

Cet atelier sur la qualité des données historiques (30 septembre-2 octobre, Université Toulouse Jean Jaurès) entend proposer un espace de débat pour réfléchir sur les processus d'acquisition, de structuration, de curation, de qualification et de partage des données informatiques sur le passé, notamment en questionnant la confiance qu'on peut leur accorder et la notion de qualité qui reste encore aujourd'hui très largement impensée.

Ces dernières années ont en effet vu le développement de projets et de réflexions mêlant histoire et informatique, à une échelle semble-t-il renouvelée depuis la fin de l'histoire quantitative¹. Dans une dynamique qui mobilise des chercheurs et des institutions de recherche sur l'ensemble du territoire français, différentes équipes ont constitué des bases de connaissance sur leurs objets respectifs : démographie historique, prosopographie ou cartographie² ; ce faisant, elles ont été confrontées au flou, à l'ambiguïté, à l'incertitude dans la manière de représenter informatiquement les traces et les manques des archives. Ces équipes n'ont pas été les premières à être dans ce cas de figure : dès le premier numéro d'*Histoire & Mesure* publié en 1986, Jean-Philippe Genet faisait état de l'aspect multidimensionnel des données en histoire, indiquant que « la donnée historique a en effet trois dimensions, puisqu'au-delà de l'information elle-même, il faudrait, idéalement au moins, indiquer la référence de la source et le niveau de valeur de certitude de l'information ; cette dernière dimension fait d'ailleurs de la donnée historique une donnée floue³. »

Plusieurs décennies après ce constat, celui-ci demeure toujours aussi pertinent : Stéphane Lamassé, Philippe Rygiel et Cédric du Mouza ne disent pas autre chose quand ils parlent de « mutation des données en contexte historien⁴ ». Alors même que l'informatique s'est développée au point de transformer l'ensemble de notre société, que la technique a changé, la question demeure inchangée : comment intégrer l'incertitude dans les données informatiques historiques ? Les enjeux sont d'importance puisque l'intégration des conditions de production des données et d'un lien avec les appareils documentaires – ce que ce même Jean-Philippe Genet nomme des métasources⁵ – est impératif pour assurer les fondamentaux de l'analyse historique. Il est rejoint dans sa conceptualisation des données informatiques historiques par Johanna Drucker pour qui le caractère construit des sources doit être pris en compte dans la définition des données. D'ailleurs, plus que d'une data, il faudrait selon elle parler d'une capta « créée à partir de phénomènes, extraite à partir d'un paramétrage dans ce qui réifie sous forme de données, puis implique des structures plus larges d'arguments interprétatifs construits sur et à travers un riche champ de discours⁶ ». Dans la mesure où il n'existe aucun « œil innocent⁷ », tout ce qui est interprété par un humain et présent dans une archive ne peut l'être sans un certain niveau d'interprétation. Montrer le processus de création de captas nécessite alors de représenter la façon dont les phénomènes sont construits dans les sources puis réinterprétés par l'historien et transformés en données informatiques historiques.

Dans cette perspective, l'enjeu final de la donnée, qu'il soit relatif à son analyse, à sa diffusion sous diverses formes carto et/ou graphiques, à son analyse, devrait comprendre une notice relative à la qualité. Pour y parvenir, l'objectif de l'atelier est double. Il s'agit dans un premier temps de

¹ On peut néanmoins voir un renouveau quantitatif en histoire française et anglo-saxonne. Claire Lemercier et Claire Zalc, *Méthodes quantitatives pour l'historien*, Paris, La Découverte, 2008 ; Steven Ruggles, « The Revival of Quantification : Reflections on Old New Histories », *Social Science History*, 45-1, 2021, p. 1-25.

² Citons les projets ANR [DAPHNE](#), [DAI-CR&TDHI](#) ou [SODUCO](#), et dans une moindre mesure les projets [PORTIC](#) ou [FermeGe](#).

³ Jean-Philippe Genet, « Histoire, Informatique, Mesure », *Histoire & Mesure*, 1-1, 1986, p. 7-18, ici p. 11.

⁴ Cédric du Mouza, Stéphane Lamassé et Philippe Rygiel, « De l'histoire numérique à l'histoire données ? », *Les Cahiers de Framespa. e-STORIA*, 42, 2023.

⁵ J.-P. Genet, « Histoire, Informatique, Mesure », *art. cit.*, p. 8.

⁶ Johanna Drucker, *Visualisation : l'interprétation modélisante*, Paris, Éditions B42, 2020, p. 82.

⁷ *Ibid*

conceptualiser ce que peut être la qualité des données historiques informatiques, c'est-à-dire leur conformité aux usages prévus en histoire, archéologie..., dans les modes opératoires, les processus, les prises de décision et la planification. Il s'agit en parallèle de produire une connaissance qui va servir de base à l'élaboration d'un livre blanc par le consortium Projets Time Machine⁸.

Plusieurs questionnements ont été préalablement identifiés pour servir d'assise à la réflexion :

- *un travail de sémantique pour définir, clarifier et caractériser ce que l'on entend par données en histoire* : de quoi parle-t-on lorsque l'on évoque ces « données » ? D'informations présentes dans les archives, comme le fait abondamment la littérature scientifique depuis – au moins – le XIXe siècle ? S'agit-il au contraire de données numériques portant sur le passé, c'est-à-dire de la représentation d'une information historique dans un programme informatique ? Ou peut-être de l'un ou de l'autre, dans un mélange des genres qui ne peut être compris qu'en interprétant le contexte ? La polysémie du terme empêche de le penser dans l'ensemble des contextes de production de l'information historique, génère des biais et bloque sa conceptualisation.
- *un travail sur la forme des données* : à quelle question est-ce que ça doit permettre de répondre ? Utilisation ? Forme ? Flou général sur la qualité. Comment définir des critères de qualité. Poser à plat les différentes visions ? Est-ce que c'est FAIR ? Est-ce qu'on a une qualité en soi ? Quel lien entre les informations présentes dans les archives et leur constitution en données informatiques ? Comment identifier ce qui relève du langage étique et du langage émique ? Il y a une confusion entre précision, incertitude, et qualité de la donnée : comment y faire face ?
- *une question sur les faits historiques, en lien avec la question précédente* : comment les construire ? Comment documenter l'articulation entre factioïde et fait construit ?
- L'analyse des données historiques impose de raisonner en contexte d'information imparfaite. Lacunes, ambiguïtés et incertitudes sont inhérentes aux sources, et doivent être intégrées non seulement dans la structuration des données, mais aussi dans les méthodes d'analyse. Cela suppose d'adopter des formes d'inférences prudentes, de documenter les hypothèses, et de rendre visibles les marges d'erreur, afin que l'incertitude devienne un paramètre maîtrisé du raisonnement historique, plutôt qu'un obstacle.
- *un travail sur la reproductibilité de la recherche* : quels formats génériques adopter pour permettre la réutilisation des données ? Quels choix épistémologiques ? Quelle accessibilité technique pour les praticiens d'une discipline qui ne se pense pas comme instrumentée – l'histoire – et quels sont les apports d'une autre (l'archéologie) qui a fait le choix opposé ? Comment documenter la production de connaissances en conservant la traçabilité des sources et des méthodes mobilisées.

⁸ <https://ptm.huma-num.fr/>