arbido

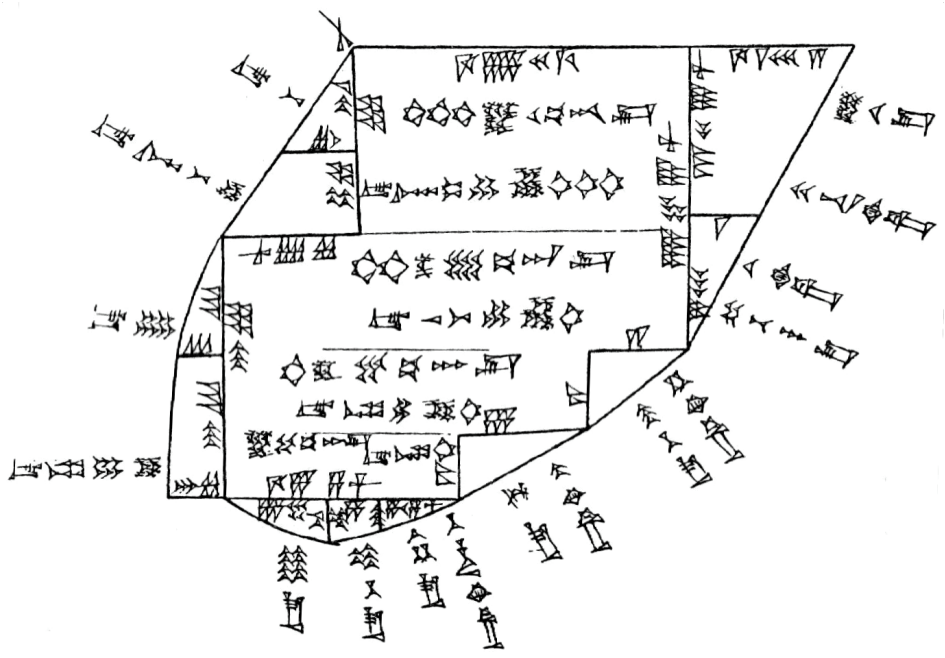
2021/4 Le document, c'est quoi?

Ourednik André, Collaborateur scientifique, Service des analyses historiques, Archives fédérales

Tags: archives big data, électronique,

L’écologie documentaire et l’inconscient réticulaire des institutions

Les documents numériques sont de plus en plus des ensembles de données imbriquées, qui doivent être compris comme une articulation de plusieurs fichiers qui en représentent les divers aspects. Leur contextualisation exige l’élaboration d’un réseau de connaissances qui tombera toujours davantage sous la responsabilité des archives, de manière à épargner les ressources des clients et des équipes de recherches. Les historiennes du futur s’intéresseront peut-être moins au contenu de nos échanges qu’à notre manière de les structurer.



Plan de terrain d’Ur III montrant la région autour de la ville de Šulgi-sipa-kalam-ma. Source: Thureau-Dangin F. 1897. «Un Cadastre Chaldéen.» RA 4 (1977 reprint), 13-27; p. 13, fig. 1.

Aujourd'hui, déjà, la recherche se tourne vers des réseaux entiers de références et de concepts partagés par une pluralité de documents[1]. Même en en considérant un seul, ses liens à d'autres sources s'imposent à l'analyse. Aucun document n'est un atome. En empruntant au vocabulaire d'Ilya Prigogine, il faut plutôt l'apparenter à une «structure dissipative»: une entité vivante et signifiante seulement dans l'échange avec son environnement. L'interaction des documents forme des écosystèmes de sens, conditionne le fait social et se cristallise dans sa trace historique.

Subjectivité des données statistiques

Une part croissante des documents actuellement produits sont d'ailleurs eux-mêmes des structures de données: tableaux (CSV, Excel), bases relationnelles (SQL), données imbriquées (JSON, XML) autant de témoignages d'une *pratique cadastrale* qui génère des isolats mesurables qu'elle ordonnance et articule. Le sociologue Armin Nassehi montre bien en quoi nos sociétés fondent leur existence sur l'*autovisibillité* garantie par un *système* de catégories statistiques résultant d'une telle pratique[2]. Il exagère seulement la modernité du dispositif. Des tablettes de la troisième dynastie d'Ur attestent déjà d'un esprit démarcatoire, de cette même pratique cadastrale, que les empires d'Eurasie finirent par imposer aux innombrables aspects de nos vies comme aux espaces nomades de leurs colonies. Le Code de Hammurabi stratifie la société sumérienne en hommes libres nobles, communs, femmes et esclaves. Aux USA jusqu'en 1865, un esclave comptait pour ? d'un homme libre dans les additions du *Federal Census Bureau*[3]. En Suisse, l'Office fédéral de la statistique (OFS) distingue à ce jour les individus «selon le sexe, le lieu de naissance ... la catégorie de nationalité et l'autorisation de résidence»[4]. Dans les années 1930 et 1940, la fiche créée pour les réfugiés par la Division de la police fédérale contenait la rubrique «race»; une directive stipulait que cette rubrique «doit être inscrite selon la loi de l'État d'origine – aryen, non aryen, nègre»[5]. Les statistiques américaines retiennent encore la «race» et «l'ethnie»; les statistiques françaises nient cette distinction sans remédier pour autant aux inégalités flagrantes induites par le rôle que lui fait jouer la *société* française.

Ces quelques exemples bien connus montrent que les documents numériques — fichiers structurés, bases de données — n'offrent pas une vision d'ensemble plus objective que d'autres sources. Comme n'importe quel document, une base de données atteste d'une perspective; elle transporte un imaginaire social[6] sédimenté dans son ontologie, c'est-à-dire dans l'ensemble des individus et des catégories avec lesquelles elle opère. Son interprétation implique forcément un contexte[7]. Comme tout document, un tableau statistique n'enrichit notre connaissance qu'à l'issue d'un processus herméneutique circulant entre l'expression singulière que représente ce tableau, et notre connaissance générale de la société qui l'a produit[8].

Traitement algorithmique des documents numériques

Ce qui distingue les documents numériques des ressources plus traditionnelles est la possibilité *immédiate* d'un *traitement algorithmique*. L'écueil persistant d'un tel traitement est le positivisme qui s'installe dès que l'historienne-programmeuse oublie qu'elle n'opère pas avec des faits mais avec des constructions sociales; lorsque, obnubilée par la technicité des tests de corrélation et de scalabilité des modèles de l'intelligence artificielle, elle cesse de questionner les logiques de production de ses «données brutes». Mais l'historienne-programmeuse est libre de procéder autrement. Les outils contemporains permettent de développer une véritable herméneutique des données tabulaires fondée précisément dans l'ancrage du document dans un *réseau contextuel*. Cela en commençant par articuler les intitulés des colonnes de tableaux aux significations de ces intitulés, à l'identité de l'instance collectrice, à la méthodologie de l'étude… Rien n'empêche ensuite d'étendre cette contextualisation à l'échelle internationale, en précisant les liens de correspondance entre les nomenclatures nationales, tenant même compte de leurs mutations asynchrones. Offrir des métadonnées d'une telle complexité implique l'adoption de *bases de données de type graphe*, qui permettent non seulement d'archiver des structures imbriquées et disparates (le sens de telle variable varie avec les années, tel document comporte plusieurs tables, tel autre est expliqué par un schéma…), mais aussi de lier, entre elles, les données associées à plusieurs documents conservés dans plusieurs archives et bibliothèques distinctes. L'adoption des modèles du web sémantique (*Records in Contexts*, *Resource Description Framework*, *Open Linked Data*) par nos institutions répond exactement à ces besoins.

Explorer l'inconscient

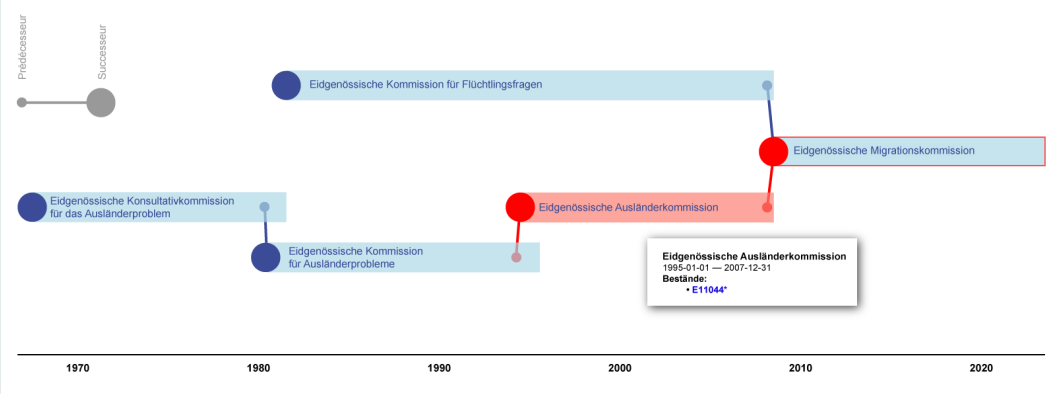
Et alors que la formalisation rhizomique du contexte permet d'éclairer l'évolution du regard conscient des administrations sur les administrées, d'autres méthodes numériques plongent jusque dans leur inconscient. Songeons notamment aux techniques du *text mining* et du *remote reading*. L'ensemble du fonds des «rapports politiques et militaires»[9], par exemple, peut être étudié en termes des fréquences et des cooccurrences de mots; et dévoiler l'effet de la formation des ambassadeurs sur le degré d'émotion qu'ils laissent, ou pas, s'exprimer dans leur écriture[10].

La ramification du contexte et le traitement par lots n'attestent pas d'une simple évolution technique, mais s'inscrivent dans un élargissement du regard, embrassant une pluralité articulée de documents, déjà initié par l'École des Annales. Leur technicité permet seulement de préciser des exigences en termes de stockage et d'accès.

Pour une préservation de la structure d’origine des documents

Le *remote reading* dépend notamment des téléchargements massifs. De trop nombreuses interfaces web d’archives conduisent encore à la commande de documents individuels, alors que des usagers toujours plus nombreux demandent des fonds entiers! Et même ceux-ci demeurent incomplets si l’on omet de conserver la structure d’origine des textes. Le format PDF – prisé pour sa fidélité photographique à l’original – détruit intégralement cette structure: des sauts de ligne typographiques introduisent des coupures logiques, les titres courants et les numéros de page se mélangent au corps du texte, les titres des sections et des chapitres deviennent des paragraphes indistincts du reste. Cette perte nuit à l’interprétation d’un document juridique pourvu d’une structure stricte (chapitres, dates, références à d’autres lois et articles de jurisprudence…), comme à celle d’un volume médiéval doté de gloses. L’analyse n’est possible qu’au prix d’un prétraitement coûteux, souvent refait indépendamment par chaque groupe de recherche. Pour remédier à ce travail de Sisyphe, les Archives fédérales suisses proposent entre autres une version XML du Recueil officiel des lois fédérales, publié en parallèle des documents PDF[11], et permettant d’archiver des informations sémantiques à même le corps du texte brut résultant d’une simple reconnaissance optique de caractères. L’archivage systématique de fichiers XML permettrait aussi de préannoter les *named entities* (patronymes, toponymes, dates…), voire de lier ces derniers, une nouvelle fois, à des sources de données tierces. Dans les archives numériques, *un document devient une articulation de plusieurs fichiers* qui en représentent les divers aspects.

Héritage institutionnel de la Commission fédérale des migrations, visualisé par l’application des Archives fédérales suisses «Réseau historique des autorités». Avec lien au fonds archivistique d’une autorité précédente. Chaque élément de ce graphe pourrait être lié à d’autres données.



Vers une sémantisation des contenus et la mise en réseau des données

Le futur de l'archivage sera celui d'une sémantisation des contenus et de la mise en réseau des métadonnées et des données primaires. L'imbrication des documents dans un *réseau de connaissances* tombera toujours davantage sous la responsabilité des archives. La collaboration d'icelles sera essentielle pour le développement de ce réseau. Les usagères, quant à elles, seront invitées à participer par l'intermédiaire d'interfaces collaboratives. Imaginons les gains en temps de recherche documentaire, le gain en rigueur scientifique dans un futur où la fiche informative d'un document listera aussi tous les articles qui le citent, ainsi que les autres ressources portant sur les mêmes personnes physiques ou morales. Imaginons que les métadonnées d'un document donnent d'emblée accès à ses diverses représentations, de même qu'au contexte de sa rédaction et de son interprétation!
Et ce faisant, réjouissons-nous de ce que notre manière d'organiser les documents en réseaux sémantiques révélera, à son tour, de notre inconscient institutionnel aux historiennes du futur.

1 Songeons notamment aux centaines recherches consacrées aux réseaux bibliométriques et leurs outils, tels VOSViewer ou CiNetExplorer.

2 2019, *Theorie der digitalen Gesellschaft*, C. H. Beck.

3 Desrosières A., 2000, *La politique des grands nombres : histoire de la raison statistique*. La Découverte, p. 234.

4 OFS, Population résidante permanente selon le sexe, le lieu de naissance, la durée de résidence, la catégorie de nationalité et l'autorisation de résidence, de 2010 à 2020.

5 Koller G., 1999, Der J-Stempel auf schweizerischen Formularen, in: *Schweizerische Zeitschrift für Geschichte*, 49/3, p. 371-374.

6 Castoriadis, C. (1975). *L'institution imaginaire de la société*. Éd. du Seuil.

7 Par exemple celui des opérations d'assignement des individus à des catégories dans les registres administratifs Cf. Desrosières (*ibid*.).

8 Dilthey, W. (1910). *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften* (1970th ed.). Suhrkamp.

9 AFS, fonds E2300* *Eidgenössisches politisches Departement: Politische und militärische Berichte der Auslandvertretungen* (1848–1965).

10 Ourednik, A., Koller, G., Fleer, P., & Nellen, S. (2018). Feeling Like a State. The Sentiments-tide of Swiss Diplomacy through the Eye of the Algorithm. Administory. *Zeitschrift Für Verwaltungsgeschichte*, 3(1). https://doi.org/10.2478/ADHI-2...

11 Amtliche Sammlung des Bundesrechts (BS / AS) 1948-2018

André Ourednik

Il est *data scientist* aux Archives fédérales suisses et enseignant au Collège des Humanités de l'EPFL et en représentation visuelle du territoire à l'Université de Neuchâtel. Il a publié notamment le *Wikitractatus*, un essai hypertextuel (Hélice Hélas, 2014), ainsi que les essais *Hypertopie: de l'utopie à l'omniscience* (La Baconnière, 2019) et *Robopoïèses: les intelligences artificielles de la nature* (La Baconnière, 2021).

Résumé

Français

Dans les archives numériques, un document devient une articulation de plusieurs fichiers qui en représentent les divers aspects. Une part croissante des documents eux-mêmes consiste en données imbriquées. Leur contextualisation exige l'élaboration d'un réseau de connaissances qui tombera toujours davantage sous la responsabilité des archives, de manière à épargner les ressources des clients et des équipes de recherches.