

L'IA au service de la Data : une approche par le cycle de vie des données

Document de travail élaboré dans le cadre du comité d'experts d'Opendatafrance.

Le groupe de travail est composé de Simon Chignard (Observatoire Data Publica), Pauline Asselin (Région Sud), William Michel (Cannes Pays de Lérins), Jean-Marc Lazard (Huwise), Laurent Toustou (IGN) et Gabriela Martin (Opendatafrance).

Les relations entre l'intelligence artificielle et les données sont riches et nombreuses. Les données servent par exemple à entraîner les modèles d'IA, qu'il s'agisse d'IA générative (pour créer du texte ou des images) ou d'IA prédictive (pour prédire par exemple la consommation énergétique d'un bâtiment public ou la fréquentation d'un tronçon routier).

Dans le cadre du comité d'experts d'Opendatafrance, la discussion a été engagée sur une autre dimension : celle de l'usage de l'IA *au service* de la donnée et de ses métiers. Comment les outils d'IA peuvent-ils faciliter la production, la diffusion ou encore l'exploitation des données, notamment des données ouvertes ? Ce document propose une première approche, en partant du cycle de vie des données. Le groupe de travail a ainsi listé une trentaine de cas d'usages (ou de tâches) qui peuvent potentiellement être réalisées à l'aide de l'IA. Cette liste n'est pas exhaustive et doit être utilisée avec discernement. Cette liste ne constitue pas non plus un appel à utiliser l'IA sans considération ou prise en compte de ses impacts éthiques et environnementaux. Pour certains cas d'usages, il existe peut-être des alternatives qui n'ont pas recours à des solutions d'intelligence artificielle.

Le présent document est un point de départ d'une discussion plus large avec les membres d'Opendatafrance sur l'usage de l'IA pour la Data. Une prochaine étape, à discuter lors du prochain Comité d'experts, pourrait être de recueillir les témoignages et les retours d'expérience des acteurs publics qui ont déjà expérimenté certains des cas d'usage présentement listés.



Les étapes du cycle de vie de la donnée et les usages de l'IA associés

1. Production et collecte de données

La première étape du cycle de vie correspond à la production et la collecte de données. L'IA peut intervenir en amont de cette étape (par exemple pour concevoir un plan de collecte de données à partir d'un cas d'usage), pour extraire des données existantes dans des documents de type et de format variés. Enfin, l'IA peut aussi faciliter la création ex-nihilo de jeux de données synthétiques.

Les tâches auxquelles l'IA peut utilement contribuer

- Préparation d'un plan de collecte de données (à partir d'un cas d'usage)
- Extraction de données existantes dans un PDF structuré ou non (OCR, ...) ou à partir d'une visualisation
- Récupération de données à partir d'un corpus (ex. analyse de questions ouvertes, analyse de sentiment, analyse de textes législatifs ou réglementaires) ou d'une image (computer vision)
- Retranscription automatique de réunions (audio, vidéo, ...)
- Extrapolation à partir d'un échantillon de données existantes
- Création de données synthétiques

2. Mise en qualité, standardisation, nettoyage

La deuxième étape du cycle de vie correspond à la mise en qualité des données : standardisation, détection d'erreurs, nettoyage des données, mais aussi des métadonnées. L'IA peut aussi contribuer à enrichir facilement des jeux de données, une tâche qui demandait auparavant un minimum d'expertise technique.

Les tâches auxquelles l'IA peut utilement contribuer

- a. Audit qualité de manière simple (sans outil expert)
- b. Détection et nettoyage d'erreurs et d'anomalies
- c. Validation automatisée - conformité à des schémas de données
- d. Enrichissement de jeux de données existants
- e. Aide à la préparation des métadonnées (titre, description, mots-clés)
- f. Anonymisation - pseudonymisation de données ou d'images
- g. Aide à la création d'un schéma de données ou métadonnées

3. Publication et circulation

La troisième étape du cycle de vie correspond à la publication et l'exposition des données : préparation des données selon le public cible, aide à la description des données et des métadonnées (titre, mots-clés, description optimisée pour les moteurs de recherche), ...

Les tâches auxquelles l'IA peut utilement contribuer

- a. Reformatage de données pour de nouveaux publics (étudiants, grand public, experts métiers, ...)
- b. Description des jeux de données selon le public cible
- c. Traduction multilingue (interface, métadonnées, données)
- d. Détection d'informations sensibles avant publication

4. Accès et diffusion

La quatrième étape du cycle de vie correspond à l'accès et à la diffusion de la donnée. Et si l'IA nous amenait à repenser les modes d'accès à la donnée, traditionnellement sous la forme d'un portail de données et de la mise à disposition d'API ?

Les tâches auxquelles l'IA peut utilement contribuer

- a. Proposition de nouveaux modes d'accès à la donnée (ex. protocole MCP, agents IA, dans des outils existants, interrogation des données depuis des outils métiers ou bureautiques)
- b. Amélioration de la trouvabilité, recherche sur les portails de données

- c. Personnalisation de l'accès selon le profil utilisateur (suggestions, filtres intelligents)
- d. Amélioration par la mesure de sentiment sur les commentaires des jeux de données
- e. Observabilité des usages - création de persona - création d'une enquête usagers

5. Valorisation, exploration et médiation

La cinquième étape du cycle de vie correspond à la valorisation, l'exploration et la médiation; comment l'IA permet-elle d'exploiter plus facilement les données, par exemple en créant des visualisations, des tableaux de bord voire des applications complètes (approche vibe coding) ?

Les tâches auxquelles l'IA peut utilement contribuer

- a. Exploration d'un jeu de données en langage naturel ("poser" des questions à un jeu de données - NLQ)
- b. Génération d'applications, de tableaux de bord ou de visualisation
- c. Analyse prédictive ou prescriptive (modélisation, prévision, simulation)
- d. Identification ou création de scénario d'usage et de médiation (kit de médiation prêts à l'emploi)

6. Archivage, sauvegarde, destruction

La sixième et dernière étape du cycle de vie correspond à l'archivage, la sauvegarde et la destruction in fine des données.

Les tâches auxquelles l'IA peut utilement contribuer

- a. Détection de doublons avant archivage
- b. Aide à l'anonymisation avant archivage
- c. Audit de conformité des archives (intégrité, traçabilité etc.)
- d. Destruction automatisée de données en fin de vie