Plateforme des données de l'éducation : rapport de préfiguration

Federica Minichiello, Nicolas Roussel, Philippe Ajuelos

Plateforme des données de l'éducation : rapport de préfiguration

Federica Minichiello & **Nicolas Roussel**
Cheffe de projet & Directeur scientifique
Programme Éducation et Numérique
Inria

Philippe Ajuelos
Administrateur ministériel des données, des algorithmes et des codes sources
Direction du numérique pour l'éducation
Ministère de l'éducation nationale et de la jeunesse

22 novembre 2023

Table des matières

1 - Synthèse

L'expression « plateforme des données de l'éducation » est utilisée dans ce document pour décrire un ensemble de moyens numériques permettant à des acteurs publics ou privés de collecter, héberger, traiter, cataloguer, partager et réutiliser des données scolaires dans un cadre de confiance préalablement établi. Les « données scolaires » sont ici celles décrites sous ce terme par le rapport IGEN-IGAENR n° 2018-016, i.e. toutes les données pouvant être produites ou collectées dans le cadre scolaire, telles que :

- Des informations administratives sur les élèves et leurs parents (e.g. établissements et classes d'affectation, filière et spécialités, disciplines, ville de résidence, catégories socioprofessionnelles) les enseignants et personnels administratifs, les intervenants extérieurs, etc. ;
- Des « traces d'apprentissage », des données produites par des instruments physiques, des logiciels ou des personnes dans le cadre d'un processus d'apprentissage et qui documentent ce processus : données de capteurs, enregistrements d'interactions, ressources consultées et temps de consultation, réponses fournies à des questions et temps passé pour répondre, productions, annotations par des observateurs, etc. ;
- Des productions d'élèves ou de professeurs : productions écrites et supports d'activités réalisés par les élèves, supports pédagogiques produits par les enseignants, etc. ;
- Des données de vie scolaire : emploi du temps, absences, retards, sanctions, passages en infirmerie, demi-pension, notes et bulletins, cahiers de texte, etc. ;
- Des données d'utilisation des logiciels de vie scolaire et des environnements numériques de travail ;
- Des données des collectivités concernant la connectivité, les équipements informatiques, les ressources et manuels numériques, etc. ;
- Des données concernant l'orientation (e.g. données de Parcoursup) ;
- Des données statistiques telles que celles produites par la Direction de l'Évaluation, de la Prospective et de la Performance (DEPP) et les services statistiques académiques.

1.1 Contexte et origines du projet de plateforme des données de l'éducation

Ce projet tire en grande partie son inspiration d'initiatives portées, soutenues ou suivies par la Direction du Numérique pour l'Éducation (DNE) du MENJ, telles que les projets Hubble (ANR, 2015-2018), METAL (e-FRAN, 2016-2021) et LOLA (depuis 2018, LORIA+DNE), le groupe thématique « Analyse des traces d'apprentissage / Learning analytics » (2017-2018), ou les projets lauréats du Partenariat d'innovation intelligence artificielle (depuis 2019), par exemple.

Les recherches menées dans le secteur du numérique éducatif ont mis en évidence les potentialités majeures offertes par le recueil, l'analyse et le traitement des données scolaires. Ces recherches ont cependant aussi montré les difficultés causées par l'hétérogénéité des sources de données, l'absence de catalogues, l'accès parfois difficile aux données, et les enjeux juridiques, éthiques et de souveraineté liés à leur collecte et à leur exploitation. Ce projet vise à mettre en œuvre les moyens nécessaires pour résoudre ces difficultés afin que l'ensemble des acteurs puissent se saisir des opportunités offertes pour transformer en profondeur le système éducatif au bénéfice des élèves et de leurs parents, des enseignants et cadres de l'enseignement, des décideurs, de la filière EdTech et de la recherche.

Une première étude préparatoire conduite d'avril à juillet 2020 par la DNE a permis de confirmer, à travers une douzaine d'entretiens avec des partenaires du ministère (scientifiques, entreprises EdTech ou autres) le diagnostic sur la situation actuelle : des données dispersées, des impératifs de sécurité difficiles à gérer, des échanges encore rares, entraînant un déficit de valorisation. Cette étude a également permis de commencer à lister des données qui pourraient être mieux exploitées (traces d'apprentissage, données d'organisation et de vie scolaire, données d'examens et de concours, données d'orientation, données d'infrastructure numérique, données de gestion, statistiques, etc.) et des nouveaux usages qui pourraient en être faits.

C'est sur ces bases qu'a été proposée en novembre 2020 la création d'une plateforme des données de l'éducation dans le cadre des États généraux du numérique pour l'Éducation, pour « faire avancer la recherche en éducation, éclairer les décisions, construire des services plus performants ».

Une deuxième étude préparatoire conduite de mai à juillet 2021 par la DNE et Inria a permis de préciser un peu plus l'ambition, la cible et la trajectoire du projet. L'ambition du projet se déclinait à ce stade autour de quatre axes :

- Soutenir la constitution/consolidation du patrimoine de données, à travers la promotion de standards d'interopérabilité et de briques logicielles les mettant en œuvre, la proposition de clauses juridiques types ou la production de jeux de données, par exemple ;
- Jouer le rôle de tiers de confiance dans l'ouverture et le partage de ces données ;
- Accompagner l'écosystème éducatif au sens large dans le développement de nouveaux usages des données partagées (visualisations, analyses, exploitation par des méthodes d'IA, etc.), à travers notamment des projets pilotes ;
- Assurer le lien avec les élèves, enseignants et citoyens, à travers l'information, une acculturation progressive aux usages des données et la co-construction de services.

Depuis, le projet a été inscrit dans la stratégie nationale d'accélération pour l'enseignement et le numérique (comme volet du PEPR « Enseignement et numérique »), dans la feuille de route 2021-2023 sur la politique de la donnée du MENJ, dans sa stratégie du numérique pour l'éducation 2023-2027 et dans sa doctrine technique en cours d'élaboration.

1.2 Travail de préfiguration

Le projet de plateforme des données de l'éducation est entré en novembre 2022 dans une phase de préfiguration associant de manière beaucoup plus importante les acteurs de l'écosystème à travers un nombre conséquent d'auditions avec l'ambition de :

- Comprendre la situation actuelle pour les différents acteurs cités : quelles sont les données produites, quelles sont celles partagées, quelle est leur utilité ? ;

- Identifier des besoins ou désirs de ces acteurs qui ne peuvent être aujourd'hui satisfaits, les freins ou verrous associés, et ce qui permettrait de les lever : qu'est-ce qui limite la production, le partage ou la réutilisation de données ? Quels services faudrait-il mettre en place, dans un monde idéal ? ;
- Délimiter sur ces bases et des contraintes additionnelles — réglementaires, par exemple — les contours de la plateforme souhaitée, à long terme, en matière de services ;
- Identifier les briques technologiques disponibles à court terme et les verrous technologiques ou scientifiques à lever pour pouvoir mettre en œuvre l'ensemble des services souhaités ;
- Définir un programme de mise en œuvre échelonné permettant de fournir rapidement les services qui peuvent l'être et d'initier les développements technologiques ou recherches nécessaires aux autres ;
- Identifier des projets pilotes sur lesquels la plateforme pourrait être testée à mesure de sa mise en œuvre.

Ce travail de préfiguration a été mené conjointement par le MENJ et Inria.

Des auditions menées jusqu'en juin 2023 ont permis de rencontrer 231 personnes représentatives des contributeurs et utilisateurs potentiels de la plateforme (enseignants et cadres de l'éducation nationale, entreprises du secteur EdTech, collectivités, scientifiques, etc.), des acteurs proposant ou ayant mis en œuvre de potentielles solutions techniques (hébergement de données, environnements logiciels, moyens de calcul, etc.) ainsi que des directions, services et autorités administratives (DINUM, ANSSI, comité d'éthique des données et CNIL, notamment).

Les auditions menées confirment la pertinence et le bien-fondé du projet dans la perspective des quatre grandes ambitions citées précédemment (soutenir la constitution/consolidation du patrimoine de données ; jouer le rôle de tiers de confiance dans l'ouverture et le partage de ces données ; accompagner l'écosystème éducatif dans le développement de nouveaux usages des données partagées ; assurer le lien avec les élèves, enseignants et citoyens). La nature et la raison d'être de cette plateforme sont néanmoins comprises de manières différentes selon les personnes interrogées et doivent donc être clarifiées.

Les perspectives, préoccupations et intérêts des personnes interrogées sont divers, tout comme le niveau de maturité de leur expression de besoin. Deux grands besoins sont toutefois largement partagés :

- Pouvoir comprendre ce qui est et ce qui est fait — Le pilotage du système éducatif s'appuie sur des données insuffisantes en quantité et qualité, dispersées entre acteurs et difficiles à consolider aux différents niveaux d'un territoire éducatif. Un exemple typique concerne le suivi du déploiement et de l'impact des équipements et ressources numériques destinés aux élèves, alors que les dépenses publiques dans ce domaine sont importantes et en constante progression [1]. Sur ce sujet comme sur d'autres, l'idée n'est pas d'aller vers un pilotage automatique par la donnée mais vers une aide à la décision et à l'évaluation s'appuyant sur des données plus facilement obtenues à travers la plateforme (dans l'exemple, des données issues du GAR, des systèmes de gestion internes aux académies, des ENT des écoles et EPLE — avec l'indispensable concours de leurs éditeurs et des collectivités).

1. Voir le rapport « Le service public numérique pour l'éducation » de la Cour des comptes sur la période 2013-2017, par exemple

- Se donner les moyens d'innover — Les capacités des différents acteurs à collecter des données couplées aux avancées en science des données et intelligence artificielle laissent entrevoir de nombreuses possibilités pour aider les enseignants dans les tâches d'évaluation, pour faciliter l'organisation des enseignements et des temps périscolaires, pour améliorer la détection anticipée des décrochages, pour une orientation, une éducation et un accompagnement adaptés et personnalisés des apprenants, etc. Les idées de choses à améliorer ou à créer ne manquent pas. La plateforme pourrait aider à réaliser les potentiels envisagés en facilitant l'accès à des données à des fins d'exploration, de prototypage et d'expérimentation. Il ne s'agit pas de tomber dans une forme de data-angélisme (l'accès à des données ne garantit en rien que quelque chose « de bon » pourra en être tiré) mais de lever un certain nombre de freins actuels à l'innovation pour augmenter les chances que ce qui a du sens et est techniquement possible puisse être réalisé.

Le principal enjeu identifié est celui de la plus simple circulation des données. Recenser les données existantes ne change pas nécessairement grand-chose à leur disponibilité pratique. Les rassembler dans un entrepôt sécurisé ne facilite pas nécessairement leur exploitation. L'objectif doit être de faciliter la circulation de données, jusque là où elles peuvent être utilisées pour répondre à des questions et besoins pré-identifiés ou pour expérimenter de nouvelles choses. Une plateforme de données n'est pas une fin en soi mais un moyen, sa valeur ne réside que dans ce qu'elle permet de faire de nouveau.

Les personnes interrogées ont exprimé des attentes élevées envers le ministère concernant l'orchestration de la gestion des données notamment en termes de normalisation et de définition de standards (doctrine technique) et d'équilibre public-privé. Ces attentes reflètent la confiance placée dans l'État dans son rôle de garant du système éducatif et de la protection des élèves, des enseignants et de toute la communauté éducative. Face aux questions complexes liées à la sécurité, à la confidentialité et aux responsabilités juridiques, par exemple, il convient d'adopter une approche politique réfléchie pour répondre aux attentes, sans décevoir, tout en respectant les cadres légaux et éthiques.

Les auditions ont aussi mis en évidence le besoin de portage politique et de cohérence avec les autres actions de l'État sur le numérique pour l'éducation. Comme rappelé précédemment, le projet de plateforme des données est inscrit depuis plusieurs années dans la vision stratégique portée par le ministère. La continuité du portage politique d'un tel projet est importante pour que l'ensemble des acteurs s'en saisissent (les directions ministérielles, notamment) et qu'il puisse être articulé avec d'autres initiatives (le programme IDEE ou la doctrine technique en cours d'élaboration par le ministère, par exemple).

Les auditions ont surtout permis de discuter avec des acteurs de terrain de ce que devaient être les grands principes fondateurs d'une telle plateforme.

1.3 Conclusions et perspectives

Le travail de préfiguration s'est terminé par la rédaction d'un rapport transmis à la direction du numérique pour l'éducation du MENJ et à la direction générale d'Inria fin juillet et discuté lors d'un comité de pilotage conjoint mi-octobre.

A travers la diversité des perspectives, préoccupations, intérêts et niveaux de maturité des personnes auditionnées, la plateforme de données apparaît comme un objet protéiforme, source d'imaginaires variés et présentant des risques de dispersion des efforts et de déceptions futures.

Certaines attentes exprimées pourront être comblées à court/moyen terme par différentes initiatives déjà lancées. Le programme IDEE devrait grandement simplifier l'accès

des scientifiques aux données de la DEPP. La doctrine technique en cours d'élaboration par le ministère devrait aussi faciliter la circulation des données entre acteurs publics et privés dans une logique de plateforme, à travers un cadre d'architecture et des règles et standards communs. Le ministère travaille également à la mise en œuvre d'un tableau de bord partagé pour un meilleur pilotage du numérique éducatif et envisage une plateforme d'échange dédiée aux collectivités, sur le modèle de ce qui existe dans d'autres domaines comme l'intérieur ou les finances publiques.

L'administrateur ministériel des données, algorithmes et codes sources joue évidemment un rôle clé dans l'accompagnement et la coordination des acteurs en vue d'une meilleure circulation des données et d'usages respectueux du droit et de l'éthique. Au-delà des initiatives citées, il est important que la démarche prospective amorcée à l'occasion de ce travail de préfiguration se poursuive sur le long terme et soit accompagnée d'expérimentations sur des sujets ciblés, difficiles à traiter mais à fort impact, en collaboration avec des acteurs de la recherche.

Lorsque des réponses auront été apportées aux besoins urgents ou élémentaires — démontrant concrètement l'intérêt et la valeur d'une plus grande circulation des données — et que les conditions nécessaires pour aller plus loin seront réunies, les besoins identifiés et principes fondamentaux donnés en annexe pourront aider à mettre en œuvre des moyens supplémentaires pour collecter, héberger, traiter, cataloguer, partager et réutiliser des données scolaires dans un cadre de confiance établi.

2 - Annexe : structures rencontrées

231 personnes ont été auditionnées de novembre 2022 à juin 2023, majoritairement lors de visioconférences organisées spécifiquement sur le sujet, mais aussi parfois en présentiel en marge d'autres réunions ou événements (e.g. enseignants en NSI réunis par l'Académie de Versailles, agents de la DNE dans le cadre d'un séminaire interne et participants au colloque des acteurs du numérique organisé par la DNE en juillet 2023).

Les structures rencontrées se répartissent de la manière suivante :

- **Collectivités territoriales** (27 personnes) : ADF, AVICCA, Bordeaux métropole, Commune d'Allonne, Commune de Meudon, Commune de Trilport, Commune de Villejuif, Département de la Manche, Région Grand-Est, Région Ile-de-France, Seine-et-Yvelines Numérique
- **Acteurs associatifs** (3 personnes) : La Quadrature du Net, SNES
- **Acteurs économiques** (59 personnes) : Belin Education, Beneylu, CGI, Cleyrop, Cohecitiz, CozyCloud, Domoscio, Editis, Edulib, EvidenceB, Google, Google - Sens, Hachette Livre, Humensis, La Poste / Docaposte, LDE, Le Monde, Librairie eMLS, Métapolis, Oppscience, Orange, Plume, Prometheus-X, Quantmetry, Scaleway, Sejer, StellIA, Wordline
- **Opérateurs publics et GIP** (10 personnes) : France Éducation international, GIP PIX, Onisep, Réseau Canopé
- **Éducation nationale** (89 personnes) : Académie de Paris, Académie de Poitiers, Académie de Versailles, Académie Nancy-Metz, IH2EF, Inspection générale, MENJ/DNE, Région académique Bourgogne-Franche-Comté, Région académique Bretagne, Région académique Grand-Est, Région académique Hauts-de-France, Région académique Nouvelle-Aquitaine, Région académique Occitanie, Enseignants
- **Enseignement supérieur, recherche & innovation** (20 personnes) : Programme IDEE, Membres d'unités de recherche (e.g. Coast, Flowers, LaBRI, LIP6, LISN, LORIA, Mnemosyne, PErSEUs, Potioc, Semagramme, Techné, XLIM), Inria, Université de Poitiers, Université de La Rochelle
- **Autres institutions** (23 personnes) : ANSSI, CNIL, Comité d'éthique pour les données de l'éducation, DGE, DINUM, Ekitia, Health Data Hub, SGPI

3 - Annexe : besoins identifiés

Le cycle d'auditions a permis de recenser différents besoins qui constituent des cas d'usage possibles (souhaités ou souhaitables) pour la plateforme. Certains besoins répondent à des enjeux d'organisation ou de pilotage, d'autres relèvent de la constitution de jeux de données inédits. On peut notamment évoquer une meilleure utilisation des traces d'apprentissage, l'une des motivations historiques du projet de plateforme, dans la volonté de faire avancer les travaux de modélisation nécessaires pour une meilleure personnalisation des apprentissages. La liste ci-dessous récapitule de manière synthétique ces besoins.

1. **Un annuaire fédéré des élèves et des enseignants dès la rentrée scolaire**

   Mettre à disposition des collectivités, régions académiques, académies et établissements, un annuaire fédéré des élèves et des enseignants dès la rentrée, pour organiser l'année scolaire et suivre les évolutions tout au long de l'année.

2. **Un tableau de bord du numérique éducatif**

   Mieux connaître l'usage et piloter le déploiement d'équipement et des ressources numériques, pouvoir poser un diagnostic de maturité numérique des établissements (en termes d'équipement, de ressources, de compétences numériques des élèves, des enseignants et d'autres personnels) pouvoir s'informer des politiques mises en œuvre sur d'autres territoires et s'en inspirer.

3. **Remplacement de courte durée**

   Améliorer la remontée de l'information sur l'absence d'un enseignant, organiser le processus d'affectation des remplaçants par des enseignants volontaires au sein du même établissement ou en provenance des bassins de remplacement concernés.

4. **Traces d'apprentissage – Solutions P2IA** [2]
5. **Traces d'apprentissage – Capytale** [3] (Académie de Paris)
6. **Traces d'apprentissage – Plateforme Avenir(s)** [4] (Onisep)
7. **Traces d'apprentissage « Moodle » – M@gistère, Eléa** [5]

   Améliorer la modélisation des apprenants dans des « profils » sur la base des compétences acquises, des habitudes d'usages, des préférences cognitives, et de leurs

2. Solutions en français et mathématiques développées via le P2IA : Partenariat d'innovation et d'intelligence artificielle.
3. Service numérique pédagogique développé par la DANE et la DSI de l'académie de Paris permettant la création et le partage d'activités de codage entre enseignants et élèves.
4. Programme pour améliorer l'accompagnement à l'orientation des jeunes, et le développement de compétences pour la construction de leur projet professionnel. L'ONISEP coordonne le programme et pilote la plateforme et le portfolio de l'enseignement secondaire. L'université Savoie Mont-Blanc est pilote coordinateur du portfolio du volet enseignement supérieur, avec d'autres partenaires et laboratoires de recherche.
5. M@gistère : plateforme Moodle pour la formation à distance de tous les personnels de l'éducation nationale ; Eléa : plateforme Moodle pour créer des parcours pédagogiques scénarisés pour les élèves.

processus d'apprentissage. Améliorer la finesse en termes de recommandation personnalisée et d'exercices proposés au sein des solutions, contribuer à une meilleure connaissance des activités des enseignants et des élèves à travers des solutions numériques, faciliter la mesure d'impact pédagogique des solutions.

8. **Jeux de données en langue française pour le traitement automatisé du langage**

   Constituer des corpus de données en français et en français langue étrangère pour des finalités de recherche (et, éventuellement, de recherche et développement) en expression orale, notamment pour des enfants, et en expression écrite.

9. **Outil d'aide à l'orientation personnalisée**

   Mieux informer et aider à construire les projets d'orientation des élèves en amont de leur choix d'orientation, proposer une aide personnalisée à l'orientation, mieux articuler l'aide à l'orientation avec les services proposés au niveau des collectivités, mieux suivre le parcours de scolarité et d'insertion professionnelle des élèves, dans le temps.

10. **Outil d'aide à la prévention du décrochage**

    Améliorer, au plus tôt, le repérage des élèves exposés aux facteurs de risques pour permettre une intervention plus rapide et détecter les signaux d'un potentiel décrochage, mettre en place des mesures d'accompagnement individuel dès les premiers signes de décrochage, sur les temps scolaire, péri et extra-scolaire.

11. **Personnalisation de parcours de formation au profit des enseignants et des cadres**

    Porter une action prédictive sur la formation initiale et continue des enseignants et des cadres, améliorer la personnalisation des contenus de formation, la recommandation de ressources et de parcours, améliorer le pilotage à travers une carte des formations.

12. **Santé et éducation : état de santé des enfants et parcours scolaire**

    Analyser le lien entre les besoins en soins des enfants et leurs résultats scolaires, identifier l'impact de l'état de santé des enfants sur leur parcours scolaire et contribuer à des réflexions sur l'amélioration du soutien en milieu scolaire des enfants avec des besoins de santé, selon leur milieu social.

13. **Mesurer l'impact des politiques publiques – réforme des lycées**

    Mesurer l'impact de l'introduction de l'enseignement des spécialités et de la réforme des modalités su baccalauréat, étudier la relation entre les résultats des élèves, leurs choix de spécialité et leur réussite au baccalauréat

14. **Optimiser des services des collectivités territoriales**

    Améliorer l'offre de services proposée par les collectivités locales, au plus près des besoins effectifs des citoyens. En particulier : améliorer les projections en termes de carte scolaire et la prévision en termes de gestion immobilière, ajuster l'offre en termes de restauration scolaire et de transport scolaires, optimiser la consommation énergétique, notamment pour le chauffage des salles.

4 - Annexe : principes fondateurs retenus

Les capacités actuelles à collecter de grandes quantités de données couplées aux impressionnants progrès des dernières décennies en matière de détection, classification, modélisation, raisonnement, prédiction, etc. laissent entrevoir de nombreuses possibilités dans le domaine de l'éducation. Certains espèrent ainsi pouvoir disposer à terme de modèles précis et prédictifs pour l'analyse de compétences, la personnalisation des apprentissages et de l'orientation, ou le pilotage d'investissements, par exemple. Nous en sommes encore loin aujourd'hui, pour des raisons assez clairement établies. Les principales difficultés ne se situent pas au niveau des traitements de données eux-mêmes, mais en amont (hétérogénéité des sources, absence de catalogues, accès difficile à certaines données, etc.) ou dans la compréhension des cadres juridique et éthique dans lesquels les traitements doivent être faits.

L'Intelligence Artificielle permettra-t-elle de révolutionner à grande échelle l'enseignement ou le pilotage du système éducatif ? Nul ne le sait vraiment mais il est important de se donner les moyens d'essayer, pour des raisons de souveraineté notamment : la maîtrise des technologies à fort impact potentiel dans un domaine aussi important est cruciale. Si l'on surestime souvent l'impact d'une technologie à court-terme, on sous-estime aussi souvent son impact à long terme. Il est donc important d'envisager les moyens évoqués comme une infrastructure évolutive sur laquelle on pourrait expérimenter et construire progressivement un ensemble de services, dans la durée, plutôt qu'une construction supposée répondre, une fois livrée, à un ensemble de besoins clairement spécifiés.

La plateforme des données de l'éducation doit être conçue comme une infrastructure habilitante destinée à faciliter l'accès à des données et leur exploitation dans le cadre juridique et éthique approprié. La suite de cette annexe présente les principes fondateurs sur lesquels nous avons convergé dans la perspective d'une mise en œuvre effective ultérieure de la plateforme.

4.1 Données

Les données « ouvertes » (publiques, déjà disponibles, comme celles de data.education.gouv.fr) doivent faire partie de celles accessibles à travers la plateforme, mais l'intérêt principal résidera évidemment dans l'accès aux données jusqu'ici non disponibles et qui n'ont pas vocation à être rendues « ouvertes ».

Toutes les données produites ou collectées dans le cadre scolaire devraient pouvoir être rendues accessibles à travers la plateforme. En pratique, toutes ne le seront pas, bien sûr. Mais dès lors qu'il y aurait un intérêt avéré ou supposé à les rendre accessibles, ce devrait être possible - dans des conditions à préciser (voir infra) - pour pouvoir développer de nouveaux services ou expérimenter.

Il est important de souligner qu'il n'existe pas à ce jour de cartographie partagée de l'ensemble des données susceptibles d'être partagées. Il n'existe pas non plus de référentiel

commun pour les données éducatives. Il ne s'agit donc pas de rendre accessibles des sources ou données précédemment identifiées, mais de permettre de rendre accessible demain toute source de données utile à des acteurs de l'éducation pour répondre à des besoins ou pour pouvoir innover. Le catalogue des sources de données accessibles via la plateforme sera construit progressivement, au fur et à mesure de son utilisation.

Les données accessibles via la plateforme doivent pouvoir être de toute nature : plus ou moins structurées ; collectées une unique fois ou de manière répétée (séries temporelles, flux de données) ; non identifiantes, à caractère personnel voire sensibles ; etc.

4.2 Conditions d'accès aux données

La plateforme devra évidemment respecter les règlements et lois en vigueur en la matière. Sa raison d'être étant de faciliter l'accès aux données et leur exploitation, sa mise en œuvre devra se faire selon une approche de type « aussi ouvert que possible, aussi fermé que nécessaire ». La plus grande ouverture sera possible pour les données « ouvertes » citées plus haut. La fermeture partielle ou complète sera parfois nécessaire, par exemple pour des données à caractère personnel. L'objectif est d'encadrer l'accès et de ne surtout pas le restreindre par défaut.

L'accès à des données à travers la plateforme est à comprendre dans un sens large incluant l'accès direct, lorsqu'elles sont hébergées sous la responsabilité de la plateforme, et l'accès indirect via un tiers vers lequel la plateforme redirige. Les conditions de ces accès devront dépendre de la nature des données, de la finalité de leur traitement initial, des conditions de collecte, et de la finalité de la demande d'accès/d'utilisation. Pour les données à caractère personnel, cela inclut la vérification de la base légale et de l'ensemble des principes sous-jacents au traitement (licéité, loyauté, transparence, proportionnalité/minimisation) et, lorsque nécessaire, une analyse d'impact (AIPD). Ces conditions devront être claires et contractualisées. Certaines seront définies en amont par le ministère et rendues opposables (via la doctrine technique, par exemple). D'autres pourraient être liées à la signature préalable d'une charte d'utilisation propre à la plateforme. D'autres seront négociées entre le fournisseur et le demandeur des données. Il ne sera pas nécessairement possible de garantir ex ante que l'utilisation des données respectera ces différentes conditions, mais celles-ci pourront faire l'objet d'audit ex post suivis d'éventuelles actions juridiques.

L'ambition de la plateforme doit être de faciliter l'accès à des données en ajustant les conditions de cet accès à la finalité de la demande. Lorsque les finalités de collecte et d'utilisation seront compatibles (but recherché, légitimité du demandeur, proportionnalité des données sollicitées, etc.), le demandeur pourra avoir accès aux données telles que collectées, y compris celles à caractère personnel. Lorsque les finalités ne seront pas compatibles, il serait souhaitable que la plateforme permette un accès « filtré » aux données permettant de les réconcilier : suppression d'informations personnelles pour respecter le principe de minimisation, pseudonymisation ou anonymisation, substitution des données réelles par des données synthétiques plus ou moins proches (e.g. avec des distributions similaires sur certaines dimensions), agrégation, etc.

Entre les données collectées que l'on peut progressivement appauvrir et des données synthétiques que l'on peut progressivement enrichir pour les rapprocher des données réelles, il faut ainsi imaginer un continuum dans lequel on aimerait pouvoir se placer en fonction des demandes. Il faut comprendre que les mécanismes de filtrage évoqués ne sont pas simples à concevoir et à mettre en œuvre, surtout si l'on veut pouvoir faire du suivi longitudinal dans le cas de séries ou flux de données.

Les conditions d'accès aux données devront également intégrer une dimension temporelle. Si le temps de conservation communément choisi pour de nombreuses données

est aujourd'hui d'un an pour la finalité de gestion administrative de l'année scolaire, il faudra permettre aux différents acteurs de l'éducation de s'inscrire dans d'autres temporalités : de quelques années pour mieux entraîner certains modèles, jusqu'au suivi sur des durées beaucoup plus longues d'individus ou de cohortes.

4.3 Architecture technique

Une manière de faciliter l'exploitation de données provenant de diverses sources est de les connecter à un point centralisé à travers lequel l'accès peut se faire de manière unifiée et sécurisée. Une manière de sécuriser plus encore le traitement des données consiste à ne l'autoriser qu'au niveau du point où sont rassemblées les données. Le traitement se fait ainsi dans une « bulle sécurisée » d'où les données ne peuvent éventuellement pas sortir. Ce modèle est depuis longtemps utilisé dans les domaines de la santé (c'est celui de la plateforme technologique du Health Data Hub, notamment) mais aussi de l'éducation (pour les données de la DEPP, pour celles de Parcoursup, ou dans le cadre du projet LOLA, par exemple).

Ce modèle de bulle sécurisée centralisant à la fois les données et les calculs présente néanmoins plusieurs inconvénients dans le cas qui nous concerne. Tout d'abord, le niveau de sensibilité des données produites ou collectées dans le cadre scolaire est très variable et ne nécessite pas forcément de telles précautions. Puisque l'esprit est « aussi ouvert que possible, aussi fermé que nécessaire », il ne faut fermer que si c'est réellement nécessaire, d'autant plus que la sécurisation a un coût important. Ensuite, quelles que soient les possibilités de traitement offertes par une bulle centralisatrice, celles-ci ne seront pas nécessairement suffisantes pour expérimenter de nouveaux usages (les chercheurs en sciences et technologies du numérique doivent pouvoir utiliser les méthodes et outils de leur domaine à l'état de l'art, par exemple). Les traitements, s'ils sont légitimes, doivent être le moins contraints possible d'un point de vue technique pour favoriser la recherche et l'innovation. Enfin, compte tenu de la diversité des acteurs de l'éducation, il semble illusoire de penser qu'une unique bulle pourra rassembler toutes les données et qu'elle pourra aussi recevoir celles d'autres domaines avec lesquelles on aimerait pouvoir les croiser, e.g. des données de santé.

Un modèle décentralisé des sources de données et des traitements doit être envisagé. Au lieu d'un entrepôt de données unique, il faut envisager de multiples entrepôts accessibles de manière standardisée. Chaque fois que ce sera possible, on préférera laisser les données là où elles se trouvent plutôt que les centraliser, et ne mettre en place que les moyens techniques minimums nécessaires pour permettre à d'autres d'y accéder. Au lieu d'une bulle sécurisée unique dans laquelle on peut mettre en œuvre des traitements, il faut envisager de multiples points de traitement dont certains nécessiteront la construction sur mesure de bulles de traitement. Une bulle de traitement est à comprendre comme un environnement logiciel spécifique à un traitement permettant de rassembler les données et outils logiciels nécessaires à celui-ci et mettant éventuellement en œuvre des mécanismes de sécurité plus ou moins contraignants (contrôles d'accès, enregistrement des activités, restrictions sur ce qui peut être fait ou ce qui peut sortir de la bulle, etc.). Dans ce modèle, la localisation de la plateforme est un non sujet, aussi bien dans l'espace physique que dans l'espace numérique, les entrepôts et bulles de traitement pouvant être opérés par différents acteurs.

La raison d'être de la plateforme étant de faciliter l'accès aux données et leur exploitation, différents éléments techniques pourront être fournis pour faciliter sa construction collective par les acteurs. Des briques logicielles pourront ainsi être fournies pour faciliter la collecte de données, leur stockage, leur catalogage, leur partage ou transmission sécurisés, et éventuellement leur traitement, leur analyse et leur visualisation dans le respect d'exigences plus ou moins élevées en matière de sécurité. Ces briques seront

idéalement développées dans une démarche de communs numériques, sur la base de technologies souveraines libres et dans un esprit communautaire.

Les caractéristiques souhaitées pour la plateforme devront déterminer les technologies utilisées pour la mettre en œuvre, et non l'inverse. Les briques logicielles évoquées devront aussi être réalisées en prenant soin de séparer les préoccupations (i.e. en isolant les aspects indépendants ou faiblement corrélés et en les traitant séparément), afin de ne pas surspécifier les mécanismes techniques et de pouvoir facilement changer leur politique d'utilisation si besoin. Les mêmes briques logicielles de base (mécanismes) devront pouvoir être utilisées pour construire des bulles de traitement respectant des cahiers des charges très divers (politiques d'utilisation).

4.4 Services

Les freins actuels au pilotage et à l'innovation sont divers, parfois techniques, parfois liés à une insuffisante connaissance et compréhension des cadres juridiques et éthiques, par exemple. Les services à mettre en place dans le cadre de cette plateforme seront donc de natures diverses.

La plateforme devra évidemment fournir les moyens techniques d'accéder à des données, mais sa réussite reposera tout autant, voire plus, sur sa capacité d'accompagnement (essentiellement humain) des projets. Cet accompagnement contribuera au renforcement de compétences pérennes de tout l'écosystème en matière de traitement de données et à la construction dans la durée d'une stratégie partagée sur ce sujet.

La liste ci-dessous donne un aperçu des services que la plateforme devra très probablement proposer dans ce sens :

- Actions de sensibilisation ou formation aux cadres juridiques et éthiques et sur la plateforme elle-même pour améliorer la compréhension de l'espace des possibles, de l'intérêt de la plateforme et de ce qu'on peut en attendre ;
- Catalogue des sources et des données disponibles à travers la plateforme ;
- Accompagnement dans la définition de projets : identification des données nécessaires, état des lieux de leur disponibilité, identification des pré-traitements éventuellement nécessaires (e.g. mise en qualité), aide à la constitution de groupes de contributeurs/bénéficiaires potentiels (pouvant inclure des élèves, enseignants ou parents) ;
- Procédures pour exprimer des demandes d'accès à des sources de données existantes ;
- Accompagnement technique et juridique pour la mise à disposition de nouvelles sources de données à travers la plateforme ;
- Accompagnement pour la contractualisation de l'accès aux données, par le biais de contrats type, d'assistant automatique ou de conseil juridique personnalisé ;
- Procédures et moyens techniques pour informer les personnes concernées par les données, gérer leur consentement, exercer leurs droits sur les données à caractère personnel, bénéficier de l'éventuelle connexion à des espaces numériques individuels, etc. ;
- Briques logicielles de base pour faciliter la collecte de données, leur stockage, leur catalogage, leur pré-traitement (e.g. filtrage), leur partage ou transmission sécurisés, et éventuellement leur traitement, leur analyse et leur visualisation ;

- Infrastructure permettant la création de bulles de traitement combinant les briques de base précédentes à d'autres outils logiciels pour garantir le respect de conditions de traitement contractualisées ;
- Audit de projets pour vérifier leur conformité à la réglementation et aux aspects contractualisés.

Ces services pourront être fournis par l'ensemble des acteurs impliqué dans la construction ou utilisateurs de la plateforme, dans une démarche communautaire, avec un rôle prépondérant du ministère.

4.5 Développement de la plateforme

La stratégie de développement de la plateforme importe tout autant que son architecture technique ou les services envisagés. La plateforme n'est pas une fin en soi. Elle ne doit pas être un objet que l'on construit et dont on espère ou cherche ensuite l'utilité. Elle doit être construite pour être utile. Compte tenu de l'architecture et des services envisagés, cette construction sera le résultat de l'agrégation progressive d'éléments sociotechniques interopérables. Il est important que cette construction soit guidée par des besoins concrets et des objectifs clairs, appuyée par des personnes directement concernées qui seront les premières à utiliser les services mis en place et pourront ensuite en être les ambassadrices.

Le développement initial de la plateforme doit se faire sur la base de projets pilotes, avec l'aide de personnes motivées désireuses d'utiliser des données et/ou prêtes à les partager. La phase de préfiguration a permis d'identifier une série de projets pilotes potentiels. Ceux-ci ne couvrent très probablement qu'une petite partie de ce qu'il serait possible de faire si les données de l'éducation circulaient beaucoup plus facilement. Mais l'ambition portée par la plateforme ne pourra monter qu'à mesure qu'elle démontrera son utilité.

Une part importante des besoins exprimés relève d'une transformation numérique incomplète des acteurs auditionnés : pour différentes raisons, le numérique reste insuffisamment utilisé en appui des processus de travail, et la marge de progression reste importante. Ces besoins doivent être traités en priorité. Tant que les acteurs auront des difficultés à répondre des besoins immédiats et conceptuellement simples (disposer d'une vue consolidée au niveau d'une région académique ou d'une communauté de collectivités, disposer des effectifs à jour et à temps pour organiser la rentrée, etc.) il leur sera difficile d'imaginer autre chose et l'intérêt de la plateforme sera limité. Les freins ou obstacles actuels doivent être perçus comme franchissables pour pouvoir aller plus loin.

Les projets pilotes visent à apporter des réponses concrètes à des attentes, mais aussi et surtout à amorcer la construction de la plateforme. Le portefeuille de projets initiaux doit donc être choisi avec attention et suivi attentivement pour que les solutions apportées participent effectivement à cette construction et ne répondent pas uniquement aux attentes locales exprimées.

4.6 Forme juridique

Le choix de la forme juridique dépendra de la volonté politique du ministère et des moyens actionnables.

Il pourrait être décidé d'internaliser l'administration (complète ou partielle) de la plateforme : cela passerait nécessairement par un renforcement de ressources internes au

ministère — sous réserve de la capacité d'attirer, de recruter et de pérenniser les compétences humaines nécessaires — et la création d'une équipe dédiée, probablement au sein de la DNE, pilotée par l'AMDAC. Une alternative serait de confier la gestion de la plateforme à un opérateur public ou de créer une entité tierce dotée d'une personnalité morale pouvant avoir une forme partenariale publique-privée (un GIP, par exemple). Cette nouvelle entité serait chargée d'organiser l'ensemble des services et de jouer le rôle central d'accompagnement de tout l'écosystème dans le portage des projets, sous le contrôle et la supervision stratégique du ministère. Cette alternative (l'externalisation) est mentionnée car elle semble être la forme privilégiée par d'autres projets similaires, nationaux comme internationaux.

Dans un cas comme l'autre, il semble indispensable de doter la plateforme d'une gouvernance organisée en différents comités pour permettre l'indispensable supervision stratégique et politique, mais aussi maintenir une dynamique opérationnelle d'accompagnement des projets, avec des temps de procédure qui doivent rester réduits, pour ne pas décourager l'adhésion des différents acteurs.

4.7 Gouvernance

La dimension de la gouvernance politique et partagée a été largement évoquée dans le contexte de la régulation de l'accès aux données éducatives. Pour garantir une utilisation responsable et bénéfique de ces données, il est nécessaire d'établir des règles de fonctionnement claires et concertées entre différents acteurs clés tels que le ministère, les collectivités territoriales, les communautés de recherche et les EdTechs.

La gouvernance politique implique la coordination et la coopération entre ces différentes entités. Le ministère doit jouer un rôle central en fournissant des orientations stratégiques et en fixant les objectifs globaux pour l'utilisation des données éducatives. Il peut également assurer la coordination entre les différentes parties prenantes et faciliter le partage d'informations et de bonnes pratiques.

Les collectivités territoriales, en tant qu'acteurs de proximité, peuvent contribuer à la gouvernance en adaptant les politiques et les pratiques à leurs besoins spécifiques. Elles peuvent jouer un rôle clé dans la collecte et la gestion des données éducatives au niveau local, en veillant à leur protection, à leur confidentialité et à leur utilisation éthique.

Les acteurs de la recherche seront à la fois contributeurs et utilisateurs de la plateforme : ils pourront à la fois partager des données (issues d'expérimentation ou d'analyses, par exemple) et accéder à des données produites par d'autres pour alimenter leurs recherches. Ces acteurs doivent pouvoir apporter leur expertise scientifique à la gouvernance de la plateforme.

Le comité d'éthique de la donnée du ministère doit également jouer un rôle crucial dans la gouvernance des données éducatives en rendant des avis sur la production, la collecte, l'utilisation et le partage des données.

Enfin, les EdTechs et éditeurs de solutions de vie scolaires jouent un rôle croissant dans le domaine de l'éducation en fournissant des solutions technologiques et des plateformes pour la collecte et l'analyse des données éducatives. Dans le cadre de la gouvernance, il est important d'établir des partenariats et des accords clairs avec les EdTechs pour garantir le respect des normes de protection des données, la transparence et la responsabilité dans l'utilisation des données éducatives.

4.8 Modèle économique

La plateforme est à considérer comme « une fabrique à projets ». Elle doit aider les acteurs à concevoir des projets et à les lancer, mais elle n'a pas vocation à les financer. Il apparaît indispensable de mobiliser des fonds publics pour financer son démarrage, les premiers développements techniques nécessaires et l'amorçage de quelques projets pilotes (sur une durée de 12 à 24 mois maximum par projet).

Le budget initial estimé pour trois ans autour de 4 M€ couvre les lignes de dépense suivantes :

Nature des dépenses	**Coût**
Coordination, accompagnement des projets (sur trois ans)	1 400 000 €
Infrastructure initiale (coût initial)	600 000 €
Accompagnement technique (sur trois ans)	1 000 000 €
Accompagnement juridique (sur trois ans)	455 000 €
Ingénierie pédagogique (sur trois ans)	180 000 €
Communication (sur trois ans)	200 000 €
Projets pilotes (sur trois ans)	210 000 €
TOTAL (sur trois ans)	4 045 000 €

Le développement et la mise en œuvre de briques logicielles de base et de la fabrique à bulles de traitement pourront impliquer des acteurs de la recherche pour leur outils et savoir-faire. Si de nouvelles actions de recherche sont nécessaires, elles pourront être soutenues sous forme de projets collaboratifs via les moyens de financements habituels (MENJ-GTnum, ANR, France 2030, etc.).

Au-delà des trois premières années détaillées ci-dessus, nous estimons le coût de maintien en conditions opérationnelles (MCO) de la plateforme à environ 1 M€ par an. Nous considérons que ce coût doit être supporté par la structure publique dans la mesure où le coût d'investissement et le MCO de l'infrastructure échappent à tout modèle économique de rentabilité.

L'amorçage de nouveaux projets devra être financé par chaque porteur pour un coût marginal estimé à 30 k€ par projet, avec de possibles aides financières extérieures. Après la phase d'amorçage, le coût de maintenance d'un projet devra être assumé par son porteur (probablement autour de 15-20 k€ par an, avec des spécificités liées à chaque projet). Concernant la capacité des porteurs à financer le coût dans la durée des services rendus par leurs projets, des auditions (des collectivités territoriales, en particulier) laissent penser qu'il serait possible d'introduire des mécanismes payants, à condition d'apporter de réels services et une vraie simplification pour les acteurs de l'écosystème.