Besoins / Qualité des données.

Animateur / Rapporteur : Alexandre Defossez

Groupe B

- Toutes les données comportent un biais.
- Les besoins en terme de qualité de la donnée dépend de la question que l'on se pose.
- Les conditions de collecte des données est aussi un facteur très important à prendre en compte.
- Prendre plus de temps pour préparer les données avant analyse, réaliser des prétraitements, peut aussi contribuer à améliorer leur qualité (et donc l'analyse dans son ensemble).
- Il est aussi essentiel de bien formuler ses hypothèses et son plan d'échantillonnage pour assurer la qualité des données produites.
- Cela étant dit, il y a toujours une certaine limite à l'amélioration de la qualité des données. Il peut exister des contraintes qui font que l'on n'a pas toujours la main sur la qualité du jeu de données que l'on utilise (notamment lorsque l'on est pas soi-même producteur des données).

Groupe A

- La qualité des données est liée à une vision à long terme, avec la perspective d'une continuité, d'un suivi temporel.
- La reproductibilité est un gage de la qualité des données.
- La jugement sur la qualité des données est très liée à la question que l'on se pose.
- « Qualité » : repose aussi sur une certaine subjectivité.
- Plus on remonte dans le temps, plus il est difficile de d'évaluer la qualité des données. Comment utiliser des données « anciennes » ?
- Quelle place donner aux « anecdotes » issues par exemple de conversations avec des acteurs (hors protocole d'enquête) mais qui pourtant comportent des informations valables et utiles au chercheurs ?
    - Mentionner l'information comme « dire d'acteurs » (ou d'expert) est une solution possible dans le cadre d'une publication.
- Le jugement sur la « scientificité » de la donnée dépend aussi du champ disciplinaire.
    - Dans le domaine des SHS, les paroles d'un acteur peuvent être considérées comme une donnée « scientifique » alors qu'il ne s'agit pas d'une mesure « physique » (dans la mesure où l'entretien se déroule selon un certain protocole).

Groupe C

- Importance majeure des plateformes « unifiées » pour accéder à la donnée.
    - D'où l'importance d'initiative comme Théia / Data Terra.
- Dans un « monde idéal » on aurait une annuaire centralisé avec toutes les coordonnées des chercheurs à l'échelle de la communauté internationale.
    - Dans les faits, peut-être que cela sera permis par l'IA générative ?
- Il existe un frein à la démarche de partage des données, de la mutualisation des efforts des chercheurs : la compétitivité très forte qui favorise une forme d'individualisme à diverses échelles (l'unité, nationale, internationale).
    - Tension entre tendance (ultra)individualiste et collaboration au sein de la recherche.

    - Or cet individualisme nuit à la qualité de la recherche.
    - Démarche de mutualisation permet aussi d'éviter les doublons.
    - D'un autre côté, on peut comprendre cette attitude dans le cas des jeunes chercheurs, en début de carrière (doctorants, post-docs) : beaucoup d'efforts à investir dans sa recherche et besoin de faire ses preuves (quelque part nécessité de s'attribuer les mérites de sa recherche pour pouvoir se faire une place dans ce monde ultra compétitif).
- Actuellement beaucoup de volonté de mutualisation des données, de la recherche en général, mais dans les faits d'importantes difficultés existent encore.
    - Exemples : accès aux données de biodiversité, certaines données d'aménagement du territoire.

Groupe D

- L'accès à de nombreuses données demeure encore liée à la discipline d'origine du chercheur (demande certaines compétences techniques, par exemple la donnée météo). En dépit de la démarche interdisciplinaire.
    - Importance des initiatives de type Théia / Data Terra.
- Au sujet de la qualité des données : c'est aussi une question de bonne pratique, notamment en ce qui concerne le renseignement des métadonnées.
- Difficile d'échanger sur les besoins en termes de données car ils sont extrêmement divers au regard de la diversité des profils des participants : données textuelles annotées, données d'imageries satellitaires, données SHS, données de biodiversité, etc.
- Dans certains cas, ce qui pose question ce n'est pas tant le besoin concernant la donnée elle-même mais le besoin de savoir comment traiter la donnée.
- Un autre type de difficulté lié à l'accès à la donnée : le blocage politique et institutionnel dans le cas des données « sensibles » (ex. Données de santé).
- Comment valoriser la donnée « experte » (provenant des acteurs) ? Comment l'utiliser dans une démarche de validation ?
- Aujourd'hui il existe une diversité de données « brutes » et données thématiques « finies » : il n'est pas toujours facile d'identifier les jeux données les plus appropriées pour répondre à la question que l'on se pose, face à la diversité de jeux de données et de produits finis existants. Ex : quelle carte de land cover choisir ?