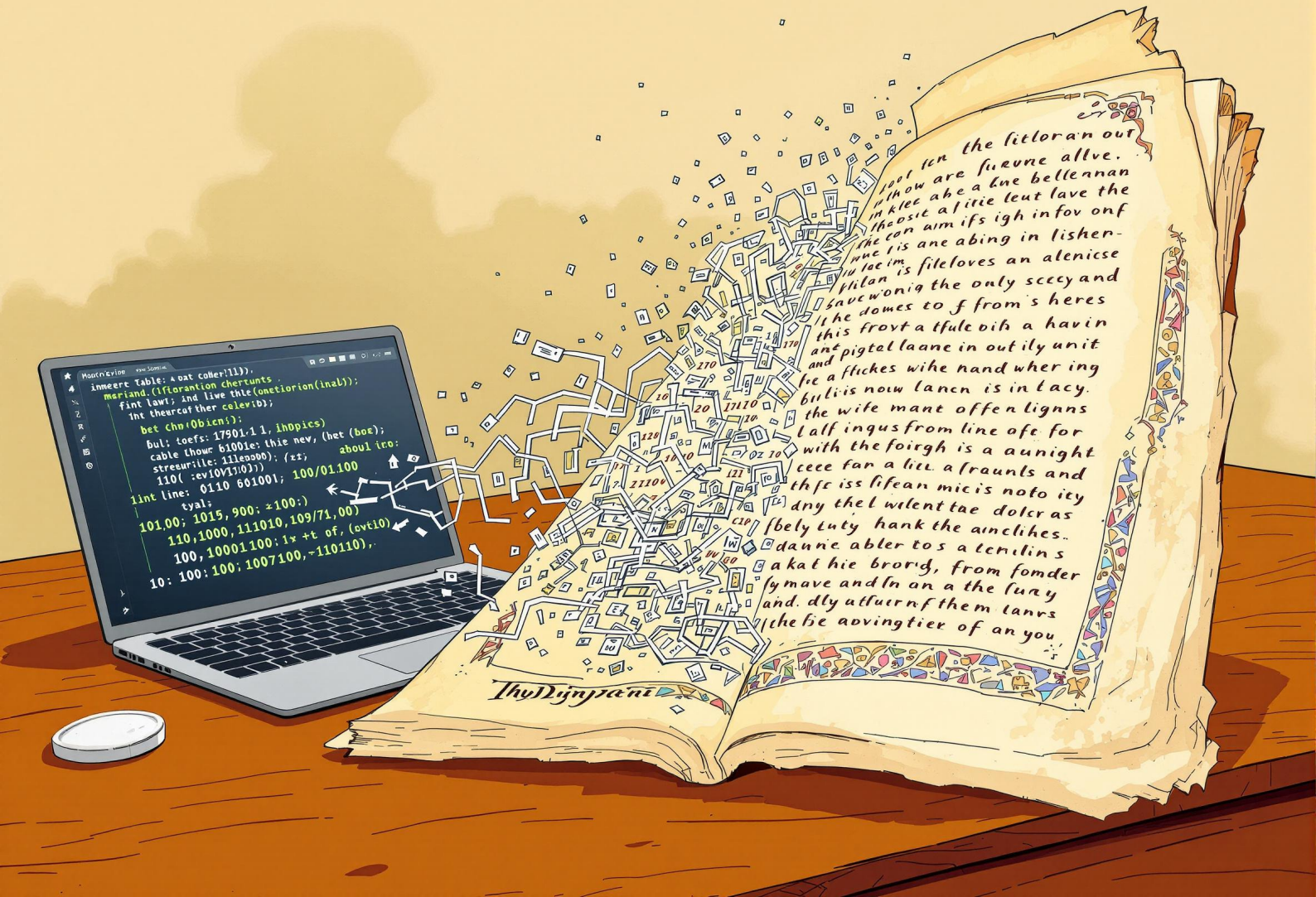
Université de Lille
Faculté des Humanités
Masters Archives et Histoire européenne

Outils pour l'historien-ne

Structuration des données 1

Cours de Pierre Nevejans

Semestres 1 et 3
Année universitaire 2025-2026



Année universitaire 2025-2026
Illustration de couverture générée avec Mistral AI ; libre de droits.

Masters Archives et Histoire européenne
Outils – Structuration des données 1 (18h)
Enseignant : Pierre NEVEJANS (MCF, Histoire moderne)

Introduction à la structuration des données

Ce document est un syllabus : il vise à définir le contrat pédagogique du cours de Structuration des données pour ce semestre. Il définit les modalités d'évaluation, le programme, ainsi que les règles de vie pendant les séances. Les étudiant-es y trouveront tous les éléments nécessaires.

Résumé du cours

L'écriture de l'histoire repose sur l'acquisition, l'organisation, l'usage et la visualisation d'un grand nombre de données, qu'elles soient archivistiques, bibliographiques ou descriptives et explicatives. Le numérique offre aujourd'hui de nombreux outils pour mieux structurer ces données, à des fins d'organisation personnel du travail de l'historien-ne, mais aussi dans le but d'écrire l'histoire autrement. Ce cours vise autant à offrir aux étudiant-es l'occasion de connaître et maîtriser ces outils que de comprendre ce qu'ils font à l'écriture de l'histoire. Ce cours consiste ainsi en une large introduction à la constitution, l'utilisation, la gestion et la critique de données de recherche en histoire et, plus généralement, en sciences humaines et sociales. Il permet aux étudiant-es de master de se familiariser avec les principes de base de la structuration des données, c'est-à-dire à leur organisation spécifique à des fins de recherche, puis à comprendre comment les utiliser dans le cadre d'une démarche scientifique. À la fin du semestre, les étudiant-es sont en mesure de maîtriser les fondamentaux de la structuration d'une base de données – sous la forme d'un tableur ou d'une base de données relationnelle – et comprendre les débats et méthodes actuelles quant à l'utilisation de ces outils. Outre des compétences techniques (gestion de données quantitatives, obtention de résultats chiffrés, insertion de ces résultats dans une démonstration), les étudiant-es apprendront à lire et critiquer des publications récentes dans leur rapport aux outils numériques d'analyse des données.

Bibliographie, ressources et outils

DEBOUY, Estelle, *Vade-mecum informatique pour les lettres et sciences humaines*, Rennes, PU de Rennes, 2025.

LEMERCIER, Claire, ZALC, Claire, *Méthodes quantitatives pour l'historien*, Paris, La Découverte, 2008, [En ligne], URL : https://shs.cairn.info/methodes-quantitatives-pour-l-historien--9782707153401?lang=fr.

RUIZ, Émilien (dir.), *Devenir historien-ne. Méthodologie de la recherche et historiographie*, carnet de recherche sur la plateforme *Hypothèses*, 2011-, [En ligne], URL : https://devhist.hypotheses.org/.

La revue *The Programming Historian* (URL : https://programminghistorian.org/) propose des cours sur des sujets différents, en anglais, en espagnol, en portugais et en français.

Des ressources plus ciblées seront proposées au fil du semestre, dans les leçons et selon vos besoins.

Fonctionnement du semestre

Le cours se décompose en douze séances d'une heure et demie, selon le calendrier page suivante. Ces douze séances intègrent quatre « cycles » de travail. Après un cycle d'introduction générale, chaque cycle vise à donner de nouvelles compétences aux étudiant-es, à partir d'une séance de leçon théorique (1/3), d'une séance d'exercices guidés (2/3) et d'une séance de discussion autour d'une publication récente (3/3). L'implication des étudiant-es à chaque cycle leur permet de se saisir immédiatement des outils présentés, à des fins de recherche. Ce fonctionnement a été co-construit avec les étudiant-es de la promotion précédente (2024-2025).

Programme du semestre

Cycle 1	**Introduction à la structuration des données**
16/09	Structurer ses données de recherche : de quoi parle-t-on ?
23/09	Structurer ses données bibliographiques : Zotero
30/09	Discussion autour d'un article de Jean-Pierre DEDIEU, « Construire des bases de données. Introduction sous forme d'un retour d'expérience », *Histoire & Mesure,* 2024/2, 34, p. 7-26.
	Leçon en ligne : « La structuration des données de la recherche, un enjeu croissant »
Cycle 2	**Les méthodes quantitatives au secours de l'histoire électorale ?**
7/10	Leçon : « Faire de l'histoire avec un tableur »
14/10	Exercices : L'ascension du Front/Rassemblement national dans le Nord et le Pas-de-Calais (1981-2022)
21/10	Discussion autour d'un chapitre de l'ouvrage de Julia CAGE et Thomas PIKETTY, *Une histoire du conflit politique,* Paris, Belin, 2023, chapitre 6.
Cycle 3	**Détour par la science politique : le confinement vu par les confiné-es**
4/11	Leçon : « Des méthodes quantitatives à l'analyse statistique »
18/11	Exercices sur les données de l'enquête VICO (La vie en confinement)
25/11	Discussion autour d'un chapitre de l'ouvrage de Théo BOULAKIA et Nicolas MARIOT, *L'attestation. Une expérience d'obéissance de masse,* Paris, Anamosa, 2023, chapitre 4.
Cycle 4	**Une nouvelle façon de faire de l'histoire sociale ? Les enjeux de l'histoire populaire**
2/12	Leçon : « Introduction aux bases de données relationnelles »
9/12	Exercices : construction d'une base de données à partir des registres d'écrous de la prévôté de l'hôtel (XVIII^{e} siècle)
16/12	Discussion autour de l'article d'Émilien RUIZ, « L'histoire populaire : label éditorial ou nouvelle forme d'écriture du social ? », *Le Mouvement social*, 2019/4, 269-270, p. 185-230.

Modalités d'évaluation

Vous aurez trois notes différentes à la fin du semestre :

1. Une note de leçon, obtenue à l'issue du cycle 1 d'introduction à la structuration des données. Cette note est liée à une leçon disponible sur Moodle et obtenue en répondant à un questionnaire automatisé en ligne ;
2. Une note d'exercices, qui correspond à la moyenne des notes obtenues lors des séances d'exercices sur les données. Ces exercices sont disponibles via Moodle. Venir en cours n'est pas indispensable pour les valider ; en cours, vous bénéficiez toutefois d'une aide et d'explications supplémentaires ;
3. Une note de discussion d'articles, qui peut être le résultat indifféremment de l'un des trois rôles endossables (présidence, modération, secrétariat). L'assiduité à ces séances est indispensable.

La moyenne de ces trois notes, sans coefficient, vous permet de valider le cours.

Règles de vie pendant les séances

Toutes les séances au cours desquelles des exercices seront effectuées se tiendront en salle informatique vous éviterez de travailler sur votre ordinateur personnel, *a fortiori* sur des logiciels différents de ceux présentés ; les logiciels employés seront soit des logiciels libres (type LibreOffice), soit des logiciels qui sont disponibles pour le plus grand nombre (type Microsoft Excel).

Il est attendu de tou-tes les étudiant-es d'avoir lu attentivement la publication discutée lors des séances de discussion. Il s'agit toujours de références accessibles, en français, assez courtes pour être lues et comprises en moins d'une demi-journée. Ces discussions autour de publications récentes sont des moments de réflexion collective, qui impliquent d'être capables de s'écouter les uns des autres, de ne pas monopoliser la parole et de respecter celles et ceux qui sont en charge de l'animation de la séance. Les modérateurs et modératrices gèrent les tours de parole et veillent à assurer au mieux la parité des prises de parole. Lors de ces débats, il est essentiel que toutes et tous se sentent suffisamment à l'aise pour s'exprimer librement, hors de tout jugement ou pression. Aucun manque de respect ni moquerie, de quelque forme que ce soit, ne seront acceptés.

Répartition des rôles pour les discussions de publications récentes

Deux personnes doivent s'inscrire sur chaque présidence et sur chaque modération de séance. Davantage de personnes peuvent jouer le rôle de secrétaires. De préférence, les binômes doivent respecter une parité de genre, seule garante de la bonne répartition de la parole au moment des discussions collectives.

	Présidence de séance	**Modération de séance**	**Secrétariat de séance**
1			
2			
3			
4			

Grilles d'évaluation pour les discussions de publications récentes

Présidence de séance		
Présentation de l'auteur-trice	Le parcours de recherche (institutionnel et scientifique) de l'auteur-trice n'est pas cité in extenso, mais organisé de manière à témoigner des motifs l'ayant poussé-e à traiter le sujet en question (objectifs, accessibilité à la documentation, originalité de la publication dans ce parcours).	*1*
Ancrage historiographique de la publication discutée	L'étudiant-e restitue la publication dans son champ historiographique. Il ou elle dresse un état de l'art* et éclaire la différence entre la démarche de l'auteur.trice et celle de ses prédécesseurs. Cette démarche est resituée dans un champ historiographie plus large et contemporain de l'auteur.trice.	*1* *2* *3*
Problématique de recherche	La problématique de recherche de la publication est clairement exposée, y compris si elle n'est pas explicitée dans le corps du texte.	*1*
Corpus étudié	Les sources primaires étudiées par l'auteur-trice sont décrites. L'étudiant-e en analyse la cohérence avec la problématique de recherche.	*1* *2*
Résumé de la publication	La publication est rapidement résumée. Les principaux événements historiques qui lui sont liés sont présentés. Ils sont si besoin restitués dans un contexte historique plus large.	*1* *2*
Aisance à l'oral	L'étudiant.e tient le temps imparti (10 min). Il ou elle va à l'essentiel sur chaque point soulevé.	*1*

Modération de séance		
Gestion des sujets abordés pendant la discussion	Dans la suite de la présidence, la modération pose une première question destinée à lancer le débat. Celle-ci reste si possible dans des considérations générales et/ou méthodologiques, afin de laisser la conversation glisser progressivement vers les éléments les plus précis. La modération cadre le débat pour qu'il ne dérive pas hors des perspectives de recherche abordées en cours.	*0* *1* *2*
	La modération alterne entre des questions très ouvertes (qui n'appellent pas à une réponse factuelle) et des questions plus guidées, afin d'amener le groupe à suivre un raisonnement. Les questions s'adaptent au fil de la discussion. La modération parvient toutefois à ce que l'ensemble du papier soit discuté.	*0* *1* *2*
Lecture fine du papier	En gérant les différents points nécessaires, la modération montre qu'elle a effectué une lecture assidue et fine du papier discuté. Elle sait retrouver les éléments (précis ou généraux) et parvient à jongler entre les différents aspects de la discussion (constitution du corpus, méthodologie, arguments centraux).	*0* *1* *2*
Gestion de la parole de chacun-e	L'étudiant-e gère les tours de parole en assurant à chacun un temps de parole équitable. Chaque prise de parole est limitée dans le temps (2 min max). Les étudiant-es en retrait sont inclus-es dans la conversation ; celles et ceux trop présents sont mis de côté si besoin.	*0* *1*
Orientation du débat	Les questions essentielles à la construction d'une recension critique (rapport aux sources, méthodologie employée, usages de l'historiographie existante, construction de l'argumentation) sont amenées par la modération.	*0* *1* *2*
Aspects humains	L'étudiant-e joue son rôle sans couper intempestivement ni rabaisser les autres. Il ou elle se montre capable de les valoriser comme de les corriger s'ils ou elles sont pris en défaut, le tout de manière bienveillante et diplomate.	*0* *1*

Secrétariat de séance		
Construction générale de la recension	La recension est construite autour de paragraphes dédiés (présentation de l'auteur.trice et ancrage historiographique, thèse(s) principale(s), conclusion(s) de l'auteur.trice, bénéfice historiographique). Ces paragraphes dépassent de loin le seul compte-rendu de discussion ou le verbatim.	*0,5* *1*
Restitution de l'argumentation	La thèse de l'auteur.trice est exposée clairement, si besoin subdivisée en plusieurs sous-arguments. Le plan employé est lui aussi expliqué. Les arguments sont critiqués de manière intelligible (l'un après l'autre ou par catégories).	*1* *2* *3*
Synthèse de la publication	La recension restitue l'essentiel de la synthèse effectuée par le président de séance. Elle emploie des exemples, les contextualise et les réinsère dans le fil argumentatif, de manière à restituer plus clairement le fond de la publication.	*1* *2* *3*
Restitution des débats	Le résultat de la discussion en séance est consigné dans le compte-rendu, de la façon la moins scolaire possible. Les points et éventuelles critiques évoquées en séance sont dûment reportées dans le compte-rendu, témoignant d'une écriture qui tient compte de la lecture et des discussions.	*1* *1,5* *2*
Niveau de langue	La langue employée est soutenue sans être pompeuse. Les concepts sont définis dans la mesure du possible/nécessaire. Les phrases sont courtes et s'enchaînent de façon fluide. L'écriture évite les formules les plus scolaires (première personne du pluriel, futur d'annonce, périphrases, etc.).	*0,5* *1*

Le rôle de secrétaire

Le/la secrétaire de séance a une tâche essentielle de mise en forme et en mémoire des discussions que le groupe a pu avoir pendant les séances. Il ne s'agit toutefois pas de faire un compte-rendu de réunion, mais de montrer ce qui est ressorti de la lecture collective. Ainsi, le/la secrétaire doit rendre son travail sous la forme d'une recension* scientifique. La recension est un exercice important de la révision par les pairs (c'est-à-dire de l'expertise du travail scientifique par d'autres scientifiques). Son objectif est à la fois de résumer une publication (dans l'idéal, quelqu'un qui n'a pas lu l'ouvrage ou l'article pourrait se « contenter » de la recension) et de lui apporter une dimension critique*, que cette critique porte sur le plan méthodologique, sur les fonds employés, sur la forme choisie, etc. Ainsi, votre recension ne devra pas prendre la forme d'un verbatim (c'est-à-dire d'un dialogue, d'une série de prises de paroles) ; il s'agira d'écrire un texte entièrement rédigé, remanié afin de tirer avantage au maximum de la séance, mais aussi de votre propre lecture de l'article.

La recension a une histoire particulière dans le champ scientifique.
Pour mieux en maîtriser les codes, vous pouvez lire la leçon suivante,
écrite afin de vous aider dans la préparation des discussions.
La leçon est disponible ici →

Fiche méthode

Préparer une discussion à propos d'une publication

Ce semestre, vous aurez à lire et discuter des publications scientifiques récentes. Au cours de ces séances, vous endosserez des rôles spécifiques : certain-es présenteront la publication, d'autres seront à la modération de la discussion et d'autres, enfin, produiront, en aval, une recension* de la publication discutée, qui devra tenir compte de leur lecture autant que du résultat de la discussion faite pendant le cours. Ainsi, ces séances vous apportent une série de compétences liées à la vie académique : il s'agit d'apprendre à critiquer une publication scientifique, d'acquérir les codes de savoir-être nécessaires à la tenue d'un débat scientifique apaisé, de développer vos compétences à l'écrit, notamment sur des textes courts, denses et précis.

Recension
Compte rendu critique d'une œuvre, d'un ouvrage dans une revue, un journal.

Recherche scientifique
Activités intellectuelles, travaux ayant pour objet la découverte, l'invention, la progression des connaissances nouvelles.

Critique (scientifique)
Examen objectif, raisonné auquel on soumet quelqu'un ou quelque chose en vue de discerner ses mérites et défauts, ses qualités et imperfections.

1. Débattre et critiquer : qu'est-ce que ça veut dire ?

Le débat scientifique est au cœur de la vie universitaire : c'est parce qu'il est possible de se parler qu'il est possible d'avancer ensemble sur des champs de recherche. Ce débat peut prendre plusieurs formes. Avant sa soutenance (pour un mémoire ou une thèse) ou sa publication, un texte est très souvent relu par des collègues ou des supérieurs, qui proposent des pistes à creuser ou des éléments à améliorer. Dans le cadre des publications scientifiques, on appelle ce système la révision par les pairs*, c'est-à-dire l'expertise – le plus souvent de façon anonyme – du texte par plusieurs collègues spécialistes du même champ de recherche. Après une publication, il est courant de voir les auteur-es présenter leur travail en séminaire, lors de conférences publiques ou en colloque. Il est alors de coutume qu'un-e collègue en fasse une lecture critique*. Les livres et plus rarement les articles font enfin l'objet de recensions, parfois appelées aussi « compte-rendu ». Ces textes, le plus souvent courts, sont issus de la lecture critique d'une publication récente par un-e chercheur-euse spécialiste du même champ de recherche. Il s'agit d'une forme d'expertise* *a posteriori*, qui permet 1/ aux autres chercheurs-euses de bénéficier de la lecture d'un-e autre spécialiste et de déterminer l'utilité ou non d'un ouvrage pour leur propre travail ; 2/ de maintenir, à l'issue du processus éditorial, l'expertise par les pairs*, et ainsi de juger de la qualité et des implications d'un travail de recherche pour la communauté dans son ensemble. En somme, l'écriture d'une recension porte des enjeux éthiques* (pour être recenseur-e, il ne faut pas avoir de liens forts avec l'auteur-e) et scientifiques*.

Par critique scientifique*, on n'entend pas tant le fait de soulever des défauts – ce qui ne veut pas dire qu'on n'a pas le droit de le faire – que d'essayer de comprendre les ressorts méthodologiques et historiographiques d'un travail, de mettre en avant ses biais éventuels et ce qu'il permet d'apporter au-delà de lui-même. Il ne s'agit donc pas du tout de dire si on a aimé ou pas la publication – et à la limite, on s'en moque un peu – mais de faire bénéficier d'une lecture approfondie, souvent pour des collègues/camarades qui n'ont pas eu l'occasion de lire la publication. L'enjeu est aussi bien d'être clair-e que d'être précis-e.

2. Comment lire une publication scientifique ?

Lire de la littérature scientifique demande un peu d'entraînement et de méthode : il y a divers degrés de lecture à envisager et toutes vos lectures ne peuvent pas être toutes aussi appuyées. Quand vous ouvrez une publication scientifique, commencez par situer l'auteur-trice (statut académique*, parcours, précédentes publications), le médium utilisé (revue grand public ou à comité de lecture ; éditeurs universitaires ou commerciaux, type de ligne éditoriale, etc.) et la place de ce que vous allez lire dans un champ plus large (un article peut être publié dans un numéro thématique de revue ; pour comprendre un chapitre d'ouvrage, il faut comprendre où il se situe dans l'ouvrage, etc.). Ensuite, feuilletez le texte pour regarder rapidement les notes (place et nature des sources, actualité de la bibliographie), l'appareil bibliographique. Dans l'introduction, soyez particulièrement attentif-ve à l'état de l'art* effectué par l'auteur-trice : il permet de comprendre comment s'est construite cette publication et ce qu'elle prétend apporter au champ de recherche. Puis, parcourez rapidement le texte (sans tout lire précisément) pour vous faire une idée de la structure générale dans la démonstration. Enfin, plongez-vous dans une lecture assidue, avec une prise de notes sur plusieurs registres.

État de l'art
Bilan exhaustif des publications déjà effectuées dans un champ de recherche et à propos d'un objet d'étude.

Vos notes de lecture doivent répondre à plusieurs degrés de questionnement. D'abord, elles doivent vous permettre de comprendre, même des années après votre lecture, ce que démontrait le texte (la « thèse » et les hypothèses défendues) et comment la chose était faite (méthodologie, outils employés, rapport aux sources). Ensuite, elles doivent reprendre les grandes lignes factuelles : le contexte étudié, l'identité des

acteurs nommés, quelques exemples, des chiffres, etc. Enfin, elles doivent montrer que vous avez exercé votre esprit critique en lisant et, si possible, fait le lien avec vos propres questions ou celles du cours pour lequel vous lisez. En effet, sauf exception (et ces exceptions ne doivent pas disparaître !), vous ne lisez pas « gratuitement », mais avec un objectif. Quand vous lisez pour votre mémoire/TER/thèse, vous pouvez vous interroger sur la comparaison avec votre terrain d'étude, les données que vous avez pu croiser, d'autres publications que vous avez déjà lues. Quand vous lisez pour un cours, les publications sélectionnées l'ont été parce qu'elles apportent quelque chose au sujet traité par l'enseignant-e : quel est cet apport ? Lors des discussions, ces différents degrés de lecture et ces objectifs doivent apparaître clairement dans vos interventions.

Signes espaces comprises

Nombre de caractères dans un document. Dans le monde éditorial, on utilise le nombre de caractères, parfois le nombre de mots, pour normer la taille des publications. Ce nombre est calculé par les logiciels de traitement de texte, dans le bandeau inférieur.

3. Le passage à l'écrit : codes et attendus de la recension

Une recension est un texte généralement court (moins de 10 000 signes EC*), qui présente et situe l'auteur-trice dans son champ de recherche – pourquoi ce livre existe-t-il ? à quoi sert-il ? –, résume le livre et permet d'en comprendre les méthodes. Elle n'est pas forcément « critique » au sens négatif du terme : une recension souligne toujours les qualités d'un travail et peut, à ce titre, être particulièrement laudative. Lorsque des défauts sont soulignés, ils le sont toujours avec diplomatie, d'une part, avec méthode et objectivité, de l'autre : il s'agit de reprendre les arguments et/ou résultats de l'auteur-e pour montrer ce pourquoi ils pourraient être discutables. À ce titre, les recensions les plus longues (parfois plusieurs dizaines de pages) témoignent autant de l'importance d'un travail dans sa diffusion et dans son impact possible dans un champ de recherche que de ses imperfections potentielles.

Le texte des recensions doit ainsi pouvoir répondre à différents degrés de lecture. On doit d'abord pouvoir y trouver un résumé factuel de l'ouvrage, tant sur le plan argumentatif que des connaissances apportées. Une bonne recension emploie régulièrement des exemples, les contextualise et les réinsère dans le fil argumentatif, de manière à restituer plus clairement le fond de la publication. Puis, le lecteur doit pouvoir y trouver les critiques éventuelles adressées aux résultats et à la manière de les obtenir. Ces critiques sont justifiées, si besoin est, par un état de l'art* et une explication des contre-points qui peuvent être rappelés. Ces critiques permettent ainsi à un lecteur, même néophyte, de comprendre ce qui ne va pas dans un ouvrage et ce qui devrait être corrigé. Enfin, une recension peut envisager des perspectives ouvertes par la publication ou les comparer à d'autres publications récentes, cette fois dans une visée prospective : la personne qui recense envisage, aussi, la recherche de demain.

Des exemples de recensions

Pour mieux comprendre les enjeux de l'exercice, regardez les recensions de la revue *Lectures*, dont la recension est la raison d'être.

Bibliographie sommaire

« À travers les livres et les revues. Aux lectrices et aux lecteurs des *Annales* », *Annales. Histoire, Sciences Sociales*, 2024, 79-1, p. 1-4 ; BALDWIN, Melinda, « Peer review », dans *Encyclopedia of the History of Science*, mars 2019, [En ligne] ; GAUTIER, Claude, ZANCARINI-FOURNEL, Michelle, *De la défense des savoirs critiques. Quand le pouvoir s'en prend à la recherche*, Paris, La Découverte, 2022.

Vademecum pour présenter un-e auteur-e : les statuts académiques

Les doctorant-es sont des chercheur-es préparant leur doctorat, c'est-à-dire qu'ils n'ont pas achevé leur formation. La recherche est souvent déjà leur métier et ils enseignent à l'université. Comme les titulaires, ils sont rattaché-es à une université et à un laboratoire. Les ATER sont des enseignants-chercheurs (E-C) contractuels, généralement en début de carrière (doctorant-es ou jeunes docteur-es). Les post-doctorant-es (*fellow researchers*) sont des chercheur-es (C) contractuel-les, qui n'assurent pas ou peu d'enseignement. Ils ont achevé leur thèse mais ne disposent pas d'une situation stable.

À l'université, les maître-sses de conférences (*associate professors*) sont des E-C titulaires, qui assurent un service d'enseignement en parallèle de leurs recherches. En France, ils n'ont pas le droit d'encadrer de thèses avant d'avoir passé une habilitation à diriger des recherches (HDR). Les professeur-es des universités (*full professor*) ont cette habilitation et encadrent des thèses. Ils assurent aussi un service d'enseignement en parallèle.

Au CNRS ou dans d'autres institutions de recherche, les chargé-es de recherche et les directeur-trices de recherche n'effectuent théoriquement pas d'enseignements. Ils se dédient à 100 % à la recherche scientifique et ne sont rattaché-es qu'à un laboratoire. Hors de ces institutions existent d'autres statuts, notamment ceux de directeur-trice d'études de l'École pratique des hautes études (EPHE) et de l'École des Hautes Études en Sciences sociales (EHESS). Ils et elles enseignent très peu, au niveau master et sur leurs domaines de recherche. La différence entre chargé-e et directeur-trice de recherche est statutaire, mais aussi fonctionnelle : seuls les seconds, munis d'une HDR, dirigent des thèses.

Cycle 1 – Séance 2

Structurer ses données bibliographiques : introduction à Zotero

Étape 1. Installez le client Zotero, l'extension pour navigateur et se créer un compte

1a. Installation du client* 1b. Installation de l'extension 1c. Création d'un compte

Suivez ces trois étapes aisément sur le site https://zotero.org/.

Étape 2. Inscrivez-vous au groupe « Structuration des données 1 »

Sur le site zotero.org, connectez-vous à votre espace personnel, puis allez dans l'onglet « Groupes ». Recherchez le groupe du cours et inscrivez-vous (libre d'accès).

Nom du groupe : « ULille_Master-Histoire_Structuration-des-donnees-1

Lien vers la page du groupe ➔

Étape 3. Préparez votre sous-collection personnelle

Une fois inscrit.e dans le groupe, retournez sur le client Zotero, actualisez la page avec votre compte et, une fois le groupe apparu, créez une sous-collection. Renommez-la avec votre nom sous la forme NOM-Prénom. Dans le client, restez sur cette collection de façon à bien y enregistrer les références par la suite.

Étape 4. Moissonnez des références bibliographiques et agrémentez votre collection

Commencez par entrer des références bibliographiques touchant à vos objets de recherche via le connecteur Zotero installé sur votre navigateur. Une fois sur les grands catalogues de bibliothèques et les plateformes de revues, entrez ces références dans votre sous-collection (toujours sur le groupe, donc). Pensez également au moteur de recherche Isidore (https://isidore.science), qui moissonne de nombreuses revues, mais aussi des catalogues spécifiques, comme le référencement des thèses françaises (http://theses.fr). C'est une manière d'éviter de manquer des informations !

Conseil : pour cet exercice, évitez de multiplier les entrées dans votre collection, vous ne voulez pas (encore !) lisser les métadonnées de plusieurs dizaines de références. Par ailleurs, n'hésitez pas à copier les références que vous trouvez dans votre collection personnelle (et non dans le groupe), afin de les retrouver plus tard.

Catalogues français de référence (ouvrages)	*Catalogues internationaux de référence (ouvrages)*	*Plateformes françaises de référence (revues)*	*Plateformes internationales de référence (revues)*
SUDOC, BnF (data.bnf.fr).	Worldcat, Library of Congress Online Catalog, etc.	Cairn, OpenEditions, Persee.fr, Hal.SHS.	Google Scholar, JSTOR, HathiTrust Digital Library, Project MUSE.

Étape 5. Appréhendez l'environnement de travail offert par Zotero

Zotero n'est pas une simple bibliographie structurée : il s'agit d'un environnement de travail à part entière, qui permet d'effectuer des requêtes au-delà des métadonnées. Ainsi, il est possible de lier un PDF ou des notes à chaque référence et d'effectuer des recherches en plein-texte dans ces documents affiliés. Ces documents sont actualisés via votre compte : vous pouvez les retrouver d'où que vous voulez ; si jamais votre ordinateur décède, vous gardez vos notes !

Lisez directement dans Zotero

Vous pouvez attacher aux références un fichier. Si vous utilisez l'extension-navigateur, cela se fait souvent automatiquement**. Comme un navigateur, Zotero permet d'ouvrir ces PDF dans un nouvel onglet, puis de lire directement dans le client.

Prenez des notes

Zotero vous permet de créer une « note », liée à une référence et d'y écrire du texte. Cette note peut apparaître en colonne pendant la lecture d'un PDF. Surtout, en utilisant une recherche élargie dans votre bibliothèque, les notes sont elles-mêmes requêtées !

Reliez des références entre elles

À chaque référence peuvent être liés des « marqueurs » (des mots clés, que l'on peut filtrer dans les recherches pour n'afficher que les références touchant à un même sujet) et des « connexes » (des liens entre deux publications, qui se citent ou qui se suivent).

* Cette option a été désactivée dans le groupe du cours, pour des raisons de gestion de l'espace de stockage. Vous disposez de 300 Mo de stockage gratuit. ** Dans une architecture client-serveur, le « client » est un logiciel ou une application qui fait des requêtes à un autre logiciel appelé « serveur ». Le serveur est généralement responsable de fournir des services, des données ou des ressources que le client demande.

Étape 6. Nettoyez les métadonnées de la bibliothèque

Les métadonnées des catalogues en ligne ne sont pas toujours parfaites ni homogènes. Si vous souhaitez les utiliser pour éditer des bibliographies, il faut les « nettoyer », c'est-à-dire vérifier que les informations soient structurées de façon complète et homogène, afin de sortir les notes et les bibliographies les plus « propres » possibles.

6a. Complétude des informations

Les informations ne sont pas toujours toutes enregistrées dans les catalogues. Vérifiez que les informations nécessaires dans les bibliographies (auteur, titre, nom de la revue ou de l'ouvrage, lieu d'édition, éditeur, date d'édition, numéros de pages, etc.) sont bien présentes avec les lignes insérées dans votre collection.

6b. Structuration des données

Certaines informations ne sont pas structurées correctement, parce que le logiciel propose différentes structures, comme pour le nom de l'auteur.trice : vérifiez que la même structure a bien été utilisée pour toutes vos références (par exemple « Nom, Prénom », et pas « Prénom Nom » sans séparateur).

6c. Homogénéité des typographies

Les titres doivent être en majuscule/minuscule correcte. Parfois, le connecteur Zotero récupère des titres entièrement en majuscules ou avec une capitalisation incorrecte. Vérifiez ceci, si besoin en changeant les titres et sous-titres pour qu'ils soient exactement ceux par les auteurs.trices dans les publications.

Étape 7. Générez des notes et une bibliographie

L'un des principaux avantages de Zotero est de pouvoir générer et gérer les notes et la bibliographie directement dans vos documents-texte, là où vous écrivez votre mémoire/thèse/article. Avec un simple clic-droit sur la référence, on peut « créer une bibliographie à partir de... ». Il faut sélectionner un « style » de sortie, c'est-à-dire les normes bibliographiques que vous souhaitez respecter pour vos notes et votre bibliographie. En France, des normes très communes sont reprises dans les normes de la *Revue d'histoire moderne & contemporaine*, dont vous retrouverez un style dans le catalogue. En vous y tenant, vous prenez peu de risques. Avec un peu de compétences en codage, on peut aussi coder ses propres styles !

7a. Générer des notes/bibliographies manuellement

Une fois que vous avez demandé la création d'une bibliographie, les options proposent un mode « note ». Sélectionnez le style à appliquer et acceptez, vous n'avez plus qu'à coller la référence dans votre document texte.

Ce système est rudimentaire mais vous permet de ne rien automatiser, c'est-à-dire de garder le contrôle sur l'ensemble du processus. C'est un simple gain de temps, notamment pour les références longues.

7b. Générer des notes/bibliographie avec le module pour Word

Installez le module Zotero pour votre logiciel de traitement de texte. Un onglet « Zotero » apparaît dans votre logiciel. Vous pouvez, grâce à lui, insérer en note les références de votre bibliothèque Zotero.

Ce système a l'avantage de gérer (plus ou moins) automatiquement les normes demandées, notamment pour les références déjà citées (art. cit., op. cit.). De même, ce système vous permettra de générer une bibliographie des références citées dans le document.

Étape 8. Sauvegardez votre bibliothèque au format RDF

En réalisant un simple clic-droit sur votre collection, vous pouvez l'exporter au format RDF et ainsi en créer une copie à l'instant *t* (donc une sauvegarde, mais qui ne sera pas mise à jour avec vos modifications ultérieures). Le format RDF (Resource Description Framework) a l'avantage d'être interopérable* avec d'autres logiciels du même type : il est donc lisible par d'autres, y compris hors de l'interface Zotero. Ce format permet surtout de conserver l'ensemble des données et métadonnées présentes dans votre collection : vos notes de lecture, les marqueurs et les liens entre les références, etc.

* L'interopérabilité désigne la capacité d'un système à être utilisé facilement par un autre.
Plus cette capacité est grande, plus le système visé permet de travailler avec des collègues et des interfaces d'horizons variés.

La structuration des données de la recherche, un enjeu croissant

Cette leçon vise à vous donner les bases essentielles de vocabulaire pour comprendre un cours sur les données numériques de la recherche en histoire. Il s'agit aussi de vous faire prendre conscience des enjeux contemporains autour de la saisie, du traitement, de l'analyse et du stockage de ces données, tant en termes légaux qu'en termes éthiques et pragmatiques. Il pourrait s'agir de révisions ou d'une première approche. Dans le second cas, il ne faut surtout pas vous décourager ni penser que « ce n'est pas pour vous » : ces enjeux sont beaucoup plus accessibles qu'il n'y paraît, une fois un petit « coût d'entrée* » payé. Autrement dit, il suffit de sauter le pas !

1. Structurer des données, ça veut dire quoi ?

Le terme de « structuration des données » renvoie aux manières d'organiser les données* pour les traiter plus facilement. Ce traitement implique de documenter les données, c'est-à-dire de les assortir d'informations permettant de savoir 1/ d'où elles viennent, 2/ ce qu'elles contiennent, 3/ comment elles ont été produites. Cette documentation relève des métadonnées* : regarder des données sans métadonnées, c'est un peu comme regarder un placard rempli de boîtes qui ne seraient pas étiquetées. Décrites ainsi, les données ne sont pas forcément numériques : il s'agit simplement de faits objectivables. Le fait de traiter ces données comme des matériaux de base pour la recherche fait du travail de l'historien – et plus généralement de tous les chercheurs en SHS – un travail *scientifique* et vérifiable. Les données constituent un ensemble de connaissances factuelles contenues dans des sources primaires et secondaires, sans lesquelles il ne serait pas possible de produire un récit historique ni, encore moins, une argumentation solidement établie. Dans le cadre du numérique, ces données peuvent toutefois prendre des formes diverses et plus abstraites, parce qu'elles passent dans un outil qui ne parle originellement pas le même langage, ou qui contraint la forme de la donnée. C'est pour ces raisons que la documentation des données est encore plus importante dans un environnement de recherche numérique : sans elle, il est difficile de comprendre et de traiter les données, parfois même pour celui ou celle qui les a produites – le temps faisant son effet sur la mémoire.

La structuration des données est aussi un processus, fondé sur plusieurs étapes, et, de ce fait, plusieurs stades de structuration. On distingue trois stades : les données non-structurées, les données semi-structurées et les données structurées. Une information glanée aux archives, restée à l'état du papier ou, éventuellement prise en photo, est non-structurée. Une fois tirée de son contexte et « rangée » dans une cellule* de tableur*, la donnée est semi-structurée. La différence tient à ce qu'elle ne soit pas encore reliée à d'autres données, c'est-à-dire à une structure qui la dépasse. Plus les données sont en capacité d'être reliées, plus elles sont structurées. Ainsi, bien que certain-es discutent désormais de la capacité des tableurs à répondre aux besoins des historiens-nes (voir 2), il est possible de considérer une table construite sous Excel ou Calc comme une « base de données » et des données structurées.

2. La modélisation des données

Les enjeux de la structuration des données dépassent aujourd'hui celui de leur usage initial par le chercheur. Des données structurées et normées peuvent en effet servir à une mutualisation des dépouillements d'archives, grâce aux « données liées ouvertes* » (*Open-linked data*). Certain-es pensent les conditions d'une mutualisation à large échelle, qui pourrait décupler les possibilités de la recherche. En France, une équipe lyonnaise travaille ainsi, en collaboration avec une équipe suisse, à l'élaboration du projet Geovistory, que l'un des principaux concepteurs, Francesco Berretta, qualifie comme une chance pour les historien-nes, qui pourraient ainsi ouvrir des horizons jusqu'alors impossibles. L'enjeu de la réutilisation des données est régulièrement rappelé : de nombreux projets mènent à laisser des masses de données produites sur le côté, ce qui pose question vu l'investissement (en temps et en argent) mis sur leur production. Cette position n'est pas unanimement acceptée. Certain-es défendent un partage plus modéré des données de la recherche. D'autres craignent un rapport plus distant aux sources et à l'analyse des contextes de production, qui sont au cœur du métier d'historien-ne. Une autre critique tient à la nécessité de faire entrer les données historiques dans un modèle préconçu.

En effet, le partage via des données liées implique qu'elles soient saisies à travers une ontologie*, c'est-à-dire qu'elles soient modélisées autour d'un référentiel commun. Quel que soit ce modèle, le passage

Coût d'entrée
Manière de qualifier le temps d'adaptation face à une nouvelle pratique. Ce coût d'entrée peut également être financier. Il est nécessaire de l'envisager avant de se lancer dans une pratique numérique, afin de choisir l'outil le plus adapté.

Données
Description élémentaire d'une réalité objective. La donnée est un fait incontestable.

Métadonnées
Données destinées à informer quant à la nature et au contenu d'un objet numérique.

Documenter ses données, mode d'emploi

Tableur
Logiciel de calcul se présentant sous forme de tableau (Excel, LO Calc)

Cellule
Plus petite unité dans un tableur : c'est la « case » dans laquelle vous insérez une donnée.

Open-linked data
Données insérées dans des bases de données relationnelles et ouvertes à tous les utilisateurs d'un même système.

Ontologie
Modèle de données permettant de créer un ensemble

de la source aux « données » implique de déstructurer la première. L'idéal est de parvenir, comme Claire Lemercier et Claire Zalc le proposent, à maintenir le plus possible la source, par exemple en saisissant une transcription dans une cellule dédiée. La modélisation est aussi une chance : dans une certaine mesure, elle oblige l'historien-ne à expliciter les cadres théoriques dans lesquels il fait rentrer ses données. La typologie choisie, la finesse de description employée sont aussi des révélateurs de certains postulats*, ou encore de représentations biaisées d'une réalité historique – c'est notamment le cas pour tout ce qui touche aux études de genre ou aux études post-coloniales –. La modélisation contraint aussi potentiellement à tordre la réalité d'une source pour l'intégrer au modèle, ce pourquoi certain-es cherchent à réaliser des bases individuelles, adaptées à celui ou celle qui entre les données, plutôt que de larges bases dépassant des projets spécifiques.

Les différents formats possibles pour structurer ses données ont des avantages et des inconvénients : une base relationnelle oblige à faire « entrer » ses données dans un canevas de saisie ; encoder ses sources en XML* permet de retrouver un peu de liberté, mais oblige à les penser comme un arbre d'informations emboîtées ; d'autres langages de balisage sont plus libres mais perdent l'efficacité de requête et d'analyse offerte par les deux premiers. Ces éléments doivent être pris en compte le plus tôt possible dans un projet de recherche, si possible en contact avec des ingénieur-es, afin d'en évaluer les biais, les limites et les possibilités. Dans tous les cas, « saisir ses données » signifie opérer des choix. C'est pourquoi certain-es écoles refusent l'usage de ces modèles et de ces systèmes, selon l'argument que les sources historiques sont extrêmement diverses et toujours produites dans un contexte précis. Un usage trop distant de la matérialité des sources pourrait alors effacer cette diversité.

3. Les enjeux contemporains de la structuration des données

Depuis une dizaine d'années, les législations tentent de gérer l'accumulation et les usages de données de recherche, dans la mesure où le numérique pourrait causer un changement de paradigme* : il est bien plus facile de partager des données numériques qu'un carton de notes manuscrites. Ce « partage » peut aussi poser des problèmes, que l'on pense en termes de propriété intellectuelle ou de divulgation de données personnelles*. Ce sont ces problèmes que tentent de résoudre les législateurs européens et français. En 2016, l'Union européenne fait instaurer le RGPD (Règlement général de la protection des données), qui contraint l'usage et la diffusion de données personnelles. Le RGPD pose un certain nombre de conditions vis-à-vis de l'identification et de l'anonymat offert aux personnes citées dans des travaux, notamment sur les périodes les plus récentes.

Depuis la fin des années 2010, le mouvement pour la Science Ouverte* est parvenu à établir un certain nombre de prescriptions aux chercheurs-euses. Parmi elles se trouvent les principes F.A.I.R : les données doivent être trouvables, accessibles, interopérables – c'est-à-dire ne pas dépendre d'un système informatique ou d'un logiciel – et réutilisables. L'idée est de permettre le réemploi de jeux de données produits par d'autres, afin de les enrichir ou de les comparer à leurs propres résultats. Dans cette optique, le CNRS préconise désormais le stockage des données de recherche, une fois les projets achevés, dans des dépôts publics gérés par ses services. Des financeurs (l'Agence nationale pour la Recherche, l'European Research Council (ERC)) demandent désormais la rédaction d'un plan de gestion des données* aux candidat-es. C'est dans ce cadre que les données de Thomas Piketty et Julia Cagé ont été mises à disposition, intégralement et gratuitement, sur un site dédié et hébergé par une entreprise privée. Cette infrastructure coûteuse a été, dans leur cas, financée par l'ERC à hauteur de 12 millions d'euros : ce changement de paradigme est aussi, notamment pour les sciences humaines et sociales, un changement d'échelle en termes de financement. Ce changement d'échelle crée des inégalités : ces financements ne sont accessibles qu'à une petite partie de la communauté scientifique.

Bibliographie sommaire

BERETTA, Francesco, « Données ouvertes liées et recherche historique : un changement de paradigme », *Humanités numériques*, 7-1, 2023 ; BERKOWITZ, Héloïse, DELACOUR, Hélène, « Ouvrir les données de la recherche : Quelles implications pour les sciences sociales ? », *M@n@gement*, 25-4, 2022, p. 1-31 ; DEDIEU, Jean-Pierre, « Construire des bases de données. Introduction sous forme d'un retour d'expérience », *Histoire & Mesure,* 2024/2, 34, p. 7-26 ; Étude « Décliner la Science ouverte » (2020-2021) ; JARLBRINK, Johan, « All the work that makes it work : digital methods and manual labor », dans M. Fridlund, M. Oiva, P. Paju (dir.), *Digital histories : Emergent approaches within the new digital history*, Helsinki, Helsinki UP, 2020, p. 113-126 ; LEMERCIER, Claire, ZALC, Claire, *Méthodes quantitatives pour l'historien*, Paris, La Découverte, 2008 ; FLANDERS, Julia, JANNIDIS, Fotis (dir.), *The shape of data in Digital Humanities. Modeling texts and text-based resources*, Londres, Routledge, 2019.

cohérent. Le projet OntoMe les recense et aide à leur usage.

Postulat
Proposition que l'on demande d'admettre comme principe d'une démonstration, bien qu'elle ne soit ni évidente ni démontrée

XML
(Extensible Markup Language)
Langage de marquage d'un texte : les données et les métadonnées sont enserrées dans des balises, elles-mêmes emboîtées. Le langage XML est notamment employé pour l'édition numérique de textes.

Paradigme
Conception théorique dominante dans une communauté.

Données personnelles
Toute information se rapportant à une personne physique identifiée ou identifiable.

Science ouverte
Mouvement visant à rendre la science, ses procédés et ses résultats, accessibles au public.

Les principes F.A.I.R et le plan de gestion des données

1. Les notions à savoir mobiliser : quelques bases de mesures quantitatives

Afin d'étudier un phénomène social, on évite de passer par de simples moyennes, qui cachent les diversités des situations. Les économistes classent notamment les données selon des divisions égales de la population étudiée, selon une division plus ou moins fine : en dix (les déciles*), en quatre (quartiles) ou en cent (centiles) parts, selon la finesse d'analyse que l'on veut mener. Or de ces divisions appliquant une variable à un corps, la chose implique de déterminer des intervalles*. Par exemple, quand on répartit des parts de votes, l'appartenance aux intervalles s'exprime ainsi[1] :

$$0 < x < 10 \qquad 10{,}01 < x < 20 \qquad 20{,}01 < x < 30$$

Étudier l'évolution d'un phénomène quantitatif implique d'être en mesure de comparer des chiffres à l'aide d'indicateurs. En économie, on emploie souvent le **taux de variation***. Lorsqu'on dispose des données pour deux moments spécifiques, on peut alors calculer le pourcentage d'augmentation ou de diminution : il correspond à la part en plus ou en moins par rapport à la valeur de départ dans la valeur d'arrivée. On le calcule ainsi :

$$\text{taux de variation} = \frac{\text{valeur d'arrivée} - \text{valeur de départ}}{\text{valeur de départ}} \text{ x } 100$$

Lorsque le taux de variation n'est pas utile, notamment lorsque les données croissent ou diminuent dans des proportions trop importantes, on préfère utiliser un simple coefficient multiplicateur* pour exprimer l'évolution. Il est calculé en divisant la valeur d'arrivée par la valeur de départ : si le chiffre obtenu est inférieur à 1, alors la valeur a diminué et plus le chiffre obtenu est proche de 0, plus la diminution est importante ; au-dessus de 1, plus le chiffre est élevé, plus l'augmentation est importante. On lit alors « Le nombre de *x* a été multiplié par *y* entre AAAA et AAAA ». Dans le cadre d'une diminution, à des fins de clarté, il est possible de diviser la valeur de départ par la valeur d'arrivée : on obtient le nombre de fois que le phénomène a été divisé.

Dans le cadre de jeux de données très importants, il est difficile de tout analyser avec finesse. Afin de ne pas tasser la complexité d'un phénomène, on peut procéder à un **échantillonnage**. Cette pratique consiste à prendre une partie des données pour analyser un phénomène spécifique, que l'on juge ce phénomène représentatif ou exceptionnel. L'échantillon* choisi peut être aléatoire. Il peut aussi découler de l'étude générale du jeu de données : on y trouve des choses étonnantes, à partir desquelles on dresse des hypothèses. Dans ce cas, on peut passer par des tests logiques avec la fonction SI pour filtrer les données selon des séries de critères. Étayer ces hypothèses implique de faire d'autres calculs, souvent plus précis, qui font intervenir d'autres variables* ou impliquent de changer d'échelle. Employer cet échantillon na va pas de soi en revanche : il faut toujours justifier les choix opérés, tenir compte des limites des données étudiées et ne pas hésiter à suivre le fil des données plutôt que de chercher à démontrer un phénomène préconçu. En somme, *a fortiori* en travaillant avec des échantillons, il est nécessaire : 1/ d'adapter ses hypothèses aux nouvelles données prises en compte ; 2/ de se limiter à une analyse de cet échantillon ou, dans le cas d'une montée en généralités, de bien concevoir celle-ci sur le registre de l'interprétation, et non plus du dévoilement de données factuelles.

Intervalle
Ensemble compris entre deux valeurs

Taux de variation
Calcul permettant de mesure l'évolution d'une variable dans le temps. On le lit en %.

Coefficient multiplicateur
Indicateur de comparaison, qui mesure le rapport entre deux variables*.

Échantillon
Ensemble d'entités (le plus souvent des individus) représentatifs d'une situation ou d'une population.

Variable
Phénomène observable auquel on peut attribuer différentes valeurs prises dans un ensemble.

2. Quelques fonctionnalités de base des tableurs

Un tableur est un logiciel permettant de créer des tableaux et d'effectuer des calculs avec les données qui y sont insérées. Originellement, les tableurs ne servent ainsi pas à écrire du texte, mais à structurer des données de manière à les rendre opérables pour une série d'opérations mathématiques. Dans un même fichier, il est possible de structurer très finement des données et de les séparer pour plus de clarté. Ainsi, au sein d'un même classeur*, on peut trouver différentes tables*, dans lesquelles se trouvent des données de nature variée. Les opérations sont réalisées grâce à une série de fonctions*, insérées dans la cellule où l'on veut obtenir un résultat donné. Ces fonctions peuvent s'emboîter dans une même cellule.

Classeur
Ensemble des tables* contenues dans un même fichier de tableur.

Table
Sous-tableau dans un classeur. On trouve les tables dans des onglets, sur le bandeau inférieur de la page.

[1] < et > signifient « strictement inférieur ou supérieur à » ; sous Excel, on emploie aussi >= (supérieur ou égal) et <= (inférieur ou égal).

Les fonctions sont appelées par leur nom, précédé du symbole « = ». Les plages* et les critères dont elles ont besoin pour fonctionner sont contenus entre deux (), après ce titre, comme suit :

=FONCTION(plage ; critère)

Chaque fonction nécessite une syntaxe spécifique, qu'il faut respecter à la lettre. La syntaxe est en partie commune à un grand nombre de fonctions. Quand on veut que le critère soit une valeur spécifique – que cette valeur soit un texte ou un nombre – on définit cette valeur entre "valeur". De même, on appelle une plage d'une autre table que celle de la cellule contenant la fonction en écrivant le nom de la table, suivi d'un point d'exclamation et des colonnes concernées par l'opération. Imaginons avec une fonction devant calculer des données contenues dans la colonne E de la table 1. On écrit :

=FONCTION(Table1!E:E ; "valeur") (sur Excel) ;
=FONCTION(Table1.E:E ; "valeur") (sur Calc)

Fonction
Formule prédéfinie pour effectuer un calcul selon des valeurs entrées en guise d'arguments.

Plage
Espace défini au sein duquel se trouve des données. Cette plage peut être interne (au sein d'un même classeur) ou externe (hors de ce classeur)

Aujourd'hui, les tableurs utilisent des technologies d'intelligence artificielle pour aider l'utilisateur-trice. C'est par exemple le cas lorsqu'on réplique une fonction en l'adaptant à l'ensemble d'une colonne, et ce faisant en analysant un phénomène avec un nouveau critère (ici, par exemple, en regardant les mêmes variables mais pour une autre commune). Les numéros de lignes et de colonnes risquent ainsi de changer lorsque vous répliquer une fonction. Pour figer des informations et éviter ce qui peut aussi devenir un problème, on utilise le symbole $ avant l'information à figer. Ainsi, dans une fonction, la valeur « E$26 » continuera, quoi qu'il advienne, à analyser la ligne 26.

3. Les fonctions que vous aurez à employer et/ou comprendre

La fonction **NB.SI.ENS** permet de calculer le nombre de lignes qui répondent à une série de conditions. Ces conditions peuvent être opérées sur plusieurs tables* d'un même classeur*. On définit une colonne dans laquelle les lignes doivent être soumises aux critères, puis les plages de critères et les critères, un à un et autant que nécessaire, comme suit :

=NB.SI.ENS(plage_données ; plage_critère1 ; critère 1 ; plage_critère 2 ; critère 2)

La fonction **SOMME** ajoute simplement les données numériques d'une plage de données. Ainsi, l'appel de fonction « =SOMME(A1:A18) » additionne les valeurs des cellules contenues dans les lignes 1 à 18 de la colonne A. La fonction **MOYENNE** calcule la moyenne des données numériques d'une plage de données. Elle fonctionne comme la fonction SOMME. On retrouve la même chose avec la fonction MEDIANE, qui calcule la médiane des données numériques d'une plage de données (le nombre en-dessous duquel se situent 50 % des valeurs et au-dessus duquel se situent 50 % des valeurs). Comme beaucoup d'autres, ces fonctions peuvent devenir conditionnelles en ajoutant « .SI » ou « .SI.ENS » au titre de la fonction. On ajoute alors des plages de critères et des critères, comme indiqué supra.

La fonction **RECHERCHEV** permet d'aller chercher la valeur d'une cellule en fonction de conditions et sur une plage : elle ne résout donc pas une opération mathématique, mais reporte automatiquement la valeur d'une cellule dans une autre. Elle permet, avec d'importants jeux de données, de retrouver des valeurs, y compris lorsque les lignes de tables différentes ne sont pas triées dans le même ordre ou ne répondent pas exactement aux mêmes logiques. Elle fonctionne comme suit :

=RECHERCHEV(valeur_recherchée ; table_matrice ; no_index_colonne ; [valeur proche])

Dans cette fonction, la « valeur_recherchée » est une valeur exacte (comme le nom d'une commune ou d'un individu, ou le contenu d'une cellule). La « table_matrice » consiste en une plage de plusieurs colonnes (dont vous définissez la limite), dont la valeur_recherchée doit être contenue dans la première d'entre elles. Ainsi, lorsque la valeur_recherchée est un nom de commune, la première colonne de la plage doit être la colonne désignant le nom de chaque commune. Le « no_index_colonne » est un numéro désignant la colonne dans laquelle la valeur qui doit sortir de l'opération totale est contenue (par exemple, la colonne avec la part de vote pour telle ou telle formation politique). Ce numéro est déterminé par sa distance (en nombre de colonnes) à la première colonne de la table_matrice. Si la donnée souhaitée est contenue dans la colonne immédiatement à droite de la première colonne, alors ce chiffre est 2, et ainsi de suite.

4. Exercices pratiques

Téléchargez le classeur disponible sur Moodle pour l'exercice sur le vote frontiste dans le Nord-Pas-de-Calais et complétez les cellules teintées en bleu dans les tableaux contenus dans la table « Résultats ». Vous vous appuierez sur les fonctions proposées précédemment. Une fois ces tableaux complétés et maîtrisés, vous pourrez vous évaluer à l'aide du questionnaire disponible, lui, aussi, sur Moodle.

Des méthodes quantitatives à l'analyse statistique

1. Les défis de l'échantillonnage

Travailler à partir d'un échantillon* est une solution raisonnable quand les fonds d'archives sont trop importants pour être transposés entièrement en séries statistiques. Dans cette optique, on ne traite de façon sérielle qu'une partie des données disponibles : une boîte sur trois, un dossier du personnel sur cinq, une année sur quatre, etc. Le problème est que ces choix peuvent biaiser le corpus réel (par exemple, ne prendre que les dossiers d'individus dont le nom commence par une même lettre peut exclure certaines origines ethniques (Lemercier, Zalc 2008, p. 27). Le mieux est de passer par un tirage parfaitement aléatoire. Dans l'idéal, 1/ on documente tous les choix qui sont opérés, pour les expliciter aux lecteurs ; 2/ on évite les biais les plus évidents, par exemple en n'enregistrant que les personnes répondant aux critères de l'étude. Par exemple, une étude qui ne tiendrait compte que des procès pour sorcellerie intentés à des femmes au XVII^e^ siècle pourrait conclure à un phénomène uniquement féminin, alors que les procès intentés à des hommes existent aussi.

Échantillon
Ensemble d'entités (le plus souvent des individus) représentatifs d'une situation ou d'une population.

On peut aussi chercher à travailler sur un sous-groupe* de données que l'on possède déjà. Cela permet de raffiner l'étude, généralement en établissant des « sous-corpus » et en isolant ainsi un phénomène spécifique que l'on voudrait étudier. Le plus facile est de constituer des sous-groupes autour d'une typologie de situations. Les filtres avancés permettent de produire des échantillons répondant à une série de conditions. En opérant ainsi, vous pouvez créer un échantillon ciblé, dont il faudra bien préciser les contours à votre lectorat. En effet, échantillonner selon des critères signifie opérer des choix et perdre en représentativité statistique.

Pour éviter ce problème et/ou pour créer un échantillon test, les tableurs permettent également d'établir des échantillons aléatoires avec la fonction ALEA. La fonction ALEA s'écrit simplement =ALEA() et renvoie un nombre parfaitement aléatoire. En attribuant à chaque ligne un tel nombre aléatoire, puis en récupérant les x premières lignes par ordre croissant, vous créez un échantillon aléatoire de votre corpus. Ici, la chose peut vous permettre de travailler sur un nombre limité de communes, ou bien sur une population plus limitée de résultat d'une enquête (avec l'enquête VICO, l'échantillon aléatoire fourni par les auteurs représente 1/8 des réponses complètes de l'enquête et ne permet d'accéder qu'à une partie seulement des questions posées), plus ou moins localisées précisément selon vos besoins (par exemple, 10 communes aléatoires d'un département qu'on voudrait étudier, ou bien la même chose, à l'échelle d'un département et d'une région...). Évidemment, échantillonner sur 36 000 communes a plus de sens qu'échantillonner sur une centaine de départements ! La validité des échantillons dépend aussi de la finesse d'analyse et du nombre de variables que l'on veut inclure dans l'analyse.

Y compris avec un échantillon aléatoire, on n'échappe pas à l'étude de la représentativité des chiffres, qu'il est nécessaire de donner à voir au lecteur : il s'agit de faire œuvre de transparence, mais aussi d'être conscient-e des marges d'erreur possibles d'une étude. Le calcul rudimentaire de l'écart entre deux valeurs est la première des conventions. Les pourcentages de répartition entre deux sous-groupes doivent par exemple différer d'au moins 12 à 14 points pour que la différence soit réputée significative sur un groupe d'une centaine de personnes ; plus l'échantillon est large, plus cet écart peut être faible (Lemercier, Zalc 2008, p. 29). Il existe aussi des calculs plus sérieux, comme la variance* : on calcule alors la probabilité que les variables étudiées ne soient pas réellement connectées entre elles et que les chiffres soient en fait dus au hasard ou à un biais de sources.

Variance
Mesure de la dispersion des valeurs d'un échantillon ou d'une variable aléatoire

2. De nouvelles fonctionnalités des tableurs

Sous les tableurs, il est possible d'**emboiter des fonctions et des opérations** les unes dans les autres. La chose demande un peu de gymnastique d'esprit (et, souvent, de passer par le papier en amont pour synthétiser les calculs que l'on veut faire) mais permet de ne pas multiplier les colonnes préparatoires ou intermédiaires. L'emboîtement se gère comme dans une formule mathématique standard, avec des parenthèses qui ferment les opérations et les placent les unes dans les autres. Sur le principe, il n'y a pas vraiment de limite à cet emboîtement, en dehors du fait que vous risquez de multiplier les erreurs. Par exemple, ici, on emboîte des résultats de fonctions SI et NB.SI.ENS :

=SI(NB.SI.ENS(plage ; plage_critère-1 ; critère-1)>10 ; « OUI », « NON »)

la fonction renvoie « OUI » si le nombre de lignes répondant au critère 1 est supérieur à 10 ;
au contraire, elle renvoie non si ce nombre est inférieur à 10.

Pour rappel (cf. leçon précédente), les tableurs permettent aussi de gagner du temps en rendant **séquentielles** des opérations, c'est-à-dire que le logiciel transpose le calcul à d'autres cellules : on n'écrit la formule qu'une seule fois dans le cas de calculs répétitifs et prévisibles, comme des totaux en fin de tableau par exemple. Pour rendre séquentielle un très grand nombre de cellules sur une colonne, on clique sur le coin inférieur droit de la cellule avec la formule à déplacer. Le logiciel transpose la formule jusqu'à ce que les lignes soient vides ou qu'une cellule non vide ne l'interrompe. Attention toutefois, le tableur peut changer plusieurs noms de cellules à la fois en rendant séquentielle une formule. Pour « fixer » une cellule, une colonne ou une ligne, on utilise le signe $ dans la formule :

« A$1 »	La ligne demeure ; la colonne varie	« $A1 »	La colonne demeure ; la ligne varie	« A1 »	La colonne et la ligne demeurent

3. Mesurer la dispersion des données (niveau avancé)

La **dispersion statistique** désigne le degré « d'éloignement des termes d'une série les uns par rapport aux autres, ou la manière dont ces valeurs se répartissent autour d'une valeur centrale » (Bavoux 2014, p. 190). L'objectif est d'analyser l'homogénéité ou l'hétérogénéité de la population statistique étudiée et de la quantifier précisément. Dans un second temps, mesurer la dispersion d'un groupe et la comparer avec un autre groupe permet également d'interroger la représentativité d'un échantillon ou d'un groupe étudié. On utilise pour cela une série d'indicateurs. Les plus simples sont l'écart à la moyenne (calculé autour d'une valeur fixe, qui est la moyenne de l'ensemble), les coefficients de variation, la variance* et l'écart-type*. Les tableurs permettent de calculer ces indicateurs de façon simple, y compris sur des jeux de données importants comme ceux auxquels vous êtes confronté-es.

En mathématiques, l'**écart** mesure la distance entre deux valeurs, par exemple entre la note la plus basse et la note la plus haute dans un paquet de copies. Le nombre obtenu se lit selon cette nature : en calculant l'écart, on mesure une distance. Dans l'exemple du paquet de copies, l'écart se lira « L'écart (absolu) du paquet est de x points ». Cet écart absolu présente ses limites et, dans les faits, on calcule plutôt deux types d'écarts : l'écart à la moyenne et l'écart-type.

L'**écart à la moyenne** permet de mesurer précisément la sous-représentation ou la surreprésentation d'un phénomène. Sous Excel, la fonction ECART.MOYEN renvoie la moyenne des écarts absolus des observations par rapport à la moyenne arithmétique (que vous n'avez donc pas besoin de calculer, la fonction le fait pour vous).

=ECART.MOYEN(nombre1; [nombre 2];....)

On peut reprendre l'exemple des copies et des notes d'élèves. Sur un semestre, tous les élèves ont eu 5 notes. Si on calcule la moyenne de chaque élève, ce chiffre final efface la dispersion des 5 notes de chacun-e au fil du semestre : avoir 10/20 de moyenne en ayant eu 10 à cinq reprises, ce n'est pas avoir 10 de moyenne en ayant des notes écartées entre 2 et 18/20. Parce que travailler avec des moyennes dissimule ces questions, l'écart à la moyenne pour chaque évaluation rend la visibilité à cette dispersion des valeurs. Ici, on pourrait par exemple calculer cet écart pour chaque évaluation, puis pour la moyenne finale des élèves, afin de déterminer quelles évaluations ont donné lieu à la plus grande hétérogénéité des résultats.

L'**écart-type** calcule la dispersion des valeurs d'une série autour de leur moyenne arithmétique. Il est obtenu en calculant la racine carrée des écarts à la moyenne. Lorsque vous calculez l'écart-type d'un échantillon, vous mesurez cette donnée par rapport à l'ensemble de la population. Autrement dit, vous pouvez calculer l'écart-type d'un groupe de communes par rapport à la moyenne du département, de la région ou au niveau national. Ce calcul vous permet de mesurer ce que représente votre échantillon par rapport à des ensembles plus importants (le but peut-être par exemple d'éviter une trop forte dispersion des données ou, à l'inverse, de la rechercher pour retrouver de la représentativité statistique). Il s'agit ainsi de présenter les limites d'une démonstration basée sur un petit groupe d'individus, ou bien de justifier d'un protocole en montrant la forte représentativité de votre échantillon..

Sur Excel, l'écart-type se gère différemment selon qu'on calcule l'écart-type sur des échantillons de population (on utilise ECARTTYPE.STANDARD, ou ECARTYPEP sous Calc) ou sur une population entière (ECARTTYPE.PEARSON, ou ECARTYPE sous Calc). La formule est la suivante :

=ECARTYPE.STANDARD/PEARSON(nombre1; [nombre 2];....)

Dans cette formule, le « nombre1 » correspond aux valeurs (qui peuvent être une plage) dont vous voulez calculer l'écart à la moyenne : ce « nombre1 » est une plage des données de votre échantillon. Le « nombre2 » correspond aux valeurs de l'échantillon plus large dans lequel on veut mesurer l'écart. Le chiffre calculé s'exprime dans l'unité de valeur des données mises en écart (ici, probablement en nombre de suffrages). Il est forcément positif.

Introduction aux bases de données relationnelles

Vous commencez à comprendre l'importance de la structuration des données de la recherche dans la construction des résultats scientifiques et les différentes possibilités d'usages de ces données. Passons désormais à la vitesse supérieure : au-delà de jeux de données sous la forme d'une table unique, il est possible de gérer une multitude de tables simultanément et d'établir des liens entre elles. En histoire, cette possibilité est particulièrement intéressante, parce qu'on n'étudie pas un monde qu'il est aisé de figer dans une liste, elle-même couchée dans un tableau Excel. Ce monde est en partie celui des bases de données relationnelles : ce cours vous permet d'en comprendre les fondamentaux et de vous lancer pendant les séances suivantes dans la modélisation de votre première base, à partir de transcriptions des registres d'écrous de la prévôté de l'hôtel de la cour de France.

1. Ce qu'est (et permet) une base de données relationnelle

Une base de données (DB) (relationnelle ou non) est un jeu de données strictement organisé et dont la structure permet de maximiser le nombre de questions et d'angles possibles d'étude (Quamen, Bath 2016: 181). Sur le principe, une table lue par un tableur est une base de données. Toutefois, il n'est pas possible – ou, disons, difficile – d'étudier avec un tableur des données de natures différentes, parce que chaque ligne doit correspondre à un même type d'entité. Il faut, à chaque changement de focale, une nouvelle table. C'est là qu'intervient le format de la base de données relationnelle (RDB, pour *relational database*). Une RDB est un type de DB qui met en relation différentes données saisies, et non plus seulement de quantifier. Les RDB fonctionnent, comme d'autres bases de données, à partir de tables. Dans ces tables, les entités sont définies par un identifiant unique, que l'on peut retrouver dans d'autres tables, associés à d'autres entités. Chaque table est toujours une liste, dans laquelle l'entrée n'est pas systématiquement la même (la ligne pourrait désormais correspondre à un parti politique ou à une élection plutôt qu'à une commune, par exemple). Les tables sont reliées les unes aux autres quand des données se recoupent, par des identifiants (ID) uniques. De ce fait, les RDB permettent d'étudier des réseaux, à différentes échelles et sur le temps long. Elles permettent également de ne pas trop ramasser la complexité des sources, ainsi que d'éviter d'analyser les données par une entrée unique (qui seraient les lignes d'un tableau). En revanche, elles sont beaucoup plus complexes à mettre en place, et sur le plan technique, et sur le plan théorique.

La RDB n'est théoriquement qu'une interface qui permet d'analyser des données. On peut envisager d'intégrer dans cette interface des données tabulées, saisies par le/la chercheur-euse ou bien importées d'autres projets de recherche. On commence par l'étape de la saisie (qui peut se faire sur tableur, ou via un système de formulaire(s), que vous pouvez moduler à votre guise, mais qui crée lui-même des données tabulées. Cette saisie précède – de plus ou moins loin – l'encodage des données, c'est-à-dire l'éclatement de la source primaire en données structurées (et plus simplement une transcription). Une fois ces deux étapes réalisées, on peut passer à l'analyse. Celle-ci utilise des systèmes de requêtage : on *demande* au logiciel de faire des calculs spécifiques, de rechercher des liens entre des données, etc. Les interfaces sont plus ou moins élaborées et ergonomiques selon les cas. LibreOffice Base permet ainsi d'interroger les données simplement ; en revanche, en face de gros jeux de données ou de calculs élaborés, il ne suffit plus. Les systèmes utilisant le langage d'encodage SQL (*Structured Query Language*) semblent plus rudimentaires et impliquent d'avoir des compétences de codage ; ils sont toutefois beaucoup plus puissants. Le plus utilisé à l'échelle mondiale est PostgreSQL.

2. Modéliser une base de données relationnelle

Le premier enjeu lorsqu'on réfléchit à constituer une BDR, c'est de réfléchir à la modélisation des données qu'on veut lui donner. C'est une étape essentielle, chronophage et parfois difficile : le degré d'abstraction nécessaire est important, parce qu'il faut réfléchir très concrètement à ce que sont les données et à ce qu'elles représentent (des entités, des relations, etc.). La modélisation des données est une « collection d'outils conceptuels pour décrire des données, des relations entre elles, leur sémantique et leurs limites » (Flanders, Jannidis 2016: 230). En clair, il s'agit de la manière dont on représente la structure d'un jeu de données. Pour ce faire, il faut prendre de la hauteur sur les sources que vous voulez transformer en données : les personnes deviennent des entités, au même titre que les institutions ou les lieux. Chaque donnée va être intégrée dans le modèle, au sein d'un *type* : chaque personne est encodée pour être reconnue comme telle ; dans la base du SUDOC, chaque livre est reconnu comme tel, puis relié à un auteur (2e type), à un éditeur (3e type), etc. Au niveau *conceptuel*, on commence à réfléchir à sa base de données en faisant abstraction des contraintes techniques et en ne faisant que décrire le réel : vous avez des sources, qui décrivent des situations, mettent en scène des personnes, qui évoluent dans des lieux... On passe ensuite au niveau *logique*, en faisant coïncider la description du réel avec des contraintes techniques. Lors de cette étape, on définit le nombre de tables et les variables qui vont s'y trouver. Dans le cas d'archives judiciaires, les personnes impliquées vont aller dans une table avec les différentes variables affiliées (prénom, nom, âge, sexe, etc.), les crimes et délits dans une

autre, les procès dans une troisième. Enfin, on implémente ces données structurées dans un niveau *physique*, c'est-à-dire qu'on insère les fichiers dans le logiciel envisagé (le plus souvent dès le début) et qu'on peut commencer l'analyse des données. Lors de cette étape, on réfléchit aussi à l'optimisation des performances de la base. En théorie, il est alors possible de changer de solution et d'importer les données dans un autre logiciel, ce pourquoi le niveau *physique* est parfois conçu en dehors de la phase de modélisation des données à proprement parler.

S'il est possible de concevoir sa propre base de données dans son coin, y compris avec des logiciels libres (LibreOffice Base par exemple), il est aujourd'hui conseillé de veiller à l'interopérabilité et à la réutilisabilité des données (selon les principes FAIR* de la Science ouverte). L'idéal pensé par certain serait de relier toutes les données de la recherche dans un système unique, ouvert (*open-linked data*). Pour ce faire, le mieux est de suivre des ontologies* les plus utilisées. Une ontologie* est un modèle pré-conçu, généralement pour une discipline donnée, dans lequel il s'agit de faire rentrer les données. La plupart des ontologies réfléchissent par classes, types, entités et relations. Dans le cas d'un crime commis, la victime est une entité, qui a avec l'auteur des faits une ou plusieurs relations (le crime les relie, mais le mariage aussi dans le cas d'une violence conjugale par exemple). L'institution qui enregistre et gère le crime est elle aussi une entité, mais qui dépend d'un autre type (non plus une personne physique, mais une personne morale). Ces entités sont considérées comme des entités persistantes (classe), parce qu'elles existent sur la durée. Le crime est quant à lui enregistré en tant qu'événement (classe), avec des items dédiés (début, fin, lieu, nature, participants, etc.). L'enjeu de la modélisation est donc de parvenir à ce niveau d'abstraction du réel, pour faire coïncider chaque source et chaque donnée avec un modèle générique, qui pourra être interopérable avec d'autres bases. Dans ce modèle, on attribue aux entités des clés, primaires et secondaires. Ces clés sont les identifiants uniques pour chaque entité, qu'on désigne par le terme *integer* (c'est-à-dire, en anglais, un nombre entier, et par extension, un ID). Les clés primaires (*primary keys*) désignent une ligne dans une table spécifique (un ID de personne avec les variables de son identité, par exemple). Les clés secondaires, ou externes (*foreign keys*) sont des clés primaires ajoutées comme variables dans d'autres tables. Ainsi, toutes les clés sont primaires pour une table, mais elles peuvent se retrouver, comme clé secondaire, au sein d'une autre table. Ce sont ces clés secondaires qui font les liens entre les données et permettent à la machine de les comprendre comme un tout cohérent.

Les principes F.A.I.R

Ontologie
Modèle de données permettant de créer un ensemble cohérent.

Open-linked data
Données insérées dans des bases de données relationnelles et ouvertes à tous les utilisateurs d'un même système.

3. Effectuer des requêtes et des analyses avec une base de données relationnelle : le SQL

Le SQL est un langage informatique qui permet de communiquer avec ces bases de données pour extraire, ajouter, modifier ou supprimer des informations. Il permet de rechercher des informations spécifiques dans une vaste base de données, comme des événements historiques, des personnages ou des dates précises, sans avoir à parcourir manuellement chaque enregistrement. Cela facilite grandement l'analyse et l'interprétation des données. En utilisant des requêtes SQL, les chercheurs peuvent croiser des informations issues de différentes tables, par exemple, pour étudier les relations entre divers événements ou acteurs historiques. Cela offre une approche plus systématique et efficace pour explorer des questions de recherche complexes. Concrètement, on demande à la machine de faire des recherches et d'effectuer des opérations, sous la forme de macros*. Ce sont ces opérations qui permettent d'aller au-delà de la seule extraction d'informations, en sortant des analyses de réseaux, des graphiques et des cartes. Le coût d'entrée pour cela est très lourd : il faut des mois de formation et un entourage solide d'ingénieurs de recherche. Si le jeu peut en valoir la chandelle, il faut, avant de se lancer dans l'aventure, bien réfléchir à vos besoins, tant en termes de formation (et de valorisation dans votre parcours, en tant que compétences) qu'en termes de recherche (est-ce qu'une architecture complexe est nécessaire, ou bien peut-on bricoler quelque chose avec des tables individuelles ?).

Macro
Programme informatique, le plus souvent rudimentaire, destiné à exécuter des tâches prédéfinies et à les répéter autant de fois que nécessaire.

Bibliographie sommaire

BERETTA, Francesco, « Données ouvertes liées et recherche historique : un changement de paradigme », *Humanités numériques*, 7-1, 2023 ; BERKOWITZ, Héloïse, DELACOUR, Hélène, « Ouvrir les données de la recherche : Quelles implications pour les sciences sociales ? », *M@n@gement*, 25-4, 2022, p. 1-31 ; DEDIEU, Jean-Pierre, « Construire des bases de données. Introduction sous forme d'un retour d'expérience », *Histoire & Mesure,* 2024/2, 34, p. 7-26 ; SCHWANDT, Silke (dir.), *Digital Methods in the Humanities. Challenges, Ideas, Perspectives*, Bielefed, Bielefeld University Press, 2021 ; Schreibman, Susan *et al.* (dir.), *A New Companion to Digital Humanities*, Malden, Oxford, Blackwell, 2016. SCHUSTER, Kirsten, DUNN, Stuart (dir.), *Routledge International Handbook of Tesearch Methods in Digital Humanities*, London, Routlege, 2021.