Educause Annual Conference 2015. Visites Purdue University & University of Washington », rapport de la délégation française, [En ligne]. Une autre expérience, mise en place par l'université du Maryland (UMUC), tente un projet ambitieux qui, pour diminuer le taux d'échec, utilise un système complexe qui interroge le lycée et le quartier d'origine, en pondérant le risque en fonction de facteurs sociologiques (origine ethnique et sociale de l'étudiant, lieu de vie). Ces indicateurs sont pris à égale considération, voire plus, avec les indicateurs sur l'activité dans le collège (source : « Educause Annual Conference 2016. Visites UCLA, Stanford & Berkeley », rapport de la délégation française, [En ligne].

données personnelles dans les usages numériques de l'Éducation nationale », dont la portée semble pour l'instant limitée, au moins dans le temps (puisqu'elle a été conclue pour une durée de trois ans) [MENJ/CNIL, 2018].

Si l'on peut espérer que les dispositions du RGPD en lien avec la coopération renforcée entre autorités pour les traitements transnationaux permettent d'améliorer les conditions de sécurité du stockage des données, la question de l'hébergement des données et/ou de leur transfert vers un prestataire extérieur semble encore, sur le terrain, un point de fragilité très important. Comme l'observe le rapport « Données numériques à caractère personnel au sein de l'Éducation nationale », « si l'on peut admettre que des données scolaires puissent être transmises sans cryptage au sein des réseaux dédiés de l'*É*ducation nationale, il apparaît plus surprenant qu'elles puissent être transmises "en clair" vers des prestataires extérieurs » [IGEN-IGAENR, 2018, p. 10]. Cette inquiétude semble d'autant plus fondée que la mission a pu observer que, bien souvent, des données inutiles étaient transmises à des tiers. Il convient également d'être particulièrement attentifs aux évolutions technologiques dans le domaine de la portabilité et du transfert des données. Par ailleurs, le constat que le stockage des données scolaires s'effectue sans aucun regard de l'État sur la sécurité des serveurs les accueillants reste extrêmement inquiétant, en regard des menaces toujours plus nombreuses de cyberattaques. Rappelons qu'en décembre 2018, les données d'un demi-million d'étudiants du San Diego Unified School District ont été volées par un hacker [34]. Compte tenu de leur grande valeur, il est à redouter que les données de Learning Analytics ne puissent être réutilisées à l'insu des apprenants si elles venaient à faire l'objet d'un hacking. Exiger une certification ANSSI de premier niveau, au moins pour tous les contractants hébergeant des données de Learning Analytics à caractère personnel, serait judicieux. De manière plus générale, la question de la sécurité et de la souveraineté des données de Learning Analytics demeure prégnante. Les enjeux aussi bien techniques que politiques sont de taille. Or, l'assurance d'une prise en considération sérieuse des enjeux de la sécurité des données, garante de la protection de la vie privée, est une condition essentielle pour l'acceptation et l'utilisation de l'analyse de l'apprentissage par tous ceux qui veillent au respect de la déontologie.

34. Bastien L., « Les données d'un demi-million d'étudiants volées par un hacker », lebigdata.fr, 26/12/2018, [En ligne].

L'analytique des apprentissages avec le numérique

Les Learning Analytics, que nous proposons d'appeler l'**analytique des apprentissages avec le numérique**, sont un terrain d'innovations fertile : deux communautés scientifiques à l'international (EDM et LAK) et une en France (l'Association des technologies de l'information pour l'éducation et la formation – ATIEF) animent la recherche autour d'objectifs communs et d'approches complémentaires. Leur impact scientifique est désormais conforté par des conférences annuelles, la publication de deux revues et, surtout, la création, au sein des plus prestigieuses institutions de l'enseignement supérieur et de la recherche, de groupes de travail, comités de surveillance, centres de recherches chargés de veiller à l'éthique, à la déontologie et aux finalités pédagogiques de la collecte des données et de leurs traitements par des algorithmes.

L'analytique des apprentissages numériques s'attache à révéler, à différents niveaux d'analyse, les informations pertinentes pour améliorer l'expérience et les environnements d'apprentissage. Au service de projets humanistes, cet ensemble de technologies propose d'étayer les acteurs de la communauté éducative. Prédiction de la progression, analyse de l'apprentissage social, analyse de discours, tableaux de bord sont désormais des domaines d'expérimentation bien balisés. En tant qu'analytique des données résultant de l'apprentissage, l'analytique des apprentissages numériques s'apparente au plus vaste ensemble de la science des données de l'apprentissage, dont les questions de recherche vont de l'architecture physique des données aux traces issues de l'activité de l'enseignant lui-même.

Le passage de l'expérimentation à l'industrialisation de ces technologies (Larusson, 2014) est bousculé par l'arrivée impromptue des géants de l'économie numérique. En l'état actuel, la recherche expérimentale se teinte d'un certain scepticisme. Les garde-fous sont plus que jamais nécessaires, s'agissant d'un public pour partie mineur et de données cognitives sensibles. Les équipes pluridisciplinaires de la recherche publique préconisent des dispositifs responsables, transparents, sobres, confidentiels et accessibles. Le futur de l'analytique des apprentissages numériques est une équation à trois inconnues : premièrement, les décideurs (gouvernements, institutions, firmes) prendront-ils la mesure des financements et des projets encore nécessaires à l'adaptation de la recherche aux pratiques de terrain et à la dissémination d'une culture de l'analytique des données au sein du corps enseignant ? Ensuite, ces décideurs parviendront-ils à élaborer un cadre législatif viable, souple mais équitable, respectant le droit des usagers à suspendre la collecte des données et qui préserve leur confidentialité ? Enfin, ces décideurs parviendront-ils, au-delà d'une logique mercantile et de technologies propriétaires, à promouvoir des standards, des projets ouverts ?

Institutions d'enseignement et de recherche, politiques et chercheurs ont la main sur le devenir de l'analytique des données résultant de l'apprentissage numérique : de la connaissance de ce domaine et de la prise de conscience de ses impacts et enjeux sociétaux dépendra le visage de l'École numérique (Griffiths *et al.*, 2015).

BIBLIOGRAPHIE

Aguiar E., Ambrose Ambrose G. A., Chawla N. V., Goodrich V., Brockman J. (2014), « Engagement vs Performance: Using Electronic Portfolios to Predict First Semester Engineering Student Persistence », *JLA*, vol. 1, n° 3, p. 7-33.

Ahn J. (2013), « What can we learn from Facebook activity? Using social learning analytics to observe new media literacy skills », in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, New York, ACM, p. 135-144.

Aleven V., Mclaren B., Roll I., Koedinger K. (2006), « Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor », *International journal of artificial intelligence in education*, vol. 16, n° 2, p. 101-128.

Amershi S., Conati C. (2009), « Combining Unsupervised and Supervised Machine Learning to Build User Models for Exploratory Learning Environments », *JEDM*, vol. 1, n° 1, p. 71-81.

Anderson J. R. (1983), *The Architecture of Cognition*, Cambridge (MA), Etats-Unis, Harvard University Press.

Anderson T. (1996), « The virtual conference: Extending professional education in cyberspace », *International Journal of Educational Telecommunications*, vol. 2, n° 2-3, p. 121-135.

Arnold K. E., Pistilli M. D. (2012), « Course signals at Purdue: using learning analytics to increase student success », in Proceedings of the Second International Conference on Learning Analytics and Knowledge, New York, ACM, p. 267-270.

Baker R., Gowda S. M., Corbett A. T. (2011), « Automatically detecting a student's preparation for future learning: Help use is key », in *Proceedings of the Fourth International Conference on Educational Data Mining*, Eindhoven, Technische Universiteit Eindhoven, p. 179-188.

Baker R., Siemens G. (2012), « Learning Analytics and Educational Data Mining: Towards Communication and Collaboration », in *Proceedings of the Second International Conference on Learning Analytics and Knowledge,* New York, ACM, p. 252-254.

Baker R., Siemens G. (2014), « Educational Data Mining and Learning Analytics », in K. Sawyer (dir.), *Cambridge Handbook of the Learning Sciences*, New York (NY), Cambridge University Press (2e édition), p. 253-274.

Baker R., Yacef K. (2009), « The State of Educational Data Mining in 2009: A Review and Future Visions », *JEDM*, vol. 1, n° 1, p. 3-17.

Balacheff N. (1994), « Didactique et intelligence artificielle », in Balacheff N., Vivet M. (dir.), Didactique et intelligence artificielle, Grenoble, La Pensée Sauvage, coll. « Recherches en didactique des mathématiques », p. 9-42.

Balacheff N. (1995), « Conception, propriété du système sujet/milieu », in R. Noirfalise, M. J. Perrin-Glorian (dir.), Actes de la VIIe École d'été de didactique des mathématiques, Clermont-Ferrand, ARDM de Clermont-Ferrand, p. 215-229.

Balacheff N., Gaudin N. (2002), « Students conceptions: an introduction to a formal characterization », *Cahier Leibniz*, n° 65, p. 1-21.

Barazzutti P.-L., Cordier A., Fuchs B., Crémilleux B., de Runz C. (2016), « Transmute : un outil interactif pour assister l'extraction de connaissances à partir de traces », *Extraction et gestion des connaissances (EGC 2016)*, p. 463-468.

Beal C.R., Qu L., Lee H. (2006), « Classifying learner engagement through integration of multiple data sources », in *Proceedings of the 21st National Conference on Artificial Intelligence*, New York, ACM, p. 151-156.

Beal C.R., Qu L., Lee H. (2008), « Mathematics motivation and achievement as predictors of high school students' guessing and help-seeking with instructional software », *Journal of Computer Assisted Learning*, n° 24, p. 507-514.

Ben-Naim D., Bain M., Marcus N. (2009), « A user-driven and data-driven approach for supporting teachers in reflection and adaptation of adaptive tutorials », in *Proceedings of the Second International Conference on Educational Data Mining*, p. 21-30.

Benzécri J.-P. (1973), L'Analyse des données, Malakoff, Dunod.

Bienkowski M., Brecht J., Klo J. (2012), « The learning registry: building a foundation for learning resource analytics », in *Proceedings of the Second International Conference on Learning Analytics and Knowledge*, New York, ACM.

Blomer J. (2012), NSDL *Metadata Formats*, [En ligne].

Borgatti S. P., Mehra A., Brass D. J., Labianca G. (2009), « Network analysis in the social sciences », *Science*, vol. 323, n° 5916, p. 892-895.

Bourdet J. (2019), « Des données sensibles à protéger », *Data Analytics Post*, [En ligne].

Bowers A. J. (2010), « Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis », *Practical Assessment, Research & Évaluation*, vol. 15, n° 7, p. 1-18.

Bouhineau D., Luengo V., Mandran N., Ortega M., Wajeman C. (2013), « Open platform to model and capture experimental data in Technology enhanced learning systems », *Workshop Data Analysis and Interpretation for Learning Environments*.

Broisin J., Venant R., Vidal P. (2017a), « Lab4CE: a Remote Laboratory for Computer Education », *International Journal of Artificial Intelligence in Education*, vol. 27, n° 1, p. 154-180.

Broisin J., Venant R., Vidal P. (2017b), « Awareness and Reflection in Virtual and Remote Laboratories: the case of Computer Education », *International Journal of Technology Enhanced Learning, Inderscience Publishers,* vol. 9, n° 2-3, p. 254-276.

Brown M. (2012), « Learning Analytic : Moving from Concept to Practice », accessible sur le site Educause, [En ligne].

Brynjolfsson E., Lorin M. H., Heekyung H. K. (2011), « Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? », accessible sur le site *Social Science Research Network,* [En ligne].

Calvo R. A., D'Mello S. K. (2011), *New Perspectives on Affect and Learning Technologies,* New York (NY), Springer.

Campbell J. P., DeBlois P. B., Oblinger D. G. (2007), « Academic Analytics: A New Tool for a New Era », *Educause Review,* vol. 42, n° 4, p. 40–57, [En ligne].

Casado R., Guin N., Champin P.-A., Lefevre M. (2017), « kTBS4LA : une plateforme d'analyse de traces fondée sur une modélisation sémantique des traces », Atelier « Méthodologies et outils pour le recueil, l'analyse et la visualisation des traces d'interaction », ORPHEE-RDV 2017, Font-Romeu, [En ligne].

Cen H., Koedinger K., Junker B. (2006), « Learning Factors Analysis. A general method for cognitive model evaluation and improvement », in Proceedings of the 8th International Conference on Intelligent Tutoring Systems, Berlin/Heidelberg, Springer/Verlag, p. 164-175.

Cen H., Koedinger K., Junker B. (2008), « Comparing two IRT models for conjunctive skills », in 9th International Conference on Intelligent Tutoring Systems, Berlin/Heidelberg, Springer/Verlag, p. 796-798.

Champalle O., Sehaba K., Mille A. (2016), « Observation et analyse de comportements des utilisateurs à base de traces », Revue des sciences et technologies de l'information - Série TSI : Technique et science informatiques, vol. 35, n° 4-5.

Champin P.-A., Mille A., Prié Y. (2013), « Vers des traces numériques comme objets informatiques de premier niveau : une approche par les traces modélisées », Intellectica, n° 59, p. 171-204.

Charleer S., Santos J. L., Klerkx J., Duval E. (2014), « Improving teacher awareness through activity, badge and content visualizations », in Y. Cao, T. Väljataga, J. K. T. Tang, H. Leung and M. Laanpere (dir.), New Horizons in Web Based Learning: Proceedings of the 1st International Workshop on Open Badges in Education, Springer International Publishing, p. 143-152.

Chatellier R. (2018), « Privacy and ethical concerns in Learning Analytics », CNIL/LINC, [En ligne].

Clauset A., Newman M. E. J., Moore C. (2004), « Finding community structure in very large networks », Physical Review E, vol. 70, n° 6.

Clow D. (2012), « The learning analytics cycle: closing the loop effectively », in Proceedings of the LAK 2012, ACM, p. 134-138.

CNIL/LINC (2012), Cahiers IP Innovation & Prospective, n° 1 : « Vie privée à l'horizon 2020. Focus sur des transformations clés au croisement des usages, des technologies et des stratégies économiques. Quel paysage nouveau pour les données personnelles, les libertés et la vie privée ? Protéger, réguler, innover », cahier téléchargeable, [En ligne].

CNIL/LINC (2013), Cahiers IP Innovation & Prospective, n° 2 : « Le corps, nouvel objet connecté. Du Quantified Self à la M-santé : les nouveaux territoires de la mise en données du monde », cahier téléchargeable, [En ligne].

Clow D. (2013), « MOOCs and the funnel of participation », in Proceedings of the Third International Conference on Learning Analytics and Knowledge, New York, ACM, p. 185-189.

Corbett A.T., Anderson J.R. (1995), « Knowledge tracing: Modeling the acquisition of procedural knowledge », User Modelling and User-Adapted Interaction, n° 4, p. 253-278.

Crédoc (2019), « Baromètre du numérique 2019, Enquête sur la diffusion des technologies de l'information et de la communication dans la société française en 2019 », CGE/ARCEP/Agence du numérique, [En ligne].

Data Bridge Market Research (2019), « Education and Learning Analytics Market 2019. Report Highlights the Competitive Scenario By Microsoft, IBM, SAP, Tableau Software, Alteryx, Qlik, Saba Software, SkyPrep Training Software, Information Builders And Others », [En ligne].

D'Mello S., Olney A., Person N. (2010), « Mining Collaborative Patterns in Tutorial Dialogues », *JEDM*, n° 2, p. 2-37.

D'Mello S. K., Graesser A. C. (2012), « Dynamics of Affective States during Complex Learning », *Learning and Instruction*, n° 22, p. 145-157.

Dabbebi I., Iksal S., Gilliot J.-M., May M., Garlatti S. (2017), « Towards Adaptive Dashboards for Learning Analytic: An Approach for Conceptual Design and implementation », in Proceedings of the 9th International Conference on Computer Supported Education, p. 120-131.

Davenport Th. H., Haris J. G., Morison R. (2010), *Analytics at Work. Smarter Decisions, Better Results*. Boston (MA), Harvard Business Press.

Dawson S. (2008), « A study of the relationship between student social networks and sense of community », *Journal of Educational Technology & Society,* vol. 11, n° 3, p. 224-238.

De Liddo A., Buckingham Shum S., Quinto I., Bachler M., Cannavacciuolo L. (2011), « Discourse-centric learning analytics », in *Proceedings of the 1st International Conference on Learning Analytics & Knowledge*, New York, ACM, p. 23-33.

Dekker G., Pechenizkiy M., Vleeshouwers J. (2009), « Predicting students drop out: a case study », in *Proceedings of 2nd International Conference on Educational Data Minin*g, p. 41-50.

Derbel F., Champin P. A., Cordier A., Munch D. (2015), « Authentification d'un utilisateur à partir de ses traces d'interaction », in Treizièmes Rencontres des jeunes chercheurs en intelligence artificielle (RJCIA 2015).

Djouad T., Mille A., Benmohammed M. (2011), « SBT-IM: Système à base de traces-Indicateurs d'interactions Moodle », Conférence EIAH 2011, Mons.

Duval E. (2011), « Attention Please! Learning Analytics for Visualization and Recommendation », in *Proceedings of the 1st International Conference on Learning Analytics & Knowledge*, New York, ACM, p. 9-17.

Dyke G., Lund K., Girardot J-J. (2010), « Tatiana : un environnement d'aide à l'analyse de traces d'interactions humaines », *Technique et science informatique*s (TSI), p. 1179-1205.

Essa A., Ayad H. (2012), « Student success system: risk analytics and data visualization using ensembles of predictive models », in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, New York, ACM, p. 158-161.

Ez-Zaouia M., Lavoué E. (2017), « EMODA: a tutor oriented multimodal and contextual emotional dashboard », in *Proceedings of the Seventh International Learning Analytics Knowledge Conference*, New York, ACM, p. 429-438.

Fayyad U., Piatetsky S. G., Smyth P. (1996), « From Data Mining to Knowledge Discovery in Databases », *AI Magazine*, vol. 17, n° 3, p. 37-54.

Feng M., Heffernan N., Koedinger K. (2009), « Addressing the assessment challenge with an online system that tutors as it assesses », *User modeling and user-adapted interaction*, vol. 19, n° 3, p. 243-266.

Ferguson R. (2009), « The Construction of Shared Knowledge Through Asynchronous Dialogue », Thèse de doctorat, Milton Keynes, The Open University, [En ligne].

Ferguson R., Clow D. (2015), « Consistent Commitment: Patterns of Engagement across Time in Massive Open Online Courses (MOOCs*) »*, JLA, vol. 2, n° 3, p. 55-80.

Ferguson R., Cooper A., Drachsler H., Kismihók G., Boyer A., Tammets K., Martinez Monés A. (2015), « Learning Analytics: European Perspectives », in *Proceedings of the Seventh International Learning Analytics Knowledge Conference*, New York, ACM, p. 69-72.

Fortunato S. (2010), « Community detection in graphs », *Physics Reports*, vol. 486, n° 3-5, p. 75-174.

Fuchs B. (2018), « Focaliser l'extraction d'épisodes séquentiels à partir de traces par le contexte », in 29es Journées francophones d'ingénierie des connaissances, p. 213-227.

Garrison D. R., Anderson T., Archer W. (1999), « Critical inquiry in a text-based environment: Computer conferencing in higher education », *The Internet and Higher Education*, vol. 2, n° 2, p. 87-105.

Gee J. P., Green J. (1998), « Discourse analysis, learning and social practice: a methodological study », *Review of Research in Education*, n° 23, p. 119-169.

Georgeon O. L., Mille A., Bellet T., Mathern B., Ritter F. E. (2012), « Supporting Activity Modeling from Activity Traces », *Expert Systems,* vol. 29, n° 3, p. 261-275.

Goldhaber M. H. (1997), « The Attention Economy and the Net », *First Monday*, vol. 2, n° 4.

Greller W., Drachsler H. (2012), « Translating learning into numbers: A generic framework for learning analytics », *Journal of Educational Technology & Society,* vol. 15, n° 3, p. 42-57.

Griffiths D. (coord.) (2015), « *Visions of the Future. Horizon Report* », [En ligne].

Hayashi C. (1998), « What is Data Science? Fundamental Concepts and a Heuristic Example », in Hayashi C., Yajima K., Bock H.H., Ohsumi N., Tanaka Y., Baba Y. (dir.), Data Science, Classification, and Related Methods, Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Springer Japan, p. 40-51.

Haythornthwaite C., de Laat M. (2010), « Social Networks and Learning Networks: using social network perspectives to understand social learning », in *Proceedings of the 7th International Conference on Networked Learning*, Lancaster University, p. 183-190, [En ligne].

Hershkovitz A, Nachmias R. (2008), « Developing a log-based motivation measuring tool », in *Proceedings of the First International Conference on Educational Data Mining*, p. 226-233.

IGEN-IGAENR (2018), « Données numériques à caractère personnel au sein de l'Éducation nationale », rapport n° 2018-016, [En ligne].

Iksal S. (2011), « Tracks Analysis in Learning Systems: A prescriptive Approach », *International Journal for e-Learning Security (IJeLS)*, p. 3-9.

Iksal S. (2012), « Ingénierie de l'observation basée sur la prescription en EIAH », HDR en Informatique, Université du Maine, [En ligne].

Joksimović S., Dowell N., Skrypnyk O., Kovanović V., Gašević D., Dawson Sh., Graesser A. (2015), « How do you connect? Analysis of Social Capital Accumulation in connectivist MOOCs », in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, New York, ACM, p. 64-68.

Kay J., Maisonneuve N., Yacef K., Reimann P. (2006), « The Big Five and Visualisations of Team Work Activity », in *Proceedings of the International Conference on Intelligent Tutoring Systems*, Berlin, Springer-Verlag, p. 197-206.

Kizilcec R. F., Piech C., Schneider E. (2013), « Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses », in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, New York, ACM, p. 170-179.

Knowles J. E. (2015), « Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin », *JEDM*, vol. 7, n° 3, p. 18-67.

Koedinger K. R., Corbett A. (2006), « Cognitives tutors: Technology bringing learning science to the classroom », in Sawyer R. K. (dir.), *The Cambridge Handbook of the Learning Sciences*, New York (NY), Cambridge University Press, p. 61-78.

Koedinger K., McLaughlin E., Stamper J. (2012), « Automated student model improvement », in *Proceedings of the 5th International Conference on Educational Data Mining*, p. 17-24.

Koedinger K., Pavlik P., Stamper J., Nixon T., Ritter S. (2010), « Avoiding problem selection thrashing with conjunctive knowledge tracing », in *Proceedings of the 3rd International Conference on Educational Data Mining*, p. 91-100.

Kong Win Chang B., Lefevre M., Guin N., Champin P.-A. (2015), « SPARE-LNC: un langage naturel contrôlé pour l'interrogation de traces d'interactions stockées dans une base RDF », [En ligne].

Kovanovic V., Gašević D., Dawson S., Joksimovic S., Baker R. (2015), « Does Time-on-task Estimation Matter? Implications on Validity of Learning Analytics Findings », *JLA*, vol. 2, n° 3, p. 81-110.

Lallé S., Mostow J., Luengo V., Guin N. (2013), « Comparing Student Models in Different Formalisms by Predicting their Impact on Help Success », in *AIED 2013: 16th International Conference on Artificial Intelligence in Education*, Springer, p. 161-170.

Larusson J. A., White B. (dir.), (2014), Learning Analytics: From Research to practice, New York (NY), Springer.

Learning Technology Standards Committee. (2002), « IEEE Standard for learning object metadata », *IEEE Standard*, vol. 1484, n° 1.

Lesnard L., de Saint Pol T. (2006), « Introduction aux méthodes d'appariement optimal », *Bulletin de méthodologie sociologique*, n° 90, p. 5-25.

Levin D. Z., Cross R. (2004), « The strength of weak ties you can trust: The mediating role of trust in effective knowledge », *Management Science*, vol. 50, n° 11, p. 1477-1490.

Long P. D., Siemens G. (2011), « Penetrating the Fog: Analytics in Learning and Education », *Educause Review,* vol. 46, n° 5, p. 30-40 [En ligne].

Loup G., Serna A., Iksal S., George S. (2016), « Immersion and Persistence: Improving Learners' Engagement in Authentic Learning Situations », in Verbert K., Sharples M, Klobučar T., *Adaptive and Adaptable Learning (EC-TEL 2016: 11th European Conference on Technology Enhanced Learning*), Springer, p. 410-415.

Luengo V. Guin N., Bouhineau D. Daubias P., Bruillard E. *et al.* (2019), « HUBBLE, un observatoire des analyses des traces », Rapport de recherche, ANR (Agence nationale de la recherche, France), [En ligne].

Lukarov V., Chatti M. A., Thüs H., Kia F. S., Muslim A., Greven C., Schroeder U. (2014), « Data Models in Learning Analytics », in *DeLFI Workshops*, p. 88-95.

Macfadyen L. P., Dawson S. (2010), « Mining LMS data to develop an "early warning system" for educators: A proof of concept », *Computers & Education*, vol. 54, n° 2, p. 588-599.

Mannila H., Toivonen H., Verkamo A. I. (1997), « Discovery of frequent episodes in event sequences », *Data Mining and Knowledge Discovery*, vol. 1, n° 3, p. 259-289.

Marcelli D., Bossière M.-C., Ducanda A.-L. (2018), « Plaidoyer pour un nouveau syndrome "Exposition précoce et excessive aux écrans" (epee) », *Enfances & Psy*, n° 79, p. 142-160.

Martin T., Aghababyan A., Pfaffman J., Olsen J., Baker S., Janisiewicz P. *et al.* (2013), « Nanogenetic Learning Analytics: Illuminating Student Learning Pathways in an Online Fraction Game », in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, New York, ACM, p. 165-169.

Martinez-Maldonado R., Pardo A., Mirriahi N., Yacef K., Kay J., Clayphan A. (2015), « LATUX: an Iterative Workflow for Designing, Validating and Deploying Learning Analytics Visualisations », *JLA*, vol. 2, n° 3, p. 9-39.

May M., George S., Prévôt P. (2011), « TrAVis to Enhance Online Tutoring and Learning Activities: Real Time Visualization of Students Tracking Data », *International Journal of Interactive Technology and Smart Education (ITSE)*, vol. 8, n° 1, p. 52-69.

MEN/MESRI (2018), « Le numérique au service de l'École de la confiance à l'école », dossier en ligne et version téléchargeable, [En ligne].

MENJ, « CARMO, Cadre de référence pour l'accès aux ressources pédagogiques via un équipement mobile », [En ligne].

MENJ/CNIL, « Convention relative à la protection des données personnelles dans les usages numériques de l'Éducation nationale, [En ligne].

Mercer N. (2004), « Sociocultural discourse analysis: analysing classroom talk as a social mode of thinking », *Journal of Applied Linguistics*, vol. 1, n° 2, p. 137-168.

Michel C., Lavoué E., George S., Ji M. (2017), « Supporting Awareness and Self-Regulation », in Project-Based Learning through Personalized Dashboards, International Journal of Technology Enhanced Learning (IJTEL), vol. 9, n° 2-3, p. 204-226.

Mille A., Pérès-Labourdette Lembé V., « Learning Analytics : vers une éthique par construction des EIAH », [En ligne].

Ming Ming C., Nobuko F. (2014), « Statistical Discourse Analysis: A method for modeling online discussion processes », *JLA*, vol. 1, n° 3, p. 61-83.

Minh Chieu V., Luengo V., Vadcard L., Tonetti J. (2010), « Student modeling in complex domains: Exploiting symbiosis between temporal Bayesian networks and finegrained didactical analysis », *International Journal of Artificial Intelligence in Education,* n° 20, p. 269-301.

Mitrovic A., Martin B., Suraweera P. (2007), « Intelligent Tutors for All: The Constraint-Based Approach », *IEEE Intelligent Systems,* n° 22, p. 38–45.

Miyake N. (1986), « Constructive interaction and the iterative process of understanding », *Cognitive Science*, n° 10, p. 151-177.

Muratet M., Yessad A., Carron T. (2016), « Understanding Learners' Behaviors in Serious Games », *ICWL 2016 - International Conference on Web-based Learning.*

Muslim A., Chatti M. A., Mahapatra T., Schroeder U. (2016), « A rule-based indicator definition tool for personalized learning analytics », in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, New York, ACM, p. 264-273.

Naur P. (1969), « 'Datalogy', the science of data and data processes », Dans A. J. H. Morrell (dir.) *Information Processing 68, Proceedings of IFIP Congress 1968* (Edinburgh, UK, 1968), (vol. 2, *Hardware, Applications*, p. 1383-1387), Amsterdam, North-Holland Pub. Co.

Niemann K., Scheffel M., Wolpers M. (2012), « An overview of usage data formats for recommendations in TEL », in *Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 12)*, p. 95-100, [En ligne].

Niemann K., Wolpers M., Stoitsis G., Chinis G., Manouselis N. (2013), « Aggregating social and usage datasets for learning analytics: data-oriented challenges », in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, New York, ACM, p. 245-249.

Ohlsson S. (1994), « Constraint-based student modeling », NATO ASI Series F Computer and Systems Sciences, n° 125, p. 167-189.

Ohlsson S. (1996), « Learning from performance errors », Psychological Review, n° 103, p. 241.

Pavlik P.I., Cen H., Koedinger K.R. (2009), « Performance Factors Analysis. A New Alternative to Knowledge Tracing », in *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, Amsterdam, IOS Press, p. 531-538.

Peraya D., « Les Learning Analytics en question. Panorama, limites , enjeux et visions d'avenir », *Distances et Médiations des Savoirs*, n° 25, 2019, [En ligne].

Prinsloo P., Slade S. (2016), « Student Vulnerability, Agency, and Learning Analytics: An Exploration », *Journal of Learning Analytics*, vol. 3, n° 1, p. 159-182, [En ligne].

Pham Thi Ngoc D. (2011), « Spécification et conception de services d'analyse de l'utilisation d'un environnement informatique pour l'apprentissage humain », Thèse de doctorat en Informatique, Université du Maine.

Rebolledo-Mendez G., Du Boulay B., Luckin R., Benitez-Guerrero E. I. (2013), « Mining Data From Interactions With a Motivational-aware Tutoring System Using Data Visualization », *JEDM*, vol. 5, n° 1, p. 72-103.

Reffay C., Chanier T. (2003), « How social network analysis can help to measure cohesion in collaborative distance-learning », in *Computer Supported Collaborative Learning* (Bergen, Norway), Kluwer Academic Publishers, p. 343-352.

Réseau Canopé (2018), *Les Données à caractère personnel. Comprendre et appliquer les nouvelles réglementations dans les établissements scolaires*, dossier en ligne et version téléchargeable, [En ligne].

Roberge A. (2013), « Le LMS, un marché en croissance *», article*, [En ligne].

Rosenthal R., Jacobson LF. (1968), « Teacher Expectation for the Disadvantaged », *Scientific American*, vol. 218, n° 4, p. 19-23.

Santos J. L., Govaerts S., Verbert K., Duval E. (2013), « Addressing learner issues with StepUp! An evaluation », in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, New York, ACM, p. 14-22.

Schneider B., Abu-El-Haija S., Reesman J., Pea R. (2013), « Toward collaboration sensing: applying network analysis techniques to collaborative eye-tracking data », in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, New York, ACM, p. 107-111.

Sclater N. (2014), *Learning analytics: The current state of play in UK higher and further education*, Bristol, JISC, [En ligne].

Siemens G. *et. al.*. (2011), *Open Learning Analytics: an integrated & modularized platform Proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques*, SoLAR, [En ligne].

Skrypnyk O., Joksimović S., Kovanović V., Gašević D., Dawson S. (2015), « Roles of course facilitators, learners, and technology in the flow of information of a cMOOC », *The International Review of Research in Open and Distributed Learning*, vol. 16, n° 3, p. 188-217.

Slade S., Prisloo P. (2013), « Learning Analytics : Ethical Issues and Dilemmas », *American Behavioral Scientist*, vol. 57, n° 10, p. 1509-1528, [En ligne].

Snell J., Atkins M., Norris W., Messina C., Wilkinson M., Dolin R. (2012a), « JSON Activity Streams 1.0 », [En ligne].

Snell J., Atkins M., Recordon D., Messina C., Keller M., Steinberg A., Dolin R. (2012b), « Activity Base Schema (Draft) », [En ligne].

Snow E. L., Allen L. K., Jacovina M. E., Crossley S. A, Perret C. A., McNamara D. S. (2015), « Keys to Detecting Writing Flexibility Over Time: Entropy and Natural Language Processing », *JLA*, vol. 2, n° 3 , p. 40-54.

Stamper J.C., Koedinger K.R., Baker R.S.J.d., Skogsholm A., Leber B., Demi S., Yu S., and Spencer D. (2011), Managing the educational dataset lifecycle with datashop », in *Proceedings of the AIED 2011*, Berlin, Springer, p. 557-559.

Tukey J. W. (1977), *Exploratory Data analysis*, Reading (MA), Addison Wesley.

Van Leeuwen A. (2015), « Learning analytics to support teachers during synchronous CSCL: balancing between overview and overload », *JLA*, vol. 2, n° 2, p. 138-162.

Verbert K., Duval E., Klerkx J., Govaerts S. (2013), « Learning analytics dashboard applications », *American Behavioral Scientist*, vol. 57, n° 10, p. 1500-1509.

Walker E. (2012), *Primer on K-20 Education Interoperability Standards*, Washington (DC), Software & Information Industry Association, [En ligne].

Waters A., Studer C., Baraniuk R. (2014), « Collaboration-Type Identification in Educational Datasets », *JEDM*, n° 6, p. 28-52.

Wells G., Claxton G. (dir.), (2002), *Learning for Life in the 21st Century*, Oxford, Blackwell.

Wertsch J. (1991), *Voices on the Mind: ASocio-Cultural Approach to Mediated Action*, Cambridge (MA), Harvard University Press.

Wolpers M., Najjar J., Verbert K. and Duval E. (2007), « Tracking actual usage: the attention meta- data approach », *International Journal of Educational Technology and Society,* p. 1176-3647.

Yessad A., Muratet M., Carron T. (2017), « Aider à l'analyse du comportement d'un apprenant dans les jeux sérieux », EIAH 2017, Strasbourg

Zarka R., Champin P.-A., Cordier A., Egyed-Zsigmond E., Lamontagne L., Mille A. (2013), « TStore: A Trace-Base Management System using Finite-State Transducer Approach for Trace Transformation », *International Conference on Model-Driven Engineering and Software Development (MODELSWARD 2013)*.

Zhu Y., Xiong Y. (2015), « Towards Data Science », *Data Science Journal*, vol. 14, n° 8.

Zimmermann J., Brodersen K. H., Heinimann H. R., Buhmann J. M. (2015), « A Model-Based Approach to Predicting Graduate-Level Performance Using Indicators of Undergraduate-Level Performance », *JEDM*, vol. 7, n° 3, p. 151-176.

Liste des outils étudiés dans la partie 3

Abstract (LIRIS) : [Georgeon *et al.*, 2012]
Activity Streams : [Snell *et al.*, 2012a]
Contextualized Attention Metadata (Fraunhofer FIT) : [Wolpers *et al.*, 2007]
CSVtoxAPI - xAPI Lab (Projet ANR Kolflow)
D3KODE (LIRIS) : [Champalle *et al.*, 2016]
DDART+Reporting tool (LIRIS) : [Michel, *et al.*, 2017]
DisKit (DIScovering Knowledge from Interaction Traces) (LIRIS) : [Fuchs, 2018]
dmt4sp (LIRIS) : [Mannila *et al.*, 1997]
EMODA (LIRIS) : [Ez-Zaouia *et al.*, 2017]
IMS Caliper
kTBS (a kernel for Trace-Based Systems) (LIRIS) : [Champin *et al.*, 2013]
kTBS4LA (LIRIS) : [Casado *et al.*, 2017]
Laalys (Université Pierre et Marie Curie) : [Muratet *et al.*, 2016 ; Yessad *et al.*, 2017]
Lab4ce (IRIT) : [Broisin *et al.*, 2017a ; Broisin *et al.*, 2017b]
LCDM (Learning Technologies)
LEA4AP (Université Pierre et Marie Curie)
Learning Registry (SRI International)
Limesurvey
Méthode statistique de recherche de profils/typologie à partir de descripteurs (proposée dans les logiciels d'analyse statistique)
Méthodes statistiques d'appariement optimal - analyse de séquences (proposée dans les logiciels d'analyse statistique)
NSDL Paradata (Fraunhofer Institute for Applied Information Technology) : [Niemann *et al.*, 2012]
Samotraces (LIRIS) : le projet SAMOTRACES : http://sourceforge.net/ projects/samotraces/
SamoTraceMe (LIRIS) : [Derbel *et al.*, 2015]
SBT-IM (LIRIS) : [Djouad *et al.*, 2011]
SPARE LNC (LIRIS) : [Kong Win Chang *et al.*, 2015]
T-store (LIRIS) : [Zarka *et al.*, 2013]
Taaabs (LIRIS) : Site du projet TAAABS : https://projet.liris.cnrs.fr/sbt-dev/tbs/doku.php/tools:taaabs
Tactiléo Map
TraceMe (LIRIS)
Transmute (LIRIS) : [Barazzutti *et al.*, 2016]
Tatiana (ICAR) : [Dyke *et al.*, 2010]
TRAVIS (LIRIS-LIUM) : [May *et al.*, 2011]
UnderTracks (LIG) : [Bouhineau *et al.*, 2013]
UTL (LIUM) : [Iksal, 2012 ; Iksal, 2011 ; Loup *et al.*, 2016 ; Dabbebi *et al.*, 2017 ; Pham Thi Ngoc, 2011]
xCollector (LIRIS)

POUR L'ÉCOLE
DE LA CONFIANCE

académie Créteil
LIP6
TECHNE
CNED
AU SERVICE DE TOUTES LES RÉUSSITES
CECOJI
Centre d'Études et de Coopération Juridique Interdisciplinaire
EA 1383
UP
LIRIS
LIUM
Laboratoire d'Informatique
Université du Mans
Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE L'ÉDUCATION
NATIONALE ET
DE LA JEUNESSE
RÉSEAU CANOPÉ
CANOPÉ
LE RÉSEAU DE CRÉATION
ET D'ACCOMPAGNEMENT PÉDAGOGIQUES