Grandes tendances : retour sur le Salon de la Data et de l'IA

Synthèse présentée par Benjamin Chartier, consultant, pour Afigéo, Sept 2024.

Le Salon de la Data et de l'IA s'est tenu la Cité des Congrès de Nantes le 17 septembre 2024, avant veille des GeoDataDays 2024, organisé par Afigéo & DécryptaGéo. Dans le cadre de l'accompagnement par Benjamin Chartier (Opteos) de projets techniques pilotés par l'Afigéo, ce dernier a réalisé une synthèse sur des sujets d'intérêt pour les membres de l'association (données & administrations publiques, cadre réglementaire et information géolocalisées).

Sur le plan organisationnel, ce salon s'est tenu à la fois en présentiel et distanciel. Cette édition était fortement teintée d'intelligence artificielle, sujet qui mobilise fortement le secteur de la tech depuis quelques années. Il réussi à regrouper les diverses parties prenantes du monde des données : des fournisseurs de services, des utilisateurs, des acteurs privés, des représentants de la société civile, des administrations, des avocats, des chercheurs... On peut noter la présence du volet géo représenté par l'IGN, Geofit, l'Afigéo, et Isogeo.

Impacts du développement de l'IA

La généralisation des IA est telle que deux mises en abimes remarquables ont été évoquées lors du salon :

- L'évaluation de la qualité des LLM (large language models - grands modèles linguistiques) est actuellement confiée de manière automatique à d'autres LLM ;
- Il est devenu courant qu'on utilise des intelligences artificielles génératives pour interroger d'autres intelligences artificielles ; est-ce que le métier de Prompt engineer est déjà obsolète ? Les liens qui unissent IA et données sont des liens de dépendance forts : pas d'IA sans données car les données servent à entrainer, calibrer et optimiser les IA et beaucoup d'IA ont souvent pour objectif de produire des données.

De plus la qualité des données utilisées en amont se répercute inévitablement sur la qualité des résultats des IA. Il apparaît donc nécessaire que pour des usages nécessitant une certaine forme de garantie sur les résultats de disposer de jeux de données de bien meilleure qualité.
La présentation faite par Linagora de son LLM Lucie portait justement sur la manière de bien choisir les sources de données utilisées pour son entrainement au regard du public et des usages ciblés : qualification de corpus de données et équilibrage des langues (pour éviter la prépondérance de l'anglais).

Autre fait intéressant que l'avènement des IA met en évidence : jusqu'à présent on considérait que des données non structurées n'étaient pas vraiment des bases de données exploitables de manière automatique ; les IA actuelles exploitent couramment des données non structurées. Ainsi le règlement européen sur la gouvernance des données (data governance act) de cette année intègre dans le giron des données les enregistrements sonores, visuels et audio-visuels. Cela implique un changement de regard sur les données pour ceux qui sont chargés de leur gouvernance et de leur traitement dans nos entreprises et administrations.

IA et confiance

Les IA posent un certain nombre de difficultés au regard de la réglementation applicable et de la confiance que l'on place en elles :

- Certaines données ne devraient pas être utilisées pour leur entrainement sans autorisation spécifique. Cela pourrait être le cas d'œuvres protégées par le droit d'auteur pour des IA générative capable de produire des créations similaires. Cela pourrait être le cas de données à caractère personnel dont l'usage est cadré par le RGPD.
- Le RGPD pose d'ailleurs d'autres difficultés pour les IA concernant la durée de conservation des données, la notion de profilage et au regard de l'exigence de transparence des traitements réalisés.
- Même si les données utilisées pour entraîner l'IA ne sont pas des données à caractère personnel, l'utilisateur final peut lui en fournir lorsqu'il l'interroge. Le traitement de ces données, leur conservation pour des traitements ultérieurs et leur confidentialité posent question. D'ailleurs cette problématique de confidentialité n'est pas particulière au RGPD ; elle se pose également pour d'autres types de secrets.
- En France, une création produite par une machine ou un algorithme n'est pas protégée par le droit d'auteur (le créateur doit être une personne physique). Cela n'empêche pas que leurs créations peuvent enfreindre le droit d'auteur des créations qui ont été utilisées pour les entrainer. Les utilisateurs d'IA génératives peuvent donc en théorie être poursuivis pour contrefaçon. Cela pose donc le problème de leur transparence (des sources, des algorithmes) et de l'explicabilité des résultats.
- Les IA ont tendance à reproduire et renforcer les biais et les erreurs présents dans les données qui leur ont été fournies. Cela pose donc encore une fois la question de la qualité et de la confiance que l'on peut attribuer aux données d'entraînement.
- On ne peut pas non plus ignorer les usages détournés d'IA pour réaliser des actes malveillants : deep-fake, usurpation d'identité (par reproduction de la voix d'une personne par exemple).

Ref: voir article de Journal du Net, traitant d'une partie de ces sujets un peu plus en profondeur : La qualité des données est primordiale pour libérer tout le potentiel de l'IA.

Administrations publiques et données

La loi pour une République numérique a donné une impulsion forte à la publication de données par les administrations publiques. Néanmoins, seuls 16% des collectivités territoriales censées ouvrir leurs données le font réellement, soit par manque de moyens, soit par absence de volonté politique.

On constate également que même si les données sont publiées, il est fréquent qu'elles ne présentent pas les qualités nécessaires à leur réutilisation. Cela témoigne que de la donnée n'a pas encore imprégné les administrations publiques. Elles ont sans doute besoin d'être accompagnées plutôt que laissées seules face aux défis techniques que les réglementations en la matière leur imposent.

D'ailleurs, Simon Chignard de DataPublica, dans son intervention consacrée à l'état des lieux de la donnée publique, relevait que les administrations publiques s'engagent de plus en plus de manière collective sur la voie

des données, soit via des partenariats entre institutions nationales, via des mutualisations entre collectivités, via la mise en place d'espaces communs de données, voire en mettant en place des démarches citoyennes.

Il a été évoqué lors du salon le fait que les Français font partie de ceux qui exprime le plus leur défiance vis-à-vis de l'usage qu'il peut être fait de leurs données. Paradoxalement, il y aurait une forte défiance des citoyens vis-à-vis de l'usage des données personnelles par les administrations publiques alors qu'ils auraient tendance à confier facilement leurs données aux acteurs du web (de manière consciente ou non).

Cela illustre le besoin pour les acteurs publics de définir des cadres de confiance dans un contexte de crise démocratique et de défiance générale vis-à-vis des usages abusifs des données. Quelques administrations se sont emparé de cette problématique et ont présenté les démarches qu'elles ont mises en place pour impliquer leurs usagers dans la définition de la stratégie ou de la gouvernance en matière de données et d'IA. Exemples : la convention citoyenne sur l'IA de Montpellier et la concertation sur la stratégie de la donnée de Rennes. Les objectifs de ces actions sont multiples :

- Affirmer le rôle pilote des collectivités autour des données ;
- Mettre en place une réelle écoute des citoyens ;
- Établir un cadre de confiance.

Évolutions du cadre réglementaire

Le cadre réglementaire lié aux données et à l'IA évolue à grande vitesse depuis quelques mois, surtout au niveau européen. On peut citer les règlements sur les données (Data Act), sur la gouvernance des données (Data Governance Act) et sur l'intelligence artificielle (AI Act).

Pour l'essentiel, tout le monde s'accorde à dire que ces évolutions vont dans le bon sens car ils définissent un cadre commun avec la mise en place de garde-fous attendus. Néanmoins, l'enchainement de ces évolutions et l'enchevêtrement des textes rendent la chose peu lisible et ne laissent pas beaucoup de temps aux acteurs pour se les approprier et les mettre en œuvre. Deux besoins se sont faits fortement ressentir lors du salon : une simplification et à un accompagnement.