Données scolaires élèves : étude de l'existant et des possibilités de recueil et d'analyses

Khansa Ghabara[1]

[1] Université de Paris, France
khansa.ghabara@etu.parisdescartes.fr

Abstract. L'article expose les résultats d'une première étude effectuée dans le cadre de travaux de recherche visant à explorer les possibilités de recueil et de traitements conjoints des données disponibles dans les différentes plateformes numériques en éducation scolaire au niveau collège et lycée. Contrairement aux universités qui sont responsables du recueil et du traitement des données produites au sein de leurs plateformes, les institutions scolaires doivent négocier avec une multiplicité d'acteurs, ce qui est à l'origine de nombreuses contraintes notamment pour la disponibilité des données pour les travaux de recherche. Elles sont également confrontées à des interrogations majeures liées au nouveau cadre règlementaire (RGPD) dont les outils et les mécanismes sont en cours de mise en œuvre et d'appropriation. Nous présentons les constructions actuelles autour de ce nouveau cadre règlementaire de protection des données à caractère personnel en éducation scolaire et essayons de répondre aux différentes questions : quelles données sont disponibles ? Qui les détient ? Quelles articulations sont possibles entre les données issues des différentes plateformes utilisées en éducation scolaire ? Quels traitements possibles de ces données ?

Keywords : Données scolaires, Espaces Numériques de Travail, vie scolaire, accès aux ressources

1 Introduction

Nos travaux de recherche visent à explorer les possibilités de recueil et de traitements conjoints des données disponibles dans les différentes plateformes numériques éducatives en éducation scolaire au niveau collège et lycée dans le contexte français. En effet, nous nous interrogeons sur les enrichissements possibles entre les plateformes de type *Espaces Numériques de Travail* (ENT), les outils de vie scolaire et les plateformes d'apprentissage. Comment ces données de type et de niveau de granularité différents peuvent-elles s'articuler et s'alimenter mutuellement, notamment pour améliorer l'apprentissage des élèves et leur suivi ?

Alors que les universités sont responsables du recueil et du traitement des données produites au sein de leurs plateformes, les institutions scolaires ne détiennent pas l'ensemble des données élèves. Elles sont confrontées à de multiples acteurs parallèlement à la mise en œuvre d'un nouveau cadre règlementaire (RGPD), à la fois pour la collecte et les traitements. Deux grandes tendances peuvent être dégagées pour la finalité de ces traitements : une description (1) au plus près du processus d'apprentissage grâce à une analyse fine des actions de l'apprenant, ou (2) orientée vers

les acteurs de l'enseignement, notamment l'institution, pour la mise en œuvre de leviers d'intervention ou de régulation. Le choix des traitements peut être fortement lié au détenteur des données et aux raisons motivant leur exploitation conduisant à différentes interrogations autour des données disponibles, de leurs détenteurs, des articulations possibles entre les données issues des différentes plateformes et des traitements possibles de ces données.

Afin d'y répondre, nous commençons par faire le point sur la typologie des plateformes utilisées en éducation scolaire et sur les constructions en cours autour du nouveau cadre règlementaire de protection des données à caractère personnel.

Nous ferons ensuite un point sur les données disponibles dans ces différentes plateformes en France. Puis, à travers une revue internationale, nous explorons les analyses possibles de ces données et nous évoquerons, enfin, les perspectives de nos travaux en cours.

2 Les plateformes éducatives en éducation scolaire

En France, les données élèves sont produites et stockées dans trois principales plateformes : les Espaces Numériques de Travail (ENT), les applications de gestion de vie scolaire et les plateformes d'apprentissage.

Les *ENT*, au sens du Schéma Directeur des Espaces numériques de Travail pour l'enseignement scolaire, imposé comme cadre de référence par le Ministère de l'Éducation Nationale (SDET 6.3, 2019)[1], sont des projets territoriaux portés soit par une Académie (Toutatice, i-Cart, …) soit par une collectivité territoriale, département ou région (c'est le cas de la majorité des projets), implémentés par des acteurs privés et mis à disposition des établissements. Selon les plateformes, des « bouquets de service » sont fournis, en mode « intégré » et d'autres sont proposés via des connecteurs permettant l'interfaçage avec des applications externes sans réauthentification de la part de l'utilisateur.

Les *applications de gestion de vie scolaire et de vie de l'établissement* permettent la gestion des évènements de vie scolaire (absences, retards, punitions et sanctions disciplinaires) ou de la vie de l'établissement (emploi du temps et enseignements). Elles peuvent être dérivées des applications SIECLE (Système d'Information pour les Élèves des Collèges, des Lycées et pour les Établissements du Ministère de l'Éducation Nationale) et mises à disposition des établissements. Les établissements peuvent aussi faire le choix d'une solution privée. Deux d'entre-elles, Index-Education[2] et Axess Education[3], se trouvent en situation de quasi-duopole à l'échelle nationale (Cour Des Comptes, 2019). Ces différentes applications peuvent être intégrées dans les ENT mais bien souvent elles sont externes à ces derniers.

Les *plateformes d'apprentissage* sont des espaces d'interaction enseignant-élève offrant souvent des services supplémentaires tels que la gestion des résultats des tests et des réussites des élèves, des suggestions de nouveaux exercices ou de nouvelles

[1] SDET 6.3 en vigueur : https://eduscol.education.fr/cid56994/sdet-version-en-vigueur.html
[2] https://www.index-education.com/fr/
[3] http://www.axess-education.fr/

ressources, des forums de discussion, des tableaux de bord pour les enseignants et/ou élèves, etc. (Baron & Bruillard, 2018).

Ces plateformes sont mises à disposition des établissements, gratuitement ou moyennant abonnement, par des acteurs variés pouvant être des institutions publiques, ou des éditeurs privés.

Le choix des plateformes utilisées se décide à l'échelle de l'établissement et/ou de l'enseignant. L'accès à ces plateformes peut se faire via les ENT. Mais bien souvent cet accès échappe au cadre institutionnel mis en place et se fait via une connexion directe sur la plateforme de l'éditeur en question.

3 Mise en place du RGPD[4] : le GAR et les délégués aux données

La construction et surtout l'appropriation d'un cadre organisationnel, juridique et technique autour de l'exploitation des données scolaires nécessitent un temps long. En attendant, des doutes, des hésitations, des incompréhensions persistent et rendent davantage complexe l'accès aux données.

D'abord, s'agissant de l'outillage technique, le GAR (Gestionnaire d'Accès aux Ressources) est mis en œuvre par la Direction du Numérique pour l'Education. Il fournit un service d'accès non visible par les utilisateurs en établissements leur permettant d'accéder, via leur ENT, aux ressources qui leur sont affectées en transmettant aux fournisseurs de ressources, les attributs strictement nécessaires à leur fonctionnement.

Depuis le « Médiacentre », module intégré à l'ENT permettant d'afficher une liste personnalisée de ressources, l'élève clique sur le lien d'accès à la ressource. L'ENT sollicite alors le GAR pour qu'il mette l'utilisateur en relation avec la ressource.

Le GAR a la spécificité de regrouper l'ensemble des acteurs fournisseurs d'identité, fournisseurs d'attributs et fournisseurs de services autour d'accords réglementaires, juridiques, organisationnels et techniques qui garantissent la sécurité de la gestion des identités, de la gestion des autorisations et du contrôle d'accès. On parle ainsi d'accords de fédération des données.

Depuis sa mise en œuvre, le GAR gère la transmission des données à caractère personnel des élèves et des enseignants dans un cadre respectant la législation en matière de RGPD. Le ministère contractualise avec les fournisseurs qui doivent justifier, dans le cadre d'un suivi de « la conformité applicative » de leurs ressources, de leurs besoins de données à caractère personnel, en particulier quand elles permettent d'identifier directement ou indirectement l'individu. Le GAR ne transmet que les données strictement nécessaires au bon fonctionnement de la ressource et ayant fait l'objet d'une demande justifiée et validée par le Ministère (voir MENJ, 2019, pour un bilan CNIL du déploiement du GAR).

Les institutions scolaires, à l'échelle nationale ou locale, tentent de construire un cadre organisationnel facilitant la prise en compte et l'appropriation du cadre règlementaire en vigueur. Outre le Délégué à la protection des données au ministère et les Délégués à la protection des données en académie, il a été nommé un administrateur ministériel des données et un Comité d'éthique pour les données d'éducation a également été constitué.

[4] Règlement général sur la protection des données

4 Données disponibles : localisation, accès et flux

En France, les données issues des différentes plateformes utilisées en enseignement scolaire sont l'objet de travaux visant à explorer les traces d'activités des élèves et des enseignants sur les plateformes éducatives, à développer des méthodologies d'analyse de ces données ou à concevoir des tableaux de bord à destination des différents acteurs (élève, enseignant, institution) (Luengo et al., s. d., 2018), et à mettre en place un protocole de collecte et de stockage sous forme d'entrepôt de données (Boyer, 2019).

Nous avons choisi de situer notre recherche dans la lignée des travaux portés par la Direction du Numérique pour l'Éducation (DNE[5]), et notamment le projet de recherche LOLA qui vise la mutualisation, au sein de la communauté de recherche française, des corpus de données collectées, des modèles et des outils de visualisation déjà développés en open source, des indicateurs spécifiques de chaque projet et des documents d'accompagnement. Cette mutualisation passerait par la mise à disposition de la communauté scientifique d'un entrepôt de données qui serait également ouvert aux institutions éducatives qui s'interrogent sur l'exploitation des données scolaires et aux équipes pédagogiques qui expérimentent dans un contexte réel (Boyer, 2019). Néanmoins, le projet LOLA, reposant sur les travaux conduits dans le cadre du projet METAL (Boyer, 2019), malgré des objectifs ambitieux, rencontre depuis son lancement des difficultés persistantes liées à l'accès aux données. Les travaux se sont ainsi limités à l'exploitation de données simulées et à la conception d'un prototype d'entrepôt de données.

Afin de faire le point sur les données disponibles en éducation scolaire, nous reprenons les trois types de plateformes précisées plus haut et, pour chacune d'entre elle, nous ferons un focus sur la typologie des données disponibles, sur les acteurs clés pour un éventuel processus de recueil et sur les flux entrants ou sortants depuis/vers les autres plateformes.

4.1 Espace Numérique de Travail

Données disponibles : voir figure 1 qui répertorie l'ensemble des données élèves disponible au sein d'un ENT

Accès : Ces données sont gérées par l'éditeur de la solution sur sa propre plateforme. Le traitement associé est sous la responsabilité du chef d'établissement.

Flux : Une grande partie des données disponibles au sein des ENT est issue du Système d'information du ministère de l'Éducation Nationale (SIECLE) qui centralise les données sur l'identité et les coordonnées des élèves et de leurs responsables, sur la scolarité actuelle de l'élève (établissement, formation, classe, disciplines en option, répartition en groupes, redoublement, hébergement, bourses, circuit de transport) et sur les scolarités antérieures de l'élève. On y trouve, également, les évaluations (par notes et par compétence), les attestations et les diplômes obtenus par l'élève.

Elles sont renseignées par les chefs d'établissements et enrichies par les acteurs locaux (collectivité) ou nationaux (Ministère). D'autres flux d'initialisation et/ou de mise à jour des données sont également possibles par le biais d'imports/Exports

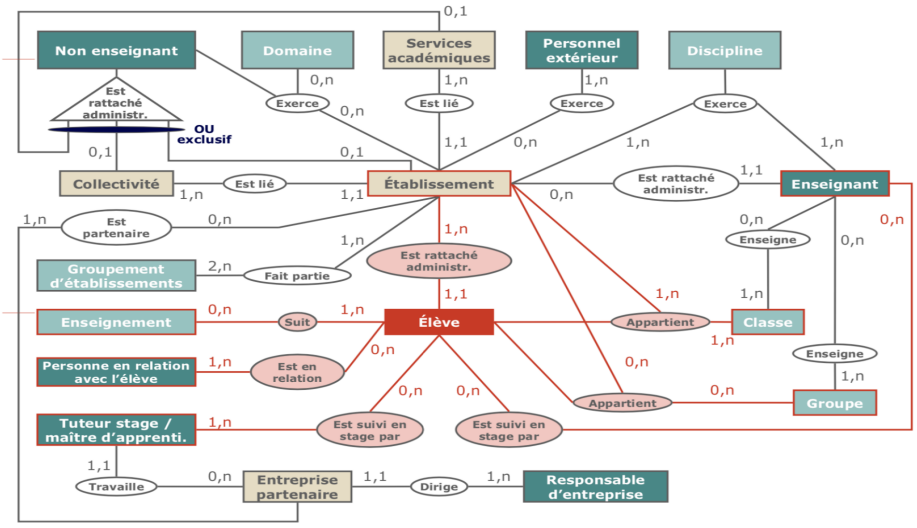
[5] https://www.education.gouv.fr/direction-du-numerique-pour-l-education-dne-9983

depuis/vers les applications de gestion de vie scolaire et de vie d'établissement, via les applications STS Web, SIECLE Bee ou encore « LSU ».[6]

Les bases élèves-établissement de SIECLE (SIECLE Bee) sont gérées au niveau local par le rectorat, pour qui la Délégation Académique au Numérique pour l'Éducation (DANE) coordonne les problématiques de droit d'accès et de protection des données personnelles.

Une partie limitée des données du Système d'Information du Ministère alimentent les ENT, en passant par les instances académiques des Annuaires Académiques Fédérateurs (AAF). Outre le vecteur d'identité [Profil (élève), Nom de l'élève, Prénom de l'élève, Numéro d'enregistrement de l'élève dans SIECLE, Code UAI de l'établissement de l'élève dans l'annuaire ENT][7], sont également transmises : des données sur l'organisation de la scolarité de l'élève (Modules d'enseignement, niveau de formation, établissements et classes associée, personnes en relations avec l'élèves, etc. Cependant, il ne s'agit pas d'une configuration unique sur l'ensemble du territoire car de nombreux ENT sont, aujourd'hui, alimentées par le logiciel d'emploi du temps. C'est bien souvent laissé à la décision du chef d'établissement.

Fig. 1. Caractérisation d'un élève dans l'ENT (Source : SDET 6.3, 2019)



4.2 Applications de gestion de vie scolaire et de vie de l'établissement

Données disponibles : Ces solutions permettent la mise en relation paramétrée dans le temps (date, horaires) et l'espace (lieu, numéro de salles, ...) des bases « élèves »,

[6] SIECLE Livret Scolaire Unique

[7] A partir de septembre 2020, ce vecteur sera remplacé par un vecteur transmis par EduConnect, le futur fournisseur unique d'identité de l'éducation nationale, qui sera enrichi avec le code BCN de l'Académie de l'élève, le degré, l'identifiant de l'élève dans l'Annuaire Fédérateur et la valeur chiffrée de l'identifiant interne EduConnect, commune à l'ensemble des projets ENT.

« enseignants », « matières » et « groupes ». Certaines solutions, comme *Pronote* (Index-Education), peuvent également gérer les données liées aux résultats scolaires ou encore à l'accès aux ressources.

Accès : En dehors des flux d'initialisation et de mise à jour, ces données sont gérées par l'éditeur de la solution sur sa propre plateforme. Le traitement associé est sous la responsabilité du chef d'établissement.

Flux : Les flux d'initialisation et/ou de mise à jour des données sont matérialisés par des imports/exports entre ces solutions et les bases SIECLE, via les applications STS Web et SIECLE Bee ou le module NetSynchor.

Un flux supplémentaire est également possible pour les solutions intégrant des services d'évaluation des élèves. C'est notamment le cas de la solution *Pronote* qui permet d'exporter les données d'évaluation présentes dans son applicatif « Bulletins » vers l'application « LSU » de SIECLE.

4.3 GAR et plateformes d'apprentissage

L'ENT transmet au GAR, l'ensemble des données des élèves, enseignants, groupes, enseignements, établissements et responsables d'affectation.

Un identifiant pérenne propre au GAR permet d'identifier de manière unique un élève par rapport à un projet ENT.

Dans le cas d'une demande d'accès à une ressource, le GAR vérifie les autorisations d'accès à la ressource pour l'élève concerné puis transmet au fournisseur de ressources un identifiant Opaque, résultant d'une jointure « élève-ressource » ainsi que les attributs nécessaires au fonctionnement de cette dernière et ayant fait l'objet d'une demande justifiée et validée par le Ministère dans le cadre d'une procédure appelée « conformité applicative des ressources » mise en œuvre dans le cadre du projet GAR. Voici les données susceptibles d'être transmis par le GAR pour un élève (RTFS GAR 3.2, 2019)[8] : Code établissement, Code projet ENT, ID Opaque, profil, Division (s), Groupe (s), Degré(s) d'enseignement, Cycle de scolarité, Dispositif(s) de formation, Niveau(x) de formation, Civilité, Nom d'usage, Prénom usuel, IDC et LRA (attributs complémentaires pour une famille de ressources)

L'ensemble de ces attributs est sous la responsabilité du Ministère de l'éducation national. Cette responsabilité a été récemment élargie à l'hébergement des données produites par les utilisateurs au sein des plateformes d'apprentissage connectées au GAR. Les éditeurs conservent les données d'interaction sur leurs propres plates-formes mais ce traitement est désormais sous la responsabilité du ministère. Pour une éventuelle analyse conjointe des données issues des ENT, des outils de vie scolaire et des plateformes d'apprentissage, le GAR pourrait constituer l'élément clé permettant d'établir le lien pour les identités des individus.

Ce que nous venons de voir montre la complexité de la gestion des données élèves et de leurs interactions dans les plates-formes de formation. Si on peut récupérer les données, quels types de traitement pourrait-on faire ? La littérature internationale donne des pistes intéressantes.

---

[8] Référentiel Technique et fonctionnel de Sécurité, https://gar.education.fr/documentation-/

5 Quelles analyses possibles ? revue internationale

Les questions liées à l'exploitation des données d'apprentissage sont, aujourd'hui des champs de recherches fertiles, explorées par plusieurs communautés de chercheurs en sciences humaines et sociales et en informatique.

Selon une étude publiée en 2016 par H. Labarthe et V. Luengo, EDM (*Educational Data Mining*) et SoLAR (*Society for Learning Analytics Research*) qui est à l'origine du LAK (*Learning Analytics & Knowledge*) sont les deux communautés les plus actives dans le domaine. EDM et LAK partagent les mêmes enjeux et adoptent des approches différentes mais complémentaires.

Il s'agit, pour ces deux communautés de déployer des solutions, méthodes et techniques aidant à révéler à différents niveaux d'analyse les informations pertinentes pour améliorer les environnements d'apprentissage au service de l'apprentissage ou de la gestion de ce dernier.

EDM puise ses sources dans le *Knowledge Discovery in Database* qui s'appuie sur des méthodes et techniques de fouille de données (Visualisation, classification, régression, modélisation des dépendances, ...). Les premières applications étaient notamment en marketing où, à titre d'exemple, les algorithmes, permettant le profilage des clients et la prédiction de leurs comportements, ont prouvé leur efficacité (Fayyad et al., 1996).

EDM cultive sa spécificité en priorisant la conception de nouveaux algorithmes touchant des micro-concepts tels que l'apprentissage du calcul. Il s'agit essentiellement d'algorithmes de prédiction ou d'exploration de stratégies d'apprentissage basés sur la recherche de modèles enfouis dans les données.

La communauté SoLAR adopte, quant à elle, une approche itérative et interactive visant à transmettre la modélisation et la visualisation des données aux acteurs de l'apprentissage (apprenants, enseignants, personnels de l'éducation).

« C'est là, la différence fondamentale entre EDM, dont le produit de la recherche alimente une machine, et SoLAR qui vise à amplifier le rôle décisionnel des acteurs de l'apprentissage » (Labarthe, Luengo 2016).

EDM et LAK mettent en œuvre plusieurs méthodes d'analyse des données :

L'analyse prédictive où sont utilisées des techniques classiques de classification et de régression, pensées soit pour la mise en œuvre d'alerte ou d'une typologie d'intervention soit pour l'estimation latente de connaissance.

La découverte de structures qui fait appel aux techniques de clustering et d'analyse factorielle, mais également à celles de l'analyse de réseaux, utilisée notamment pour étudier les travaux de groupes et en construire des indicateurs tels que le niveau d'engagement des apprenants.

La fouille de relation où la technique de découverte de patterns permet d'élaborer des systèmes de suggestion adaptés aux apprenants.

L'analyse automatisée des données textuelles (text mining, analyse du discours) ;

L'interprétation visuelle qui donne lieu à des conceptions de tableaux de bord spécifiques à chaque type d'acteurs.

LAK et EDM constituent un écosystème de méthodes et de techniques complémentaires qui visent, en premier lieu, la collecte et le traitement de données produites dans un contexte machine, et en deuxième lieu la détermination du sens de

ces données, leur interprétation et leur mise en forme pour répondre aux besoins des acteurs concernés et s'adapter à leurs usages.

6 Perspectives

Les premiers travaux présentés ici ont permis de faire le point sur la typologie de données disponibles dans les plateformes éducatives de type ENT et les *solutions* de vie scolaire et de mettre en évidence les flux échangés notamment ceux en direction des plateformes d'apprentissage. Nous n'avons par contre, pas encore, exploré les données produites dans ces ressources éducatives. Nous privilégierons, pour cela, les ressources liées à la discipline de mathématiques : les suites de calcul ou de transformations d'expression se prêtent bien à des analyses automatiques, facilitant l'élaboration de modèles. Par ailleurs, le GAR permet d'identifier les ressources utilisées et permet, également, via le processus de conformité applicative mis en œuvre, d'identifier les traitements effectués sur les données produites par les élèves et les enseignants dans chaque ressource et de cibler ainsi celles qui comportent des interactions. Nous avons identifié la ressource « Labomep[9] » comme première cible potentielle d'analyse.

Afin de pouvoir envisager la construction de mécanismes de fouille articulant entre elles des données issues de chaque type de plate-forme, il est aujourd'hui important, pour nous, de disposer d'un entrepôt de données garantissant les conditions de sécurité et de qualité nécessaire à ce type de traitement.

Références

1. Boyer, A, (2020). Quelques réflexions sur l'exploration des traces d'apprentissage, Distances et médiations des savoirs [En ligne], Consulté le 27 avril 2020. URL : http://journals.openedition.org/dms/4086
2. Bruillard É., Baron G. L. (2018), Researching the design and evaluation of information technology tools for education, in J. Voogt, G. Knezek, R. Christensen, K. W Lai. (eds), Second Handbook of Information Technology in Primary and Secondary Education, New York, Springer. En ligne : https://doi.org/10.1007/978-3-319-53803-7_79-1 .
3. Cours des Comptes (2019). *Le service public numérique pour l'éducation : un concept sans stratégie, un déploiement inachevé*. La Documentation Française, disponible sur https://www.ccomptes.fr/fr/publications/le-service-public-numerique-pour-leducation
4. MENJ (2019). *Bilan CNIL - GAR après deux ans de déploiement. Principes et orientations de la protection des données à caractère personnel*. En ligne, novembre 2019 https://gar.education.fr/medias/fichier/gar-a4-36pages-20191115_1574023611376-pdf
5. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37, 37. https://doi.org/10.1609/aimag.v17i3.1230
6. Labarthe, H., & Luengo, V. (2016). L'analytique des apprentissages numériques [Research Report]. Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-01714229
7. Luengo, V., Guin, L. N., Bouhineau, L. D., Daubias, I. P., Bruillard, S. E., Iksal, L. S., & Kuzniak, O. R. (s. d.). *Hubble, un observatoire des analyses des traces d'apprentissage* ANR 14 CE24 0015 (2015-2018).

[9] https://labomep.sesamath.net/