AXE "DATA INTELLIGENCE" (2014/2015) – lundi 20 avril 2015 IMI

Présents à cette réunion de préparation: F. Ghitalla, P. Jollivet-Courtois, Célo Collomb, V. Misseri, S. Grès, Sebastien Telebaum (ingénieur de recherche UPMC inscrit en Master sociologie politique à Paris 7), Alberto Romené (chercheur à l'université de Porto., philosophie de la technique).

Nous fixerons, lors de la prochaine réunion programmée le 26 mai 2015 à l'IMI, le calendrier prévisionnel pour l'année 2015-2016. L'idée de l'alternance des réunions orientées «discussions/présentations théoriques et méthodologiques» et des séances d'atelier proprement dit (avec manipulation de data) doit encore être validée. A priori, les deux types de séance devraient alterner chaque mois (soit une réunion tous les quinze jours). Le rythme évoqué dépendra par ailleurs du nombre de participants et des aléas de calendrier (comme les périodes de vacances UTC).

Les notes prises au cours de la séance par Pascal ont été intégrées (pour partie) dans le texte de présentation de l'Atelier qui suit. En effet, l'une des façons pertinentes de capitaliser sur nos échanges est, selon moi, de faire évoluer le texte de présentation de l'Atelier.

Evidemment, je prendrai en compte vos remarques par e-mail en plus des informations recueillies lors des réunions.

TEXTE DE L'ATELIER (en évolution continue)

La culture naissante des data et de son ingénierie tournée vers les réseaux revêt bien des aspects. Depuis les masses de données elles-mêmes (le big data et son satellite l'open data), en passant par le *data analytics* et le data intelligence (processus de traitement de l'information) jusqu'à la figure montante du *data scientist* qui annonce peut-être une nouvelle "science" hybride gouvernée par des hypothèses et des expérimentations mais aussi par une méthodologie inductive (*data driven*) de découverte de patterns statistiques - voire de "connaissances" (*knowledge discovery*) - exploitables sous forme de nouveaux services par les géants de l'information. On parle aussi d'intelligence des données pour résumer toutes les étapes des chaînes de traitement de l'information en masse (recueil, archivage, analyse, filtrage, fusion...) qui dépasse aujourd'hui l'univers du web et des moteurs de recherche (gènes, protéines, transport, téléphonie mobile, etc.) où elles sont nées dans les années 90.

Les questions que l'on pose à "l'approche data" sont aussi vastes et diverses que le champ lui-même: certaines sont profondément liées à des problématiques très anciennes de "construction des connaissances", de "limites des outils" ou de la "relativité" du point de vue de celui qui traite les données (l'ingénieur, le chercheur, l'analyste...), d'autres concernent la "plus-value" (économique ou sociétale) que l'on peut attendre des technologies issues de cette écologie des données en train de naître.

L'axe "Intelligence des données" se présente dans un premier temps sous la forme d'un atelier à forte dimension expérimentale avec le témoignage de professionnels, ce qui nous permettra à la fois de nous tenir loin des caricatures que l'on veut parfois dresser (culte des machines, obsession des statistiques, absence de recul critique sur les pratiques et méthodes) et à la fois de faire gagner les discussions en pertinence à partir d'une mise en évidence des traits saillants d'une ingénierie des données effectivement pratiquée. C'est l'objectif majeur de l'atelier intelligence des données où l'on se propose d'observer les

différents processus de traitement des données, depuis l'extraction, la curation, l'indexation jusqu'au travail de conception des services utilisateurs qui inclut le design d'interfaces ou le déploiement d'algorithmes.

La posture souhaitée relève d'une approche volontairement constructiviste, en concentrant les discussions autour de la description des opérations «élémentaires» de transformation des données en corpus intelligibles. Les questions de méthodologie y seront donc centrales (agrégation, enrichissement, croisement, filtrage...), tout comme l'expérimentation technologique et le partage des résultats de l'expérience technique. Les problématiques développées se situent donc dans un espace encore peu exploité: au delà des données mais en deçà des problématiques de recherche particulières de chacun des participants (avec des participants que l'on souhaite issus des SHS comme des sciences exactes). Ainsi, il apparaît qu'un même socle de problématiques techniques et méthodologiques alimente bien des types de recherche: comment construire une grille de description d'acteurs ou de communautés en ligne, sur Facebook ou Viadeo ? Comment croiser des noms d'inventeurs issus de notices de brevets et des profils linkedin ? Comment traiter-agréger des données pour visualiser des flux entre des points dans l'espace (par exemple pour analyser les données d'usage de réseaux de téléphonie mobile) ? Peut-on créer de nouveaux systèmes de classification des domaines de recherche d'une université à partir de l'analyse des publications ? Comment produire des données qui permettront d'analyser les phénomènes dynamiques et temporels ? Existe-t-il d'ailleurs des "modèles temporels" pour les masses de données ?

«L'atelier data» aura donc pour objectif principal d'articuler une série de problématiques théoriques (typiques de l'usage des data en recherche comme dans l'industrie des services) à un exercice de construction-déconstruction des «boîtes noires» qui fondent la capacité opérationnel des dispositifs de traitement automatiques. Cela peut passer, d'un côté, par un questionnement des informaticiens et des ingénieurs dans leur travail quotidien sur les data et les services mais aussi, d'un autre côté, par une pratique de production et de manipulation de data sets, quitte à «réinventer» une sorte de «boîte à outils» (dans l'esprit des *data hackers spaces*).

a) La distinction entre "**approche qualitative"** et "**approche quantitative**", très présente dans les débats de méthodologie en sociologie notamment, occupe une place centrale dans les réflexions actuelles sur la façon de construire les objets de connaissance à partir des données numériques. La distinction peut, ou non, être scientifiquement "rentable" selon la perspective adoptée. Ce thème constituera l'un des débats centraux de l'atelier pour observer et expérimenter les multiples façons dont les deux approches s'hybrident dans le travail sur les data en une sorte d'alchimie quali-quantitative qui peut déboucher, par exemple, sur la conception d'un algorithme. C'est l'agilité avec laquelle les chercheur.e.s et les ingénieur.e.s en data sciences manient cette alchimie qui détermine souvent le nombre et la richesse des prises que l'on se donne sur les corpus de données numériques.

b) La **nature des corpus** en data sciences, qui ne s'épuisent pas dans les critères de pertinence (ce qui n'est pas pertinent à un certain moment ou selon certaines dimensions, le sera sous d'autres angles et à d'autres moments), d'exhaustivité (l'impossible appréhension du tout et de ses parties dans l'univers des masses de données en réseau) ni de clôture (les données sont prises dans des boucles de transformation, par exemple formatées pour de nouveaux services qui eux-mêmes en produiront de nouvelles). L'atelier aura pour objectif d'éclaircir les différentes méthodologies (extraction, construction, fusion) à l'oeuvre en phase "amont" d'un travail sur les data.

c) Les rapports du "**manuel**" et de "**l'automatique**", que l'on confond souvent avec la distinction qualitatif-quantitatif. Sur ce point, la partie "expérimentation" de l'atelier intelligence des données permettra

d'éprouver l'hybridation presque systématique des deux dimensions. Le travail d'ingénierie des données montre combien la construction d'une "machine logique" comme un algorithme passe d'abord par une série de phases exploratoires et de manipulations presque "artisanales" sur les data.

d) L'articulation entre "données" et "hypothèses" et du **niveau intermédiaire du "modèle de donnée"**. Les data sciences et les méthodes d'intelligence des données se réclament en effet d'une *data driven methodology* où le travail sur les données précède la formulation d'hypothèses. On peut aussi parler de méthode inductive par opposition à une démarche hypothético-déductive où le dispositif expérimental (et donc les data) sont mobilisées avant tout pour valider ou falsifier une ou plusieurs hypothèses de départ. Plutôt que d'opposer massivement les deux approches, il paraît plus pertinent d'interroger le travail de construction des modèles de données en data sciences, une activité centrale en recherche et développement mais aussi sur le plan scientifique.

e) **Les changements d'échelle**s. Question centrale dans une démarche orientée data, l'atelier conduira une réflexion spécifique sur les effets scientifiques induits dans une démarche où l'accumulation quantitative des données peut en effet conduire, dans certains cas, à la *modification qualitative de l'objet de science*.

f) Les questions de **valorisation** des (big) data. L'exploration de l'écosystème des big data, et une bonne partie de ce que l'on désigne par intelligence des données, débouche sur la mise au point de services innovants (chez les "géants" de l'information), l'avènement de formes sociales d'organisation en ligne (comme dans le logiciel libre à travers la contribution à des langages de programmation) ou encore des formes nouvelles de codification et de protection de la propriété intellectuelle. Parmi toutes les pistes potentielles ouvertes dans cette direction, nous voudrions privilégier deux pistes de réflexion (et peut-être d'expérimentation). La première concerne la question juridique des droits d'appropriation des droits de propriété traditionnels aux formes nouvelles émergées du Libre, de l'open source, de l'open science, de l'open data. Avec en ligne de mire la question assez primordiale du statut juridique des banques de données et des investisseurs dans le formatage, storage et le traitement des banques de données. La seconde concerne la caractérisation de la nature des données échangées et du type de biens informationnels ou immatériels ou intangibles du point de vue de leur portée économique. L'idée serait de viser les multiples façons dont des scénarios de valorisation se mettent en place actuellement basés sur le principe d'enrichissement des données ou de conception de services avancés. Au delà, les discussions de l'atelier pourront porter sur les méthodes de description et d'analyse des interactions des acteurs socio-économiques dans des contextes d'innovation "ouverte" ou "ascendante" où la mesure de la valorisation autour des données en masse ne s'arrête pas à l'aspect financier mais inclut bien des formes "d'externalités positives" comme l'engagement, le potentiel de coopération entre acteurs ou la confiance.