

# Outillage numérique et statistique de l'historien

Philippe Rygiel

Master d'histoire à distance, Université Paris Ouest-Nanterre,  
2014



# **Table des matières**



# Chapitre 1

## Avant-propos

### 1.1 Organisation du cours

Ce cours est conçu comme une initiation, spécifiquement destinée à des historiens, à la culture numérique, qui est pour partie une culture mathématique et statistique. Son objectif n'est pas de transformer d'honnêtes historiens en programmeurs ou en mathématiciens, mais prenant acte des transformations intervenues au cours des dernières décennies, de permettre à ceux-ci de se forger une représentation fonctionnelle de ce qu'est aujourd'hui l'instrumentation à la disposition d'un historien au travail et de faciliter par là à la fois le repérage et l'appropriation de certains outils, mais aussi, surtout peut-être de stimuler le développement d'une agilité, d'une curiosité, permise seulement par le moyen de la maîtrise de quelques notions clés, autorisant ensuite chacun à construire son propre outillage. Il s'agit d'aider donc à la constitution d'un trousseau, ou d'une boîte à outils, par un détour réflexif.

Nos exemples seront souvent empruntés à l'histoire moderne ou à l'histoire contemporaine, simple reflet de la meilleure connaissance de ces périodes et de leur historiographie. La période d'étude choisie n'est pas sans incidences sur les pratiques numériques des uns et des autres. Les corpus de sources, particulièrement dans le cas des contemporanéistes ont des contours parfois très flous, peuvent être très hétérogènes et comporter une bonne part de sources riches, sans même parler du fait que les flux vidéos, la radio, voire les dispositifs numériques eux-mêmes, sont alors susceptibles de faire source. L'éventail des pratiques de recherche est alors très vaste, de même que sont divers les besoins d'instrumentation. Nous espérons cependant, partant des tâches de l'historien, qui sont pour beaucoup commune, que tous trouveront ici des pistes et des idées. Nous avons alors choisi, c'est pour partie

la conséquence d'un cours s'adressant à tous quel que soit le sujet traité et le matériau utilisé, non pas d'initier au maniement d'une ou de quelques technologies, mais de présenter, puisqu'intervenant en début de corpus, un panorama, incomplet bien sûr, des ressources numériques et statistiques aujourd'hui accessibles à l'historien, en tenant de préciser pour chacune d'elle ses conditions et ses contraintes d'usage, ses fonctionnalités, et en renvoyant, pour de plus amples informations et une prise en main à une littérature seconde aujourd'hui abondante. C'est à un détour par l'abstraction donc que ce cours convie, non pas à un apprentissage procédural, même si sont rappelées quelques notions et techniques fondamentales. Il nous a paru, au regard de la diversité des pratiques contemporaines mais aussi de la très forte instabilité des procédures que tel était le choix le plus raisonnable.

Ce cours d'autre part s'adresse à un public très divers tant par les sujets et les méthodes d'études choisies par les uns et les autres que par la culture informatique et statistique de départ de ses auditeurs. Le risque est réel de proposer des éléments qui sont inutiles et inaccessibles aux uns, triviaux pour d'autres. Le choix a donc été fait de proposer à tous des éléments permettant de nourrir nous l'espérons une réflexion sur vos propres pratiques tout en rappelant quelques notions de base, en offrant quelques conseils simples aussi, qui nous semblent indispensables à tous, les techniques plus avancées sont généralement simplement évoquées, à l'attention des plus curieux ou des plus experts d'entre vous, accompagnées de renvois bibliographiques permettant d'aller plus loin.

#### *Nota bene*

Dans un certain nombre de cas les références des ressources utiles, en particulier celles disponibles en ligne ne sont pas précisées. La volatilité des url fait qu'il est souvent plus efficace de donner les mots-clés nécessaires à la recherche d'un document à laquelle on procédera par l'utilisation d'un moteur de recherches. Lorsque nous donnons une url, celle-ci a été vérifiée au cours de l'été 2014.

**renseignements pratiques** Ce cours sera accompagné d'un forum, qui devrait être ouvert dès la mi-septembre et proposera quelques documents d'accompagnement et exercices.

**Evaluation** L'évaluation consistera en la production d'un devoir écrit, dont les modalités seront précisées sur le forum attaché à cet enseignement fin septembre.

## 1.2 Historiens face au numérique

J'ai jusqu'ici affirmé que les pratiques des historiens avaient changé, et continuaient à changer, du fait des transformations de leur environnement et de l'instrumentation à leur disposition et qu'il était important d'intégrer cette dimension dans sa réflexion, mais aussi d'en tenir compte lorsque l'on définit ses stratégies de recherches. Nous allons dans un premier temps tenter de préciser ce constat, en nous appuyant sur les réflexions de quelques historiens de profession. Depuis quelques décennies, un certain nombre d'historiens en effet s'interrogent sur les transformations des pratiques érudites liées à l'apparition puis à la diffusion de l'informatique, puis des technologies réseaux. Longtemps cantonnés à de petits cercles de spécialistes, ces débats se retrouvent aujourd'hui souvent dans les pages des revues d'histoire, ou alimentent blogs et carnets électroniques, celui par exemple de Frédéric Clavert. Je vous propose ici de lire un extrait d'un billet électronique daté de 2012.

« (...) toute une frange de l'histoire numérique s'intéresse aux périodes les plus récentes de notre histoire. Sans parler d'histoire immédiate, de grands chantiers historiographiques s'ouvrent en même temps que les centres d'archives mettent des sources primaires à disposition au fil de la règle des trente ans.

Or le web, le courrier électronique et de nombreux autres supports numériques s'imposent progressivement au sein des administrations, des entreprises, des associations, des particuliers... au fil des années 1990. À partir de 2020, et même si les archivistes élaborent des stratégies de conservations (c'est-à-dire qu'ils ne gardent pas tout), nous aurons de plus en plus à traiter, à côté de nos bonnes vieilles liasses de papier, des documents numériques.

Or nous ne sommes pas armés, pour le moment, pour y faire face. Rares sont ceux qui commencent à élaborer une méthodologique critique des documents numériques. Encore plus rares sont ceux qui pensent à la manière dont nous interrogerons des archives numériques massives. Nous pouvons nous préparer, en adoptant des technologies émergentes, décrites ça et là. Mais cela suppose un fort effort de formation des étudiants, ce qui nous fait retomber sur une autre discussion.

Il y a déjà eu des historiographies, qui, pendant quelques années, se taisent, temporairement étouffées par un afflux important d'archives nouvelles. Ainsi, de la biographie d'Hitler par J. Fest en 1974 à celle de Marlis Steinert en 1991, de nombreux ouvrages sur Hitler ont été publiés. Entre Steinert et Kershaw huit ans plus tard, il y a, par contre, un vide certain. Pourquoi ? En raison de la chute du Mur de Berlin, qui a permis aux historiens d'aller dans les archives des pays de l'Est de l'Europe. Mais il leur a fallu plusieurs années pour les assimiler.

Si nous ne sommes pas collectivement prêts à utiliser des sources primaires numériques d'ici une dizaine d'années, le risque est que l'histoire contemporaine, sur certains sujets, reste muette pendant une longue période, car coupée de sa matière première. Pour ces domaines de l'histoire contemporaine, les Digital Humanities et leurs méthodes me semblent être une question de survie.»

Frédéric Clavert, Les Digital Humanities, une question de survie pour l'histoire contemporaine ?, *Humanités numériques, Pensées éparses* le 24/07/2012 .

Je vous laisse le loisir de commenter ce texte. Retenons une idée simple, mais forte. L'historien vit aujourd'hui, et de fait déjà depuis plusieurs décennies, dans un environnement numérique. La première conséquence en est qu'une bonne partie des traces laissées par les sociétés contemporaines sont des traces numériques. Il sera bien difficile de mener une histoire culturelle ou politique des décennies qui viennent de s'écouler sans avoir accès au web par exemple. Or les matériaux qui y sont déposés sont d'une nature particulière, qui déterminent à la fois les conditions de leur conservation, mais aussi de leur consultation et de leur utilisation. Il est de fait difficile de les manipuler efficacement sans un minimum de compétences informatiques et de connaissances mathématiques.

Ce constat, qui fait obligation aux historiens du très contemporain de se confronter aux mondes numériques, n'en dispense pas pour autant les autres. Je vous propose de lire ici ce qu'a écrit Serge Noiret de l'importance des transformations en cours pour les pratiques, mais aussi les modes de pensée de tous les historiens.

(...)l'instabilité des textes transposés au numérique est aujourd'hui une donnée permanente avec laquelle l'historien « digital » doit se confronter. Cette mutation vers des textes fluides, soumis à des changements continus, a forcé à s'interroger (...) sur de nouvelles notions descriptives des documents et de nouvelles pratiques de leur conservation et d'accès constant dans le temps, aux nouveaux documents numériques.

L'historien assiste ainsi –souvent de manière passive– à la construction de nouveaux instruments (logiciels, bases de données) et de nouvelles pratiques (communication, lecture, publication) qui, de fait, lient son travail quotidien à des pratiques d'informatique humaniste (...). Une nouvelle dépendance vis à vis de connaissances documentaires qui se trouvent dans des lieux virtuels et nécessitent de “machines” et de programmes pour être visualisées (une connaissance que les bibliothécaires et les archivistes tentent de maintenir sur le long terme), a suscité de nouvelles pratiques dues au numérique qui ne faisaient pas traditionnellement partie du bagage de l'humaniste. (...)

En parallèle à cette transformation de ses méthodes documentaires et critiques, l'historien assiste à une déstabilisation de l'autorité –souvent de l'autorité académique comme unique détentrice de la connaissance vraie et scientifique face à l'apparition de discours d'histoires qui proviennent de tous les secteurs de la société(...)

L'auteur isolé d'un essai historiographique disparaît parfois au profit du collectif et sans attribution respective de ce qui a été écrit. Les sources primaires ne sont souvent plus reliées au contexte matériel qui leur faisait « prendre un sens » et les validait dans leurs contextes : dans le monde numérique un des grands problèmes est certainement celui de l'individuation des contextes signifiants, ce que les philologues appellent l'histoire de la construction des textes et des documents.

Il faut donc prendre acte de la nécessité de recomposer l'appareil critique et les méthodes scientifiques historiennes en fonction du web, le médium qui cannibalise tous les autres (...) nous possédons de nouvelles bibliothèques, de nouvelles sources, de nouvelles formes d'enseignement et d'apprentissage, de nouvelles formes d'écritures de l'histoire et surtout, de nouvelles formes de communication de l'histoire et, en historie contemporaine, de nouvelles formes de représentation identitaire et de reconstruction mémorielle souvent antagonistes des reconstructions des historiens.

Serge Noiret, «Y'a-t-il une histoire numérique 2.0? », in Jean-Philippe GENET and Andrea ZORZI (eds), *Les historiens et l'informatique : Un métier à réinventer. Etudes réunies par Jean-Philippe Genet et Andrea Zorzi*, Rome, Ecole Française de Rome, 2011, 235-288, Collection de l'Ecole Française de Rome, 444

Je vous laisse le soin là aussi de retrouver ce texte et d'en cerner précisément les enjeux. Retenons en l'affirmation de ce que les transformations de l'univers numériques affectent non seulement les pratiques des historiens, a minima par l'usage de ressources électroniques, à peu près aujourd'hui inévitable, mais aussi, écrit Noiret les conditions sociales de l'écriture ou de la diffusion des discours historiques et certaines des notions fondamentales liées à ce domaine de savoir, et essayons maintenant de comprendre ce qui justifie de telles affirmations. Il nous faut pour cela emprunter quelques notions et quelques exemples à d'autres domaines que l'histoire.

### 1.3 L'historien vu depuis les mondes numériques

Si nous examinons ainsi l'historien, pour ainsi dire de l'extérieur [?], nous voyons des acteurs, se nommant les uns les autres historiens, qui produisent une multitude structurée de signes dont la très grande majorité n'est portée à la connaissance de personne et ces signes entretiennent des rapports complexes avec plusieurs réservoirs de signes et de symboles. Un schéma, éclairera je l'espère ce point.

L'historien contemporain apparaît alors d'abord comme un polygraphe hypertextuel dissimulant aux regards l'essentiel des inscriptions qu'il produit. Consultant les documents conservés par les centres d'archives, les bibliothèques, les institutions muséales, ou bien ceux qu'il a lui même rassemblés, parce qu'il peut être lui même créateur d'archives, il accompagne, ou est censé le faire, ses inscriptions de l'indication du chemin d'accès au document consulté, qui peut être un vestige du passé qu'il cherche à saisir ou bien un commentaire ultérieur – extrait d'ouvrage et d'article, notice de catalogue, voire notes d'un érudit – lui-même généralement renvoyant directement ou de façon médiate à un vestige authentifié, par l'historien ou par d'autres, d'une époque passée que l'historien tend à rapporter au référent de son discours. Si nous voulons englober la plus grande diversité possible de pratiques, il nous faut considérer que les annotations produites par l'historien ne prennent pas forcément la forme de textes ou de fragments de texte en langage naturel. Elles peuvent aussi consister en graphiques, enregistrements vocaux, dessins, chaînes de caractères parfois codées insérées dans des bases de données informatisées, voire en clichés ou films. Il est probable cependant, même si nous savons très peu de choses de la façon dont les historiens de fait travaillent , que les données non textuelles qu'ils accumulent soient décrites et indexées au moyen de dispositifs textuels (séquences linéaires, listes, tableaux, arbres)

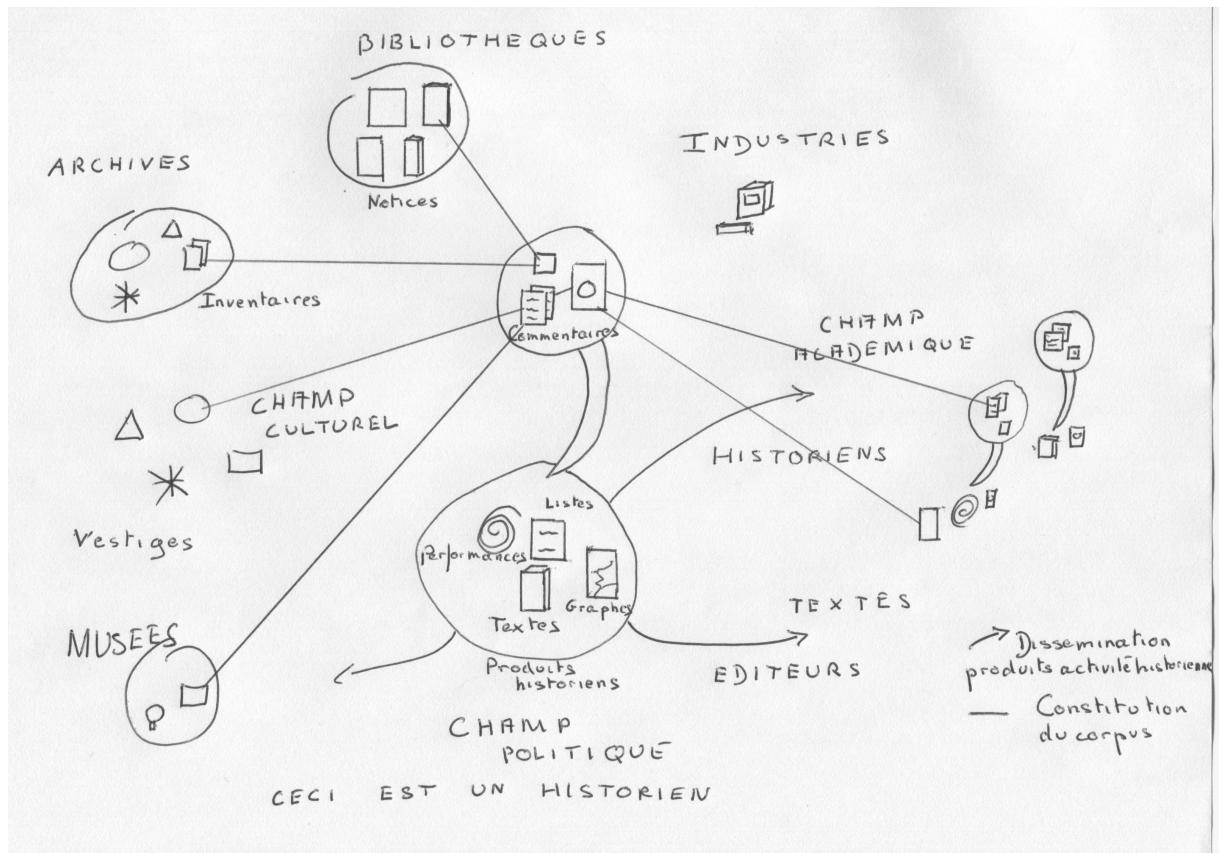


FIGURE 1.1 –

plus ou moins complexes. L'historien de ce fait se trouve à la tête d'une masse d'annotations dont beaucoup sont des données textuelles ou quasi textuelles qu'il va structurer et manipuler, produisant listes et tableaux, au moyen de requêtes plus ou moins complexes, mais aussi des textes nouveaux par concaténation ou extraction, voire calcul, et possiblement des dispositifs graphiques (cartes, plans, graphes, schémas), qui sont, dans le monde numérique, assimilables eux-aussi aux produits de l'application d'opérateurs d'écriture à des éléments textuels . Ce n'est qu'ensuite qu'il va rendre compte de son activité, par un cours, une conférence, un texte, une prise de parole dans le cadre d'un débat par exemple. Il rencontre alors d'autres types d'acteurs, éditeurs, journalistes étudiants, pour n'en citer que quelques-uns, dans le même mouvement qu'il rend publics des documents qui pourront à leur tour faire l'objet d'une appropriation et nourrir le travail d'autres historiens. L'enquête historique, dans cette perspective, devient la délimitation d'un corpus, par inclusion dans l'ensemble des matériaux de l'enquête d'inscriptions, référée toujours à une matérialité dont l'existence est antérieure à son mouvement et n'est pas affectée par celui-ci et qui peut être ou non élément d'une collection, puis application à celui-ci d'un jeu de règles opératoires - donc production de nouveaux éléments textuels.

Cette définition n'a pas vocation à définir un être ou une essence de l'histoire telle qu'elle est pratiquée aujourd'hui. Construction, abstraction, elle laisse délibérément de côté un certain nombre de propriétés de l'acte historien et plusieurs des questions qui lui sont associées, en particulier celle du sens, de la mise en récit, mais aussi de la définition même de ce qu'est le fait historique, que nous avons commencé par écarter et cela parce que sa visée est d'abord stratégique. Il s'agit de réfléchir à ce que change, ou peut changer, pour une profession particulière, les transformations contemporaines de l'outillage des professions intellectuelles, et partant à se doter des outils permettant de définir des conduites adaptées.

L'intérêt de ce détournement est triple. Il permet de rappeler d'abord, après bien d'autres, que la connaissance historique est le produit d'une chaîne de production de savoirs, prise dans des institutions, des systèmes techniques et des champs sociaux, et non accumulation d'œuvres singulières émanant de solitudes géniales. Cela nous rappelle aussi qu'il y a entre le référent auquel se rapporte le discours de l'historien et le vestige d'une activité humaine qu'il convoque à l'appui de son discours, puis entre l'incorporation de ces vestiges dans un corpus et les descriptions multiples dont il font l'objet, des médiations successives, dont nous avons parfois bien à tort postulé la transparence et qui font de la production de l'histoire une activité éminemment sociale et collective . De plus, nous pouvons alors penser le fait qu'une transformation des conditions de production de la connaissance historique est susceptible d'avoir

de profonds effets tant sur le mode de son élaboration que sur la nature des productions émergeant dans la sphère publique. En l'occurrence, l'augmentation de la masse des données accessibles, l'existence sous des modes numériques divers d'une partie croissante d'entre elles, la diversification des modes de traitement de l'information , mais aussi la possibilité de l'automatisation de certains, qui est la marque propre de l'informatique et les transformations des modes de diffusion de la connaissance constituent tant une mutation des technologies de l'intellect que des conditions de réception du savoir. Elles laissent augurer d'une mutation des produits de l'activité historienne, tant par une transformation de la demande, susceptible de se segmenter et donc de conduire à une différenciation des produits et des profils historiens, que par la possibilité offerte de manipulations nouvelles de corpus d'un ordre de grandeur nouveau. Ajoutons que le schéma proposé permet de penser « topologiquement ». S'il y a des lieux dans le monde numérique d'accumulation et de transformation de l'information, alors s'ouvre la possibilité, en même temps que d'une distribution potentiellement universelle, d'une appropriation privée des annotations et des modes opératoires – qui étant logiciels sont eux-mêmes textes – ou à l'inverse d'une dissémination modifiant tant la nature que le volume des données mobilisables.

Prenons quelque exemples pour mieux comprendre les implications pratiques de ces changements en cours. Prenons deux exemples, illustrant chacune de ces modalités et d'abord l'existence de plates-formes numériques payantes donnant accès au contenu des numéros récents des revues scientifiques. L'accès à celles-ci est affranchi des contraintes géographiques. Ces ressources peuvent être consultées de n'importe où dans le monde. La possibilité de le faire dépend cependant de l'appartenance institutionnelle des individus, de leurs capacités financières, ainsi que des formes des réseaux sociaux desquels ils participent, du degré aussi de littératie numérique auquel ils sont parvenus. La consultation d'un article déposé sur la plate-forme Cairn coûte ainsi, aux dires des intéressés, cinq dollars lorsque l'on se connecte depuis le campus de l'université de Sydney. Au moment où l'inscription des productions textuelles des chercheurs au sein de la bibliographie internationale – l'inscription à bon escient de références à celle-ci dans leurs textes autant que leur capacité à obtenir que leurs propres travaux soient mentionnées par les producteurs de textes inscrits dans les corpus de référence des agences de notation – est de plus en affirmée comme critère d'évaluation des productions scientifiques, le passage des revues les plus prestigieuses dans le monde numérique payant offre aux plus dotés la possibilité d'accroître leur productivité et de multiplier les signes de distinction.

Il pourra alors être question d'accès, de passages, donc de territoires, de pouvoirs et de conflits, et possible d'interpréter aussi parfois les postures et

les conduites des agents du champ en référence à des positions spécifiées. L'exemple peut être ici celui des débats en cours autour de la mise à disposition des bases de données ou des dossiers documentaires élaborés par les chercheurs, par exemple lors de la préparation d'une thèse ou d'un ouvrage. Cela revient, si nous reprenons notre schéma d'ensemble, à dire que, puisque les annotations produites par les chercheurs sont aujourd'hui dans une large mesure nativement numériques, alors s'ouvre la possibilité de les mettre massivement et pour un coût marginal faible à disposition de la communauté, ce qui représente de fait, par le changement d'échelle de la taille des corpus structurés disponibles, la promesse d'une augmentation de la capacité interprétative des chercheurs . Imaginons un chercheur travaillant sur la diffusion des problèmes arithmétiques dans l'Europe médiévale. Il peut espérer ainsi pouvoir raisonner sur trois cents formes de problèmes et non plus cinquante, couvrant une aire géographique un peu plus vaste ou une période un peu plus longue. Il peut espérer tant une plus grande fiabilité et pertinence des indicateurs statistiques qu'il construit qu'une interprétation plus complexe parce que l'espace des principes interprétatifs qu'il élabore se voit doté de quelques dimensions de plus, cependant que le changement d'échelle du corpus autant que la possibilité de traitements nouveaux de ses éléments, offre la possibilité du surgissement de questions nouvelles. De fait, la revendication d'une « libération des données » est portée par nombre d'acteurs, souvent ingénieurs, ou issus des centres de recherches français les plus prestigieux, au sein de la mouvance qui tente aujourd'hui de faire émerger des humanités numériques à la française et relayée par un certain nombre d'acteurs au sein des institutions françaises de recherche qui prônent, pour le moins, le libre accès aux données produites par les agents CNRS ou les thésards ayant bénéficié de financements publics, en conformité d'ailleurs avec la déclaration de Berlin de 2003, signée par les représentants de plusieurs organismes de recherches européens .

Le dossier pourtant n'avance guère et ne semble pas générer beaucoup d'écho parmi les historiens eux-mêmes. Jouent bien sûr des difficultés techniques, tant de format que de structure des données, mais il est permis de penser que les envies des uns et des autres sont aussi fonction de leur position au sein de la chaîne de traitement du signal que nous avons dessinée. Il faut, en l'état actuel des techniques, pour réutiliser efficacement, c'est-à-dire sans que l'opération prenne trop de temps, les données structurées d'une enquête que l'on n'a pas élaborée, de véritables compétences informatiques (à tout le moins la maîtrise des expressions régulières, l'habitude des bases de données et une capacité minimale à programmer), ce qui réserve l'opération aux mieux dotés en capacités cognitives ou aux membres des institutions les mieux pourvues en intelligence informatique, qui sont souvent les mêmes. Il est assez peu sur-

prenant dès lors que de tels acteurs réclament l'ouverture des gisements de données, demande que les institutions de la recherche, dans la mesure où elles sont tenues de promouvoir la productivité du travail scientifique, ne peuvent que prendre elles-aussi à leur compte. Que beaucoup de chercheurs fassent la sourde oreille ou ne voient pas l'intérêt de la chose n'est guère plus surprenant. Ils participent d'une culture et d'une organisation de la recherche qui fait de l'historien un accumulateur de fiches et de données, qu'il peut espérer transmuer en position dans l'institution au moyen de la production de textes dont ce butin est le garant. Dans bien des cas, lui proposer d'accepter que l'on accède à son trésor en échange d'un droit de visite à une infinité d'autres dépôts revient à le payer en monnaie de singe, ou, plus exactement, à le déposséder du fruit de son travail, qui ne peut plus faire office de capital. L'accumulation et l'organisation de son dépôt n'ouvrent pas en effet, en l'état actuel des choses, droit à rémunération, ni financière ni symbolique. Les seuls produits échangeables de l'activité de l'historien – qui d'ailleurs prend généralement part à différents marchés – sont les mises en récit que, sous diverses formes, il produit à partir de son activité de rassembleur et de manipulateur de signes. Ouvrir son entrepôt à tous revient à le mettre en concurrence avec les membres les mieux équipés de sa corporation, tout en lui retirant l'exclusivité de l'accès à ses matériaux. C'est lui proposer un libre accès à des données dont il ne peut faire information, à moins de consentir de lourds investissements d'appropriation, qu'il ne peut pas nécessairement assumer, tout en lui ôtant l'exclusivité d'usage d'un bien rare. De plus sa différence, ou sa compétence spécifique, qui est souvent connaissance intime d'un fond longuement pratiqué, est brutalement dévaluée, au sens où est affirmée – peut-être un peu vite – la possibilité que celle-ci soit codifiée, détachée de son porteur et échangeable et s'ouvre donc la possibilité ou la promesse d'une division du travail qui est aussi une hiérarchisation des producteurs, dont les lignes tendent à recouvrir les gradients de prestige et de ressources existant, tout en faisant de certains ingénieurs des alliés indispensables. Les termes du débat sont cependant probablement tout à fait provisoires. Les rapides progrès du traitement d'images, pratique aujourd'hui devenue une discipline autonome, sont susceptibles de modifier à moyen terme pour les historiens les conditions matérielles d'acquisition du signal, en ouvrant la possibilité d'une automatisation partielle de la numérisation et de la structuration de partie de l'information supportée par l'archive, soit, dans le langage historien, de constitution et d'organisation du corpus et donc le rapport à celui-ci. Numérisation donc ne rime pas forcément avec démocratisation et moins encore avec égalité, mais ces exemples ont du moins l'avantage d'attirer l'attention sur le fait que la transformation des outils de l'historien ne conduit sans doute pas seulement à une évolution de ses produits, mais aussi à des mutations de la

profession historienne, voire de la figure même de l'historien, qui participent d'un processus engagé depuis déjà plusieurs décennies, qui ne peut donc être assimilé à un brutal changement d'ère dont le terme serait proche.

Le détail de cet argumentaire n'est pas indispensable à retenir ici. La chose qui me semble importante est qu'il est possible aujourd'hui de définir l'activité historienne comme la manipulation de stocks de chaînes de symboles, à l'aide d'une instrumentation changeante qui fait appel à l'informatique et au langage mathématique. Il convient donc de s'équiper et de réfléchir à son instrumentation, sachant que les formes en sont, rapidement, changeantes, ce qui suppose une compréhension des logiques en cours et des possibilités, plus que la seule maîtrise de quelques procédures (ou logiciels), qui ont de bonnes chances d'être rapidement obsolètes.

Sur les transformations de l'histoire à l'ère numérique on pourra voir aussi [?], dont proviennent certains éléments de ce cours, [?], [?], [?] et surtout [?]. Sur la question de l'usage du web comme source pour la période très contemporaine, sur laquelle je ne m'étend pas ici, parce qu'elle concerne très peu d'entre vous, on commencera par noiret2005.



# Chapitre 2

## Outils. De la trace au comptage

Nous avons considéré lors du premier cours, que produire un discours historique rigoureux suppose de constituer une collection organisée de notices liées à des objets, considérés comme les garants - la garantie si vous voulez - du discours que l'on tient. Il demeure tout à fait possible de se livrer aux opérations que cela nécessite sans utiliser un ordinateur. Celui-ci cependant offre des possibilités spécifiques de manipulation et de classement, à condition d'un apprentissage plus ou moins ardu selon la complexité des instruments que l'on utilise.

Nous allons dans les pages ci-dessous, suivre ce fil, en nous arrêtant, pour chaque opération, sur certains des outils disponibles, sans prétendre à l'exhaustivité, impossible aujourd'hui.

### 2.1 Acquisition de l'information

Mode d'acquisition

#### 2.1.1 Prélever

Il est possible aujourd'hui de trouver directement depuis le web de nombreux et très divers contenus sous forme de textes, d'images, de documents sonores ou vidéos. Je vous renvoie, pour leur repérage, au cours de méthodologie générale de Monsieur Demont qui contient d'excellents conseils pour les historiens, me contentant de signaler que les archives nationales maintiennent un site permettant de retrouver les fonds numérisés par l'ensemble des archives publiques, [http ://www.archivesdefrance.culture.gouv.fr/ressources/en-ligne/](http://www.archivesdefrance.culture.gouv.fr/ressources/en-ligne/), et que pour ceux qui travaillent sur des domaines étrangers, le site europeana, quoique difficile à manier, [http ://www.europeana.eu/](http://www.europeana.eu/), permet de retrouver

de nombreuses ressources abritées par des institutions européennes. Le très précieux internet archive permet lui de trouver de très nombreux matériaux de toute nature pour l'histoire contemporaine (<https://archive.org/index.php>) . Je ne peux cependant que répéter son avis. Il est fort probable que vous serez fascinés par la richesse de ces ressources, et tentés d'y naviguer, mais qu'elles renferment au final peu de données qui soient directement utiles à votre recherche. Le volume des ressources numérisées est infime au regard de ce que renferment les bibliothèques et les centres d'archives.

Si cependant vous trouvez des ressources utiles, Il faudra r pouvoir les récupérer et les stocker efficacement, ce qui implique de savoir deux ou trois choses.

Une ressource internet est définie (si nous ne nous occupons que de ce qui nous intéresse directement), par un mode d'accès et un format. Certains sites permettent de parcourir tous les fichiers, en passant par une arborescence (c'est à dire un inventaire), ce sont généralement des sites dits statiques, d'autres ne livrent les informations qu'ils renferment qu'au prix de l'interrogation d'une base de données (c'est par exemple le cas de Gallica).

**aspirer un site** L'avantage indéniable des premiers est qu'ils est facile de récupérer la totalité de leur contenu sur le disque dur de votre machine, en conservant l'organisation du site. Cela peut-être intéressant si vous travaillez par exemple sur l'histoire d'une association ou d'un groupe militant dont une bonne partie de l'activité passe par un site internet, ou bien si un particulier ou une institution a numérisé une source importante pour vous. Vous pouvez bien sûr enregistrer les pages ou les contenus un par un, en utilisant votre navigateur, mais cela devient rapidement fastidieux, alors qu'il existe de très nombreux logiciels (nommés aspirateurs de site), qui s'acquittent très bien de cette tâche, de même que des modules d'extension pour les principaux navigateurs. On peut citer parmi ceux-ci, HTTrack Website Copier, disponibles pour toutes les plate-formes ou le petit sitesucker pour Mac, mais ce ne sont que des exemples il en existe des dizaines.

**Formuler une requête** Beaucoup de sites savants aujourd'hui proposent un accès à leur contenu par le biais d'une requête adressée à une base de données. La qualité de l'information obtenue dépend alors de votre capacité à formuler un requête pertinente et précise, cela suppose de savoir deux ou trois choses de la logique sous-jacente à ces répertoires de documents. La plupart (les sites permettant d'accéder à des collections d'articles, les sites des bibliothèques par exemple) sont en fait constitués d'une interface graphique vous permettant de poser une question (de construire une requête en termes

informatiques) qui est adressée à une base de données qui la plupart du temps est une simple table. Chaque ressource (textes, articles, image) est décrite au moyen de descripteurs standardisés (nom d'auteur, lieu d'édition, titre etc ..). Dans certains cas, l'une de ces rubriques est le texte inscrit dans la ressource (le texte complet d'un article par exemple). Pour utiliser efficacement ce type de dispositif il faut garder à l'esprit quelques principes simples :

- tous ces sites ont une même structure, ce qui permet une interrogation à l'aide des deux mêmes opérateurs fondamentaux que sont et, ou, sauf pour désigner les liens entre les conditions que vous posez
- il est généralement possible, pour les rubriques contenant du texte (le titre d'un ouvrage par exemple) de demander si la rubrique contient ou non une chaîne de caractères (attention, l'informatique ne connaît pas des mots, mais des suites de symboles, donc maison n'est pas maisons, et pas forcément Maison non plus).
- pour les rubriques contenant un nombre il est normalement possible de demander si le nombre est égal à une valeur, ou bien supérieur ou inférieur à une valeur (donc nous pouvons demander à gallica par exemple quels sont les documents proposés publiés entre 1875 et 1894 contenant "colonie" dans le texte).

Ces éléments, simples en eux-mêmes permettent de composer des requêtes très précises. Je peux ainsi demander à Gallica quels sont les documents détenus, édités entre 1834 et 1898, de type image, dont le titre contient la chaîne de caractère "Alpes", si je m'intéresse aux représentations de la montagne au XIXe. C'est là une requête formulée en langage naturel, dans certains langages informatiques cela se traduirait à peu près par date<1898 and date>1834 and type="image" and title includes "Alpes".

C'est là que les choses se compliquent un peu pour nous utilisateurs. Il y a en informatique beaucoup de façons de décrire une table de données, la requête formée doit respecter parfaitement les normes du langage informatique utilisé (si un guillemet n'est pas fermé, ou s'il manque un ; le serveur rejette la requête) et les sites internet n'offrent pas toujours au premier coup d'oeil l'accès aux normes qui les structurent (savez-vous ainsi que si, utilisant google, vous tapez Guy Mollet vous aurez toutes les pages contenant Guy et Mollet, ou bien guy seulement ou bien Mollet seulement alors que si vous tapez "Guy Mollet" vous aurez toutes les pages contenant exactement la chaîne de caractères Guy Mollet). La plupart du temps la page "recherche" ne vous propose qu'une simple cartouche d'interrogation (à la google), dont le fonctionnement est rarement expliqué. Pour arriver à la page permettant de produire une requête complexe, il faut la plupart du temps chercher un lien donnant accès à une "recherche avancée" vous permettant d'opérer des choix par le moyen de boîtes à cliquer. Si vous utilisez souvent une ressource, ou

si votre première recherche vous donne un résultat peu satisfaisant, (trop de réponses ou bien des réponses qui ne correspondent pas à ce que vous cherchiez) je vous incite à chercher s'il n'existe pas un mode de recherche avancée et à vous souvenir que votre requête a d'autant plus de chances d'être efficace que

- vos conditions sont nombreuses
- vous utilisez des mots signifiants et précis, dans le cas d'une interrogation sur champ textuel, l'utilisation des noms propres quand elle est possible est souvent très efficace
- vous avez consulté l'aide ou la notice du site indiquant le mode d'interrogation (quand elle existe, mais c'est un autre problème)
- vous n'hésitez pas si vous n'êtes pas satisfaits par les premières réponses à tenter plusieurs requêtes

**Types de contenu** Les documents qui vous intéressent, particulièrement en histoire contemporaine, peuvent être de types très divers (audio, video, textuel, images par exemple). Pour chaque type de ressources existent une multitude de formats, pas toujours compatibles entre eux et dont les noms peuvent être un peu barbares (tiff, jpg, gif pour les images par exemple, mais il en existe des dizaines). Si vous avez besoin de manipuler et de cataloguer de nombreux fichiers, (ou même simplement à convaincre votre machine d'ouvrir le superbe extrait du journal télévisé, qui avait l'air si passionnant) vous aurez besoin d'un convertisseur permettant d'obtenir des fichiers au format le plus adapté à votre équipement. Là encore il existe à la fois des sortes de couteaux suisses capables d'ouvrir à peu près tous les fichiers de tel ou tel type et des briques logicielles chargées de la conversion de tel format spécifique en un format plus commun. Les lister tous, vu le nombre n'a aucun sens, mais il est très souvent possible de les repérer au moyen d'un moteur de recherche par le biais d'un requête la plus précise possible (et pour le coup souvent en anglais).

Pour connaître le nom du format qui vous résiste il faut regarder le nom complet du fichier. Celui-ci est classiquement de la forme monfichierà-moi.nomduformat. Si donc un fichier vous résiste qui se nomme information.rar, cela veut dire qu'il s'agit d'une archive formée en utilisant rar, et si vous ne pouvez pas l'ouvrir il faut aller demander à un moteur de recherche quelque chose comme rar software open convert et le nom de votre système d'exploitation. Remarquez que la logique est la même dans l'utilisation d'un moteur de recherches que celle que nous avons évoquée tout à l'heure, plus la requête comprend de termes signifiants et précis, plus elle a de chances d'être efficace.

### 2.1.2 Numériser

L'historien est de tous les spécialistes de sciences humaines, le moins bien traité par les mondes numériques. Très souvent en effet, son matériau, qu'il s'agisse, de textes ou d'archives, n'est pas nativement numérique, n'a pas été numérisé et ne le sera pas dans un avenir proche. Les documents issus des archives publiques aujourd'hui disponibles sous forme de fichiers numériques ne représentent qu'un tout petit pourcentage des documents renfermés dans les archives.

Si l'on veut pouvoir utiliser les ressources informatiques pour traiter son matériau, il faut donc d'une façon ou d'une autre en fabriquer une empreinte numérique (insistons une fois encore, ce n'est pas toujours nécessaire et peut-être contreproductif, passer des heures à cliquer des fonds d'archives que vous n'aurez jamais le temps de consulter et de traiter, si c'est aujourd'hui un péché fréquent chez les jeunes - et moins jeunes - historiens n'est pas seulement inutile, c'est aussi perdre un temps précieux pour une recherche difficile à mener dans les temps impartis).

Il faut donc là aussi décider d'une stratégie en fonction d'un certain nombre de critères

- Quel usage pour le document que vous vous proposez de reproduire. S'agit-il d'un document iconographique que vous pourrez vouloir reproduire dans de bonnes conditions, ou bien d'un document d'archives dont vous souhaitez simplement prendre connaissance ? Aurez-vous besoin de pouvoir accéder à des détails ?
- Quels sont les conditions d'accès à la ressource ? Si un fond d'archives vous attend sagelement à la BDIC (cette bibliothèque située à deux pas du campus en comprend de nombreux et très riches), inutile d'aller passer des heures à le reproduire *in extenso*, quand vous pourrez facilement vous y reporter, si par contre vous avez des fonds importants à Aix en Provence ou à Roubaix et que vous ne pourrez que passer quelques jours sur place, vous n'aurez pas beaucoup d'autre choix que de cliquer rapidement les cartons qui vous semblent importants pour en prendre connaissance plus tard.
- Quelles sont les caractéristiques du lieu où vous devrez opérer (autorisations nécessaires, luminosité, possibilité ou non d'une alimentation électrique).

C'est en fonction de ces critères que vous choisirez le matériel dont vous avez besoin et votre mode opératoire.

**Le matériel** Sauf si vous êtes par ailleurs un professionnel de la chaîne graphique (ou en connaissez un), votre choix de matériel se limite à arbitrer

entre un simple smartphoe, dotés aujourd’hui de capteurs d’assez bonnes qualités et d’assez de mémoire pour prendre pendant une séance de travail les clichés des pièces qui vous semblent particulièrement importantes et aux- quelles vous souhaitez pouvoir vous reporter, d’un appareil photo numérique, nécessaire si vous souhaitez prendre des images permettant de préserver les détails, ou si vous envisagez de cliquer rapidement des cartons entiers parce que vous disposez de peu de temps pour consulter un fond éloigné, ou d’un scanner sur pied (mais l’équipement est onéreux), si vous avez besoin de cli- chés de grande précision (c’est le cas de ceux d’entre vous qui envisagent d’utiliser un système d’information géographique, a priori peu nombreux). Nul besoin aujourd’hui de vous ruiner pour disposer des outils nécessaires, hormis des besoins très spécifiques (travail en basse lumière, besoin de grande précision), n’importe quel appareil compact sur le marché aujourd’hui suffit à vos besoins.

N’oubliez pas par contre chargeurs et/ou batterie de recharge, une carte mé- moire de secours, particulièrement si vous travaillez dans un fond éloigné que vous n’avez que quelques jours (ou quelques heures) pour explorer. et tes- tez votre matériel à blanc (chez vous ou dans un fond d’archives), avant de procéder à une campagne de reproduction importante. Vous aurez besoin de déterminer quelles positions et quels réglages sont les plus adaptés à votre matériel et à votre manière de procéder. Vous pourrez avoir besoin, pour éviter les effets de bougé, d’un déclencheur souple - il en existe aujourd’hui qui s’adaptent à un appareil photo mais aussi à un smartphone - ou d’un pied. Là aussi il en existe une grande variété, selon la taille et la forme.

**Les formats** S’il s’agit simplement de garder trace de documents que vous souhaitez pouvoir déchiffrer si vous avez besoin de vous y reporter, vous n’avez guère à vous soucier du format. Le jpg, généralement proposé par dé- faut, par votre appareil suffit très largement. Si vous avez besoin d’une grande qualité de reproduction ou de détails, il peut être nécessaire de choisir dans vos réglages la qualité maximale, ou même un fichier RAW qui conserve toutes les informations fournies par les capteurs. La qualité des images est bien supérieure à celle offerte par un format compressé, et permet des correc- tions fines, mais son usage suppose des compétences en traitement d’image.

**Ne pas oublier** Surtout, et c’est valable dès que l’on produit des données informatiques

- Vérifier régulièrement en cours de travail la qualité de vos prises de vues
- Sauvegardez fréquemment

- Notez scrupuleusement ce que vous faites, et à quoi correspondent les clichés que vous prenez, soit sur une feuille à part, soit sur un document informatique. Des fichiers qui ne sont pas documentés sont inutilisables, car vous serez dans l'incapacité de les attribuer et donc de les interpréter et de les citer.

### 2.1.3 Liquidité de la donnée

En fonction de votre sujet et de votre dispositif de recherches, l'usage que vous ferez de vos documents informatiques peut varier considérablement. Dans la plupart des cas, il s'agit de savoir pour vous, si vous désirez simplement prendre connaissance des informations contenues dans les fichiers, en ce cas, le format image suffit ou bien vouloir explorer et étudier systématiquement un texte, si vous faites par exemple une étude de presse. Il vous faut dans ce cas reconstituer le texte à partir de vos fichiers. Le processus est la plupart du temps très long, il est un peu hasardeux de s'y lancer dans le cadre d'un mémoire de M. Certains outils peuvent cependant faciliter la tâche. Si vous souhaitez tenter l'expérience, deux possibilités s'offrent à vous.

- Si vous disposez de fichiers de bonne qualité reproduisant un texte imprimé, ou d'un fichier pdf image obtenu par internet, vous pouvez essayer d'utiliser un logiciel de reconnaissance optique de caractères (OCR). La plupart des logiciels de ce type sont payants (et chers), il existe cependant depuis quelques années quelques logiciels open source performants dont Tesseract. Malgré les progrès indéniables de ces outils au cours de ces dernières années le résultat est souvent frustrant pour l'historien, il faut pour que la reconnaissance soit de bonne qualité, une très grande qualité de reproduction et un très bon état du support physique, conditions rarement réunies. Même dans des conditions optimales, il faut vérifier la qualité des sorties avant traitement et souvent apporter des corrections.
- Une alternative est donc l'usage de logiciels de reconnaissance vocale, qui transcriront votre lecture du texte. Il en existe aujourd'hui de nombreux et de très bonne qualité, certains même en open source. Le gain de temps par rapport à une saisie au clavier est très appréciable, au prix d'un apprentissage qui n'est pas très exigeant.

Il est probable que vous vous retrouviez dans le cours de votre travail avec de nombreux fichiers (d'images par exemple), sur lesquels vous allez vouloir opérer des manipulations souvent très répétitives (changer le nom de tous les fichiers d'un dossier par exemple, ou bien changer la taille ou la définition de tous les fichiers d'un dossier). Il est possible, et efficace, de passer par des routines programmées, mais cela est rarement accessible aux

historiens, cependant des outils spécialisés existent qui permettent de traiter simultanément de grandes quantités de documents (c'est à dire d'appliquer les mêmes transformations à ceux-ci). Graphic converter par exemple, qui est disponible aujourd'hui pour la plupart des plate-formes informatiques est très efficace dès lors qu'il s'agit de traiter d'opérer des transformations sur de nombreux fichiers images, les plus difficiles à manier pour le novice.

### **2.1.4 stocker, sauvegarde et le nuage**

Pensez toujours, particulièrement parce que vous vous attaquez à un projet au long cours, à sauvegarder fréquemment le produit de votre travail et à en faire des copies. Vous disposez aujourd'hui de très nombreuses solutions, soit au moyen de supports externes, disque dur, ou clés usb, soit au moyen des services de ce que l'on appelle le cloud, en fait des parcs de disques durs gérés par un prestataire auquel votre machine est connectée par le biais d'internet, qui sont nombreux à vous proposer des espaces gratuits de stockage (dropbox par exemple et bien d'autres).

## **2.2 Inventaire et description. Structurer l'information**

Vous allez, au cours d'un master, accumuler une grande quantité d'information, parfois très hétérogènes. L'enjeu est à la fois de pouvoir retrouver rapidement l'information, mais aussi de pouvoir faire surgir des associations pertinentes, et pas toujours pré-déterminées, quand vous traitez une question ou un dossier. Les questions que vous vous poserez en fin de recherche ne sont pas forcément celles que vous aviez en tête en débutant et ne sont pas toujours prévisibles, il est donc utile de pouvoir opérer des recoupements inédits.

Cela suppose à la fois de documenter l'information engrangée (d'associer des descripteurs aux éléments que vous détenez), que cette information soit structurée (nous aurions dit autrefois classée, si les opérations sont de nature différente la fonction de la structuration de la donnée est de même nature que le classement) et que vous puissiez la retrouver (c'est à dire former une requête efficace). Si classiquement (c'est à dire au cours des deux ou trois décennies écoulées) parvenir à ces fins supposait de concevoir et de renseigner une base de données, les dispositifs aujourd'hui disponibles, parfois très spécialisés en fonction de la nature des données ou des formats, sont beaucoup plus divers et permettent de bénéficier de la puissance d'un système d'information sans posséder de compétences particulières. Les dispositifs logiciels qui permette

cela reposent tous sur une architecture de base de données, mais celle-ci n'est souvent plus directement accessible à l'opérateur et encore moins conçue par lui. On perd parfois en précision, mais la pente d'apprentissage des outils est beaucoup moins rude et les rend accessible à tout un chacun pourvu que l'on soit un peu ordonné.

### 2.2.1 Documenter la donnée, trouver l'information

Vous pouvez choisir, ou non, de disposer sous forme numérique de copies des documents que vous jugez particulièrement intéressants, d'articles ou de textes traitant de sujets proches du votre, voir des notes que vous prenez tout au long de votre travail. Répétons le au risque de lasser, une documentation est parfaitement inutile si elle n'est pas classée et décrite et s'il n'est pas possible de trouver facilement une réponse à la question que l'on se pose. C'est vrai dans l'univers du papier, ce l'est encore plus dans le monde numérique. Il faut donc à la fois que les documents que vous accumulez soient accompagnées d'éléments permettant de les décrire et que vous disposiez des moyens de trouver réponse à vos questions.

**Les expressions régulières** Dans le cas des documents stockés sous formes de texte (vos notes par exemple), les contraintes de documentation des fichiers sont aujourd'hui un peu plus relâchées. Il est aujourd'hui possible de procéder à une recherche qui permette de repérer un mot dans le contenu même du fichier. Une recherche par mots clés, associant plusieurs critères, sera cependant toujours plus rapide et plus précise que celle effectuée de cette façon.

Il est possible aussi que, travaillant un texte, vous vouliez trouver rapidement des passages qui ne peuvent pas être repérés simplement par la mention d'un mot ou d'une chaîne de caractère. Vous pouvez par exemple vouloir trouver tous les noms propres présents dans un texte, ou bien toutes les dates qu'il contient, ou vérifier lors de la rédaction que toutes vos parenthèses ou vos guillemets sont fermés, ou que toutes vos phrases débutent bien par une majuscule. Bref vous pouvez avoir besoin de formuler une requête complexe portant sur un ou plusieurs fichiers textes, ce qu'il n'est pas facile de faire depuis la fenêtre de recherche proposée par votre système. Il existe un système très puissant permettant de faire des recherches très poussées dans des documents texte auquel vous pouvez accéder simplement en utilisant certains traitements de textes usuels, en particulier celui des suite néo-office, ou open office, il s'agit des expressions régulières. Vous pouvez y accéder simplement en ouvrant la page "rechercher" de ces traitements de

textes, et en cochant la case plus d'options, vous verrez alors que vous pourrez cocher une case expression régulière, qui vous permettra, par exemple de trouver toutes les phrases contenant des chiffres, ou tous les doublons présents dans votre texte. Ce système très puissant, dont il est difficile de se passer quand on manipule de grandes quantités de données textuelles, suppose cependant un apprentissage. Il est certes facilité par la présence de nombreuses ressources dédiées sur internet (pour vous donner une idée, [https://wiki.openoffice.org/wiki/FR/Documentation/ExpressionsRegulieres\\_dans\\_writer](https://wiki.openoffice.org/wiki/FR/Documentation/ExpressionsRegulieres_dans_writer)), il n

**Systèmes d'information généralistes** Les spécialistes de l'information (journalistes par exemple), utilisent de plus en plus des systèmes d'information permettant de classer et d'annoter des documents de divers types (texte, pages web, images etc). Ces outils reposent presque tous sur de mêmes concepts. Un document, (une note, un article, un cliché), est intégré à la base documentaire accompagné de descripteurs. L'utilisateur est incité à ajouter des "tags" en fait des mots clés, permettant d'identifier le contenu de la note. Il est ensuite possible de faire des recherches complexes dans l'ensemble documentaire réuni. Ce sont des systèmes puissants et efficaces pour ceux qui doivent gérer des volumes d'information importants et des éléments disparates au cours de projets longs. La prise en main de ces outils est aujourd'hui rapide. Je vous recommande sinon de les adopter (certains sont tout à fait allergiques à ce type d'outils qui ne correspondent pas à leur mode de travail et qui demandent non pas des compétences informatiques, mais une pensée déjà solidement structurée permettant de nommer de façon pertinente les éléments engrangés, du moins de regarder de près de quoi il s'agit. L'un des plus populaires, qui a l'avantage d'être gratuit dans sa version de base (qui a de bonnes chances de très largement suffire à un étudiant en histoire) est Evernote

**Les outils dédiés** Les besoins spécifiques de certains métiers ont conduit à l'apparition de système de gestion d'information puissants spécialisés dans le maniement de tel ou tel type de ressources. Trois types peuvent particulièrement intéresser les historiens.

*Les logiciels de bibliographie*, dont les plus connus sont bibdesk et Zotero, permettent d'organiser facilement vos données bibliographiques et de conserver aussi vos notes de lectures associées à la référence complète du document que vous avez consulté. Ils présentent deux avantages indéniables. Ils permettent d'extraire automatiquement les références produites lors d'une

## 2.2. INVENTAIRE ET DESCRIPTION. STRUCTURER L'INFORMATION 29

requête à certaines grandes bases de données bibliographiques (les 25 références obtenues à l'aide du site du sudoc peuvent être automatiquement versées dans le logiciel sans que vous ayez besoin de les taper), ils permettent d'harmoniser assez simplement la présentation d'une bibliographie (vous serez certains que le lieu d'édition est toujours placé au même endroit dans votre bibliographie). Les historiens tendent à utiliser de préférence Zotero, qui a d'ailleurs été développé par un laboratoire d'histoire. De ce fait des formations dédiées existent, et la Bdic en organise à destination des étudiants d'histoire sur le campus de Nanterre

### *Les catalogueurs d'image*

La forme la plus familière de l'irruption du numérique dans les pratiques de l'historien est souvent l'usage de l'appareil photo numérique, utilisé pour cliquer rapidement des documents d'archive (parfois des cartons entiers). Il arrive assez fréquemment cependant qu'ils aient du mal à classer et interroger ce matériel, particulièrement quand des milliers de fichiers portant des noms peu amènes envahissent les disques durs. Il existe cependant des outils simples d'usage permettant de décrire, de classer et d'interroger les réservoirs ainsi constitués, prenant appui sur le fait que le fichier qui vous permet de visualiser une image sur votre écran, ou de l'imprimer est en fait un texte stockant beaucoup d'information, vous permettant d'en ajouter d'autres et d'opérer des recherches.

Parmi ces données on trouve les données Exif, généralement automatiquement générées par l'appareil numérique, et qui comportent essentiellement des informations sur les conditions de la prise de vue (date, temps d'exposition, ouverture, etc ..) et, souvent plus intéressantes pour nous, les données Iptc. Elles permettent l'ajout au fichier lui-même (et non à une base externe) de descripteurs structurés. Vous pouvez y indiquer un titre pour l'image, des mots-clés, la provenance du cliché, ainsi qu'y entrer des notes libres (du texte).

L'existence de ces données est souvent méconnues (sauf des professionnels de l'image qui les utilisent systématiquement), et il faut pour les faire apparaître soit opérer en mode console (ce qu'il est un peu douteux que vous fassiez fréquemment), soit utiliser un logiciel adapté permettant de les visualiser, de les éditer et d'opérer des recherches sur celles-ci. Les logiciels de retouche d'images, comme les catalogueurs d'images le permettent généralement aujourd'hui (Digikam, Photoshop, Grapic convertor). C'est un moyen simple (là encore pas besoin de compétences particulières), de transformer vos dossiers plein de clichés en base de données efficaces.

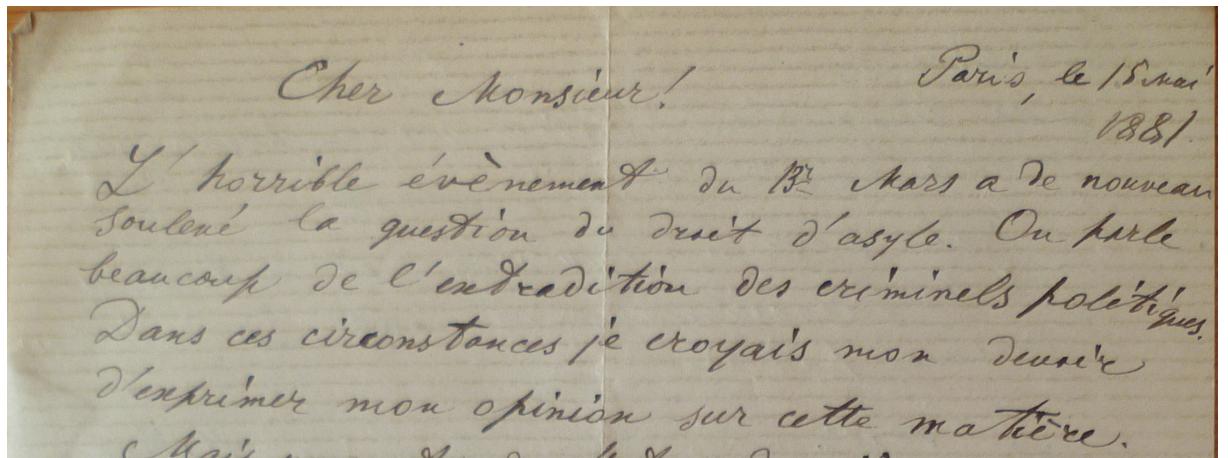


FIGURE 2.1 – Un document cliché dans un centre d'archives tel qu'il apparaît à l'écran

```
?n?6?.?wZp?T?IX~??p-l?_g????&b?Mg???A????#?#"?"k"?9?C??\[nw6?*?4j:?:*?4z&?*@?l?Mly?\?C\U??x???????"A?  
?mI????s<*/?o?q??C?:???????)?|~c*??!+?.?1?k??>?a?Q??>30????-[????:??u<?????Q?3??n[?vh{????;?vL?t?yr  
al@\????ow????S????w-?b????=?{j?G}?????{+?~)??,4(,/???)?????,??7XlZ|?5?+?o??d?DiV????e?????Qn\^?{p(?  
??^?R????SUL?H?)u[?|?_?????Gl???*??~<w?A?K]G?F}?1?c?????75?66~>;!B?w??t??Y??nIo?=v??v??1?  
?s???m?????/?+??? y???F1|' ??Y?Ks?c.0?n?}|%??>o?W?_s?v???u???oX?Bw?}?? ????&??M;??k~?{h?Px???  
????w?kd???h? ?1??_??x??x?8z??????0?????M`*?0?81?liÜ?d????i*?9?y???ty???Y?_?/z??ri.?w?k^i?????  
??^?/_?Y?F??o{|??Kz??X?^?????>N/e~?~?????0????e?  
?\\@?"???D. r?  
?\\@?"???D.??_?n???  
??????6?UhD@ ?`|?,?iTtxtXML:com.adobe.x  
mp<xpacket begin="" id="WSM0MpCehiHzreSzNTczkc9d">  
<x:xmpmeta xmlns:x="adobe:ns:meta/" x:xmptk="XMP Core 5.1.2">  
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">  
    <rdf:Description rdf:about=""  
      xmlns:photoshop="http://ns.adobe.com/photoshop/1.0/">  
      <photoshop:TransmissionReference>fonds rivier IS1977 boîte D1</photoshop:TransmissionReference>  
      <photoshop:City>Paris</photoshop:City>  
    </rdf:Description>
```

FIGURE 2.2 – Le même fichier vu depuis la console

2.2. INVENTAIRE ET DESCRIPTION. STRUCTURER L'INFORMATION31

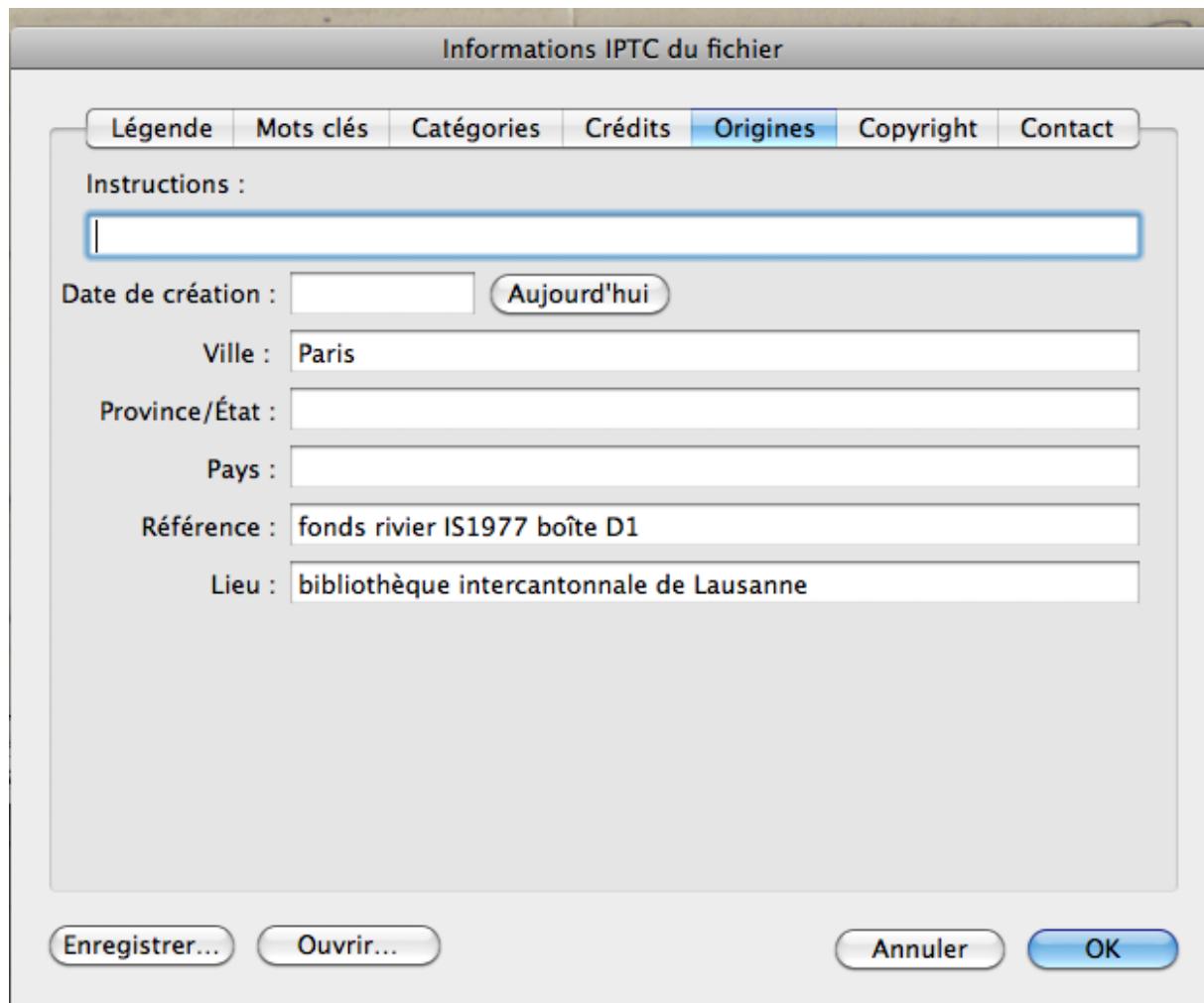


FIGURE 2.3 – Consultation des informations du fichier depuis Graphic converter

*Les systèmes d'informations géographiques*

Certains d'entre vous, mais ils sont peu nombreux, pourront avoir besoin d'utiliser des systèmes spécialisés dans la gestion des informations géographiques (SIG). Ces outils sont très puissants et très efficaces et permettent de cartographier et de visualiser des informations dont la caractéristique est qu'elles sont spatialisées (c'est à dire liées à un point dans l'espace). Ces outils, même s'il est possible aujourd'hui d'utiliser des outils robustes disponibles en open source - la communauté scientifique semble aujourd'hui beaucoup utiliser Quantum - sont très complexes (raison de leur puissance) et exigent de solides compétences géographiques aussi bien que statistiques, voire informatiques. La plupart du temps en effet l'historien ne peut pas s'appuyer sur une base existante (il n'en existe pas pour Limoges au treizième siècle pas plus que pour Nanterre vers 1880), il lui faut donc concevoir et développer un système complet. Le temps d'apprentissage n'est pas négligeable et ne se justifie que si vos données sont abondantes, précisément localisées et si les questions liées à la spatialisation des phénomènes sont centrales dans votre problématique (si par exemple vous consacrez votre mémoire à l'aménagement du port du Havre à la fin du XIXe siècle ou à la distribution des activités pastorales en haute Provence au début de la période moderne). Ajoutons qu'une sensibilisation antérieure à l'usage de ces outils (dans le cadre de votre activité professionnelle par exemple) est souvent la condition permettant de s'approprier avec succès ce genre d'outils.

Si vous voulez faire une idée des possibilités offertes par ce type d'outils aux historiens je vous incite à consulter le projet alpage dédié à l'espace parisien [http://dmap.tge-adonis.fr/alpage\\_public/flash](http://dmap.tge-adonis.fr/alpage_public/flash), ou l'atlas de la déportation des enfants parisiens de Jean-Luc Pinol ([http://tetradem.huma-num.fr/Tetrademap\\_Enfant\\_Paris](http://tetradem.huma-num.fr/Tetrademap_Enfant_Paris).

*Le texte comme donnée* Une collection de textes peut également être transformée en base de données et systématiquement structuré (voire un texte unique quand il s'agit d'éditer un document ancien dotés de nombreuses variantes). Nous sommes cependant là aussi dans le cas d'outils dont l'apprentissage se justifie si au centre de votre projet de recherches se trouve un certain type de matériau et un projet bien spécifique, essentiellement de l'édition de textes anciens. Les spécialistes en ce cas tendent aujourd'hui à structurer et documenter l'information en utilisant XML (extensible markup language). Il existe une large gamme d'outil facilitant la description xml d'un document texte. Il y a plusieurs avantages à cela. XML est aujourd'hui le langage d'entrée d'une bonne partie des analyseurs de texte. Il transforme de plus les données stockées dans le texte en éléments d'une base de données

interrogeable (à condition de passer par XML query, langage de requêtes dédiés, puissant mais loin d'être simple à maîtriser). Enfin comme la description xml revient à apposer des balises textes au sein d'un texte existant, il est possible de manipuler ces documents par le biais d'expressions régulières dans un simple traitement de texte. Ajoutons que l'acquisition des compétences permettant de structurer un document XML sont tout à fait à la portée d'un étudiant d'histoire, ne nécessitant pas de compétences mathématiques ou informatiques particulières, mais simplement beaucoup de rigueur et de patience, qualités tout à fait bénédictines. Enfin c'est un outil particulièrement adapté au travail collaboratif, facilitant les échanges entre spécialistes d'un même corpus.

J'invite ceux que ce domaine intéresse particulièrement à consulter Mattia Cavagna, Outils informatiques pour l'édition et le traitement des textes, des images, du langage, Cahiers de recherches médiévales et humanistes, numéro 20, 2010, pp. 357-390, en ligne à <http://crm.revues.org/12246>

logiciels de base de données Une large gamme d'outils, permettent de construire des bases de données, que l'on distingue par leur puissance ou bien parce qu'ils sont destinés à des usages spécifiques. Ces outils sont particulièrement adaptés (et indispensables), à l'étude de sources de type sériel (recensement, état civil, listes de toutes sortes) ou de collections d'objets (les personnes appartenant à un milieu, à une institution, à l'étude d'une procédure). Il est souvent également nécessaire de passer par la mise en place d'une base de données dans le cadre d'une étude impliquant des traitements quantitatifs.

Ce rapide panorama, qui ne concerne encore que les outils dédiés à la structuration de la documentation, conduit à rappeler avec insistance l'un des principes posés en introduction à ce cours. Les instruments disponibles sont aujourd'hui très nombreux, très divers et très précisément spécifiés. Vous ne les maîtriserez pas tous, et certains n'ont aucune utilité pour tous, même s'il est utile de connaître leur existence. Il vous faudra donc construire votre propre instrumentations en opérant des choix, ce qui revient toujours à arbitrer, en situation (quels matériaux ?) en fonction de besoins (quelles questions ?) et de ressources (quel coût, quel temps d'apprentissage).

### 2.2.2 Outils d'écriture

Une notable partie de votre temps va au cours des deux années à venir consister à produire des textes, qui auront ceci de particulier qu'ils seront

longs et nécessiteront l'usage de notes, d'où la nécessité d'utiliser des outils performants. La plupart d'entre vous opteront pour un traitement de textes, permettant de générer des tables, des index, et de bien gérer les notes. La discussion des avantages et des inconvénients des uns et des autres est un peu oiseuse. Les outils disponibles sont assez peu nombreux et l'arbitrage se résume généralement à un match entre un logiciel microsoft (word), ou un logiciel libre (OpenOffice ou NeoOffice). Les performances des uns et des autres sont assez similaires. Word a pour lui d'être un quasi standard de fait, les autres de pouvoir être utilisés gratuitement en toute légalité.

J'invite cependant les plus curieux d'entre vous, ou les plus attirés par l'univers des outils informatiques à se pencher sur LaTeX., un langage de composition de documents qui présente un certain nombre d'avantages. L'apprentissage initial est un peu plus difficile que celui d'un traitement de texte standard, mais vous disposez d'un contrôle beaucoup plus fin sur le rendu du document, d'une très bonne gestion des éléments graphiques insérés, en particulier des formules mathématiques, même si cela vous concerne peu, de plus, (et surtout pour l'auteur de ces lignes), les documents produits ainsi se révèlent pérennes et à épargner ou à peu près les désagréments nés des multiples changements de versions et de formats qui émaillent la vie des traitements de textes. Sur Latex la meilleure introduction en Français est sans doute [?], sur BibTex, son complément bibliographique, il existe une bonne introduction en Français [?].

# Chapitre 3

## L'historien et les statistiques

Un ordinateur sert aussi à compter, c'est même sa vocation première et à manipuler des données chiffrées. L'historien est souvent confronté à la nécessité de le faire et ce pour plusieurs raisons, nous explorerons dans cette section l'une d'elle, qui est le fait que les données que lui livrent les archives, particulièrement pour les périodes modernes et contemporaines prennent souvent la forme de données chiffrées, dont il faut faire source, et que l'on sera amené à citer, à représenter ou à manipuler, ce qui suppose, avant toute compétence informatique, quelques notions de leur nature et donc des usages qui en sont possibles.

Le nombre des comptages, tableaux, indices, que l'on peut retrouver dans les archives, augmente au fur et à mesure que nous nous rapprochons de la période actuelle. Ces données sont le plus souvent produites par des agents de l'état dans le cadre de leur activité, et ce d'autant plus qu'elles sont anciennes. Si l'état n'a pas le privilège de la production et de la collation de données statistiques décrivant le fonctionnement social et économique du territoire qu'il contrôle, il demeure le plus ancien et le plus constant producteur et consommateur de celles ci. Cela justifie que nous interrogeant sur ce que l'historien peut apprendre et faire des données statistiques élaborées par les contemporains des périodes qu'il étudie, nous accordions une importance particulière à la statistique publique parce qu'elle est à la fois la plus fréquemment rencontrée et le cadre de la naissance de procédures statistiques toujours à l'œuvre dont nous pouvons espérer, évoquant leur genèse, reconstituer la logique.

Il nous faudrait, si nous voulions être exhaustif nous inscrire dans un temps long et de dévider plusieurs fils à la fois. La statistique publique telle que nous la connaissons peut être vue comme le produit du croisement de plusieurs traditions

- une proprement scientifique qui est celle de la naissance et du perfectionnement des probabilités, dont les origines sont à chercher du côté de Pascal.

- Une liée au développement de l'état moderne et de l'édification d'un appareil permanent de collecte d'informations destinées au souverain, et nous pouvons là remonter au moins à Richelieu demandant aux intendants (1630) de recueillir régulièrement des informations sur leurs circonscriptions et de les faire parvenir au pouvoir central

Afin de garder à l'exposé une taille raisonnable, l'histoire des statistiques est aujourd'hui une branche en soi de l'activité historienne, et de ne pas introduire trop de complexité, nous nous intéresserons surtout ici au second de ces axes et au segment contemporain (après 1789) de cette histoire, en gardant en tête que la question fondamentale pour nous est celle des usages que l'historien peut faire des données dont nous examinons la constitution progressive.

## 3.1 La production du dénombrable

### 3.1.1 Préhistoires statistiques

Quand commence cette histoire, le terme de statistique, emprunté à l'allemand désigne une science de l'état élaborée pour les besoins du souverain et qui donne à voir à celui-ci, au moyen de descriptions littéraires l'étendue de ses états et les ressorts de sa puissance, en classant systématiquement les faits et les informations nécessaires à son appréciations, selon un schéma qui emprunte à la logique aristotélicienne. Les auteurs anciens distinguent généralement :

- - Les causes matérielles (territoire et population)
- - Les causes formelle (droit institution)
- - Les causes efficientes (moyens de l'état)
- - Les cause finales (buts de l'état)

Les manuels d'histoire des statistiques opposent souvent à cette approche l'arithmétique politique anglaise, qui supplanterait sa rivale au début du XIX<sup>e</sup>, parce que plus efficace. Celle-ci est un ensemble de procédures utilisant le calcul qui visent à proposer estimation la plus précise possible de phénomènes précisément définis (le chiffre de la population particulièrement, mais aussi assez vite l'espérance de vie).

### 3.1.2 La statistique napoléonienne

L’ Histoire en fait un peu plus compliquée. D’abord parce que longtemps coexistent au sein de l’appareil statistique d’état des approches que nous dirions littéraires et le calcul, ce dont la statistique napoléonienne fournit un bon exemple. Reprenant des projets révolutionnaires, l’ administration napoléonienne crée un service statistique permanent. Les responsables de celui-ci lancent une grande enquête demandant à chaque préfet de répondre à des questionnaires détaillés. Leurs réponses fournissent encore aujourd’hui une source très importante pour l’histoire de la France du XIXe siècle. Elles ne prennent cependant pas toujours la forme d’évaluation chiffrée et ne respectent pas toujours non plus les demandes de l’administration centrale. La présentation que fait le préfet de l’Indre du travail accompli afin de mener à bien sa tâche constitue un bon exemple de ces hésitations.

Quatorze siècles de la monarchie s'étaient presque écoulés, et la France n'avait encore que des notions succinctes sur les contributions, la population, l'étendue et les principales ressources de ses généralités. (...). Ce que durant tant de siècles on n'avoit osé entreprendre, ce que le plus puissant monarque n'avoit pu accomplir, le Gouvernement Consulaire l'aura bientôt consommé. On n'aura pas compté trois années de son existence, que la République aura la Statistique complète de es nombreux départements.

Ainsi le même génie qui préside maintenant aux destinées de la France, et qui semble avoir commandé à tous les évènemens pour la soustraire à tous les périls accélère aussi pour elle l'époque de al prospérité. Par lui, elle a vu s'accroître sa puissance et sa gloire ; par lui, elle connoîtra enfin l'immensité de ses ressources ; et c'est à lui qu'elle devra bientôt ses richesses et son bonheur.

La Statistique est sur-tout La science des faits et des calculs ; et les fais sont tout ce qu'il y a de plus difficile à recueillir, principalement lorsqu'ils sont par-tout épars, inconnus presque par-tout ; et lorsque, en les recueillant, on peut effaroucher l'intérêt.

Ce n'étoit point pour améliorer le sort des provinces, que l'ancien Gouvernement cherchoit à connoître leurs ressources, mais pour grossir la masse de leurs impôts. Ce souvenir reste encore, et rend soupçonneux et défiant.

Pendant quatre mois consécutifs, j'ai parcouru le département de l'Indre ; j'ai visité toutes ses communes, tous ses établissements, toutes ses routes : par-tout j'ai vu, j'ai interrogé, j'ai recueilli.

J'ai rassemblé dans des cadres, et j'ai mis à la portée de tous les maires, toutes les questions relatives à la Statistique. J'ai reçu toutes leurs réponses ; je les ai toutes comparées entre elles ; j'ai comparé aussi les renseignements qui m'ont été fournis, avec ceux que j'ai acquis moi-mêmes sur les lieux. Ainsi j'ai, pour ainsi dire, confronté la vérité avec elle-même ; et les résultats que je vais présenter, s'ils n'ont pas une précision mathématique auront au moins toute celle dont un travail de ce genre est susceptible

*Mémoire statistique du département de l'Indre*, imprimerie de la république, an XII, page 1

Le premier constat est que la visée statistique mêle intimement une intention pratique et une volonté de connaissance. Il s'agit de savoir certes, mais pour agir, et pour que l'Etat agisse. Cette volonté se heurte cependant à de nombreux obstacles, et d'abord à des obstacles pratiques. L'appareil d'Etat ne dispose pas encore des ramifications lui permettant une collecte efficace et uniforme de l'information, c'est un homme seul qui mène cette entreprise, qui n'est pas sans susciter la méfiance.

De plus la mise au point de procédures uniformes se heurte à l'extrême diversité des territoires et des pratiques, à la difficulté aussi des mises en équivalence - l'addition en effet suppose l'équivalence des termes - du fait de la multiplicité des mesures et des codifications. Comparer la fréquence des crimes selon les lieux suppose ainsi que l'on appelle crime d'un lieu à l'autre des évènements comparables.

### 3.1.3 Le temps de l'adunation

En ce sens l'usage systématique de la mesure, n'est pas seulement analysable en terme de progrès intellectuels et de victoire d'une arithmétique politique anglaise plus efficace et rationnelle, mais aussi comme transformation d'une pratique intellectuelle qui a sa cohérence dans le contexte, et dont la statistique contemporaine hérite, tant l'objectif de classement systématique que la technique du tableau. Le triomphe de ce mode d'investigation du social nouveau, faisant appel au chiffre est permis par vaste entreprise d'homogénéisation des espaces nationaux. En ce domaine la période révolutionnaire et impériale joue en France un rôle clé.

Une vaste entreprise de mise en équivalence des phénomènes locaux fait partie intégrante du projet révolutionnaire, qu'il s'agisse des mesures physiques, des jugements, et ces processus s'accompagnent d'une modernisation de l'appareil d'état qui est aussi renforcement de celui-ci. En somme c'est alors que sont créées les conditions de possibilités d'une statistique publique moderne.

Je vous engage à retenir plusieurs idées de cette histoire. Le produit de l'effort statistique est susceptible, comme toute source d'une lecture qui interroge conditions de sa production. La Statistique napoléonienne donne parfois des chiffres mais dont précision très loin des normes contemporaines. Et l'appréciation de ce que l'on peut dire à partir des chiffres produit suppose de savoir quelque chose de la façon dont ils furent produits.

L'Effort statistique est d'autre part liée à une visée, il a une utilité pour l'état ou le souverain telle que définie alors, plus généralement pour le producteur de ces données. En ce sens pas seulement le chiffre qui fait sens mais aussi l'organisation même de la statistique. Prenons un exemple dans la statistique

des décès telle qu'elle est pratiquée en France au dix-neuvième siècle.

Les morts violentes, volontaires ou accidentelles, ont été

	<i>Mâles</i>	<i>Femelles</i>	
1819	449	156	605
1829	489	228	717
1821	480	177	657
<i>Total général</i>		1979	

dont 45 écrasés, 10 assassinés et 7 suppliciés

Les suicides se sont montés dans le département, pour les mêmes années,

	<i>Mâles</i>	<i>Femelles</i>	
1819	250	126	376
1829	211	114	325
1821	236	122	348
<i>Total général</i>		1979	

Le genre de mort le plus commun a été la submersion, le plus rare le poison, la cause la plus fréquente les chagrins domestiques d'où naît le dégout de la vie (349 sur 1040) ; le mois d'avril est celui de tous où les suicides sont le plus nombreux, comme il est aussi, avec le mois de mars, celui où l'on compte le plus de décès. Il eût été à désirer que les tableaux eussent indiqué dans une note le nombre de suicide particulier à la ville de Paris

*Extraits des recherches statistiques sur la ville de Paris et le département de la Seine ; recueil de tableaux dressés et réunis d'après les ordres de M. Le comte de Chabrol, préfet du département etc. tirés du bulletin universel des sciences et de l'industrie publié sous la direction de M. le baron e Féribusac, Paris, 1824, page 8.*

Un lecteur contemporain est bien embarrassé de ce document, du moins aura-t-il bien du mal à y prélever des informations chiffrées qu'il pourra présenter sous forme d'un graphique à l'appui de son discours d'historien. Sont en cause à la fois la précision des données, et les catégories utilisées. Le sous-enregistrement des suicides est en effet notoire tout au long du dix-neuvième siècle, en raison de l'interdit religieux pesant sur ce geste. Quant aux causes, à la cause plutôt, mise en avant par les auteurs de ce document, elle ne

correspond guère aux catégories modernes de l'analyse sociologique des suicides, qui ne distribueraient pas la population en fonction des motivations de l'acte, mais des caractéristique des individus (âge, sexe, situation de famille, statut social, pratique religieuse, etc...). La donnée statistique est d'abord une construction sociale complexe qui parfois, c'est le cas ici, nous en dit plus sur les catégories de perception des acteurs de l'époque que sur ce que nous aimerais savoir. Avant de la manipuler et de la mobiliser, il faut donc la comprendre.

## 3.2 La naissance de l'institution statistique

### 3.2.1 La naissance de la SGF

Reprendons pour quelques instants le fil de notre récit. L'entreprise napoleonnaise avait buté tant sur conditions concrètes de production des données que sur difficultés de la mise en équivalence. Il faut attendre les années 1830 pour que naissent les institutions dont sont directement issues nos institutions statistiques.

Une Statistique générale de la France émerge en 1833 dont le premier directeur est Moreau de Jonnès. Celui-ci est un partisan d'une statistique qui compte mais guère d'une statistique qui calcule. Il est en particulier méfiant envers les moyennes statistiques popularisée à l'époque par un belge (Quetelet), auquel on reproche d'additionner des choses différents.

Son administration est un petit service, rattaché au ministère du commerce qui rassemble et publie des données produites par d'autres administrations, en particulier les résultats du recensement périodique de la population organisé par le ministère de l'intérieur. C'est là l'essentiel durant cette période de l'effort statistique d'état. Le recensement est exhaustif, car les techniques de sondages ne sont pas au point, et l'hétérogénéité du territoire encore prononcée ne permet pas la constitution d'un échantillon de commune), ppl source d'info sur démographie et distribution géo de la population.

Il faut ajouter à cela la statistique sur le mouvement de la population (naissance mariage décès) et des données issues de l'activité de l'administration (les statistiques criminelles par exemple qui sont un sous produit de l'activité des tribunaux). De fait longtemps les données produites par l'état sont d'abord des sous produits de son activité, des comptages de ses actes.

Peu de changements se produisent durant un demi siècle, sinon un enrichissement des questionnaires des recensements (profession et nationalité en 1851). En 1890 SGF est encore un très petit service dont l'utilité est assez régulière.

lièrement contestée malgré le soutien d'un petit milieu de spécialistes et de réformateurs sociaux souvent associés à la société de statistique de Paris créée en 1860

### **3.2.2 Le moment de Lucien March**

Le tournant années 90 marquées par plusieurs ruptures

- - d'institution d'abord, avec la naissance au sein du ministère du commerce d'un office du travail auquel la SGF est rattachée
- - techniques ensuite avec l'exploitation mécanographique du recensement de 1890 mise au point par un jeune ingénieur Lucien March qui importe et améliore des machines américaines
- - fonctionnelle enfin ; La Sgf se saisit questions nouvelles, travail et main d'œuvre en particulier. De nouvelles questions sont ajoutées dans les formulaires de recensement (industrie et employeur). Cela permet, couplée au capacité de traitement accrue fournie par les machines de vastes enquêtes dans un contexte crise et de naissance du chômage. La SGF durant l'entre-deux-guerres emploie une centaine de personnes et gravitent autour d'elles des universitaires (Simiand et Halbwachs en particulier), des politiques et des scientifiques. C'est Encore modeste au regard de ce que deviendra l'appareil statistique français après 1945 mais c'est un changement sensible.

## **3.3 L'industrie statistique**

### **3.3.1 Naissance de l'Insee**

L'après 1945 est marqué par un brutal changement d'échelle que symbolise la naissance de l'INSEE (1946), corps spécialisé, relativement autonome et nombreux (on parle en milliers de fonctionnaires et non plus en centaines). Cela se traduit par la multiplication des enquêtes et des données bien sûr, mais aussi par l'emploi de techniques nouvelles.

### **3.3.2 Méthodes nouvelles**

On pratiques désormais systématiquement l'enquête par sondage, et les indices synthétiques se multiplient (taux de chômage par ex, pas systématiquement calculé et commenté avant les années cinquante).

### 3.3.3 Interprétation d'un changement

La création de l'insee est contemporaine de l'émergence d'un état nouveau, qui est à la fois régulateur des risques sociaux (naissance de la sécurité sociale), acteur central de l'économie (nationalisations), coordinateur de l'effort national (planification), et régulateur de la conjoncture (politiques keynesiennes).

Cet appareil nouveau est à la fois destiné à fournir les moyens de l'intervention de l'état, et possible parce que ce rôle nouveau suppose une activité de codification et des actes administratifs (on ne compte pas les chômeurs mais les demandeurs d'emploi bénéficiaire d'une aide d'état, quand ils ne sont pas aidés il n'y a ni besoin, ni moyen de les compter). Si ses formes ont changées, la statistique publique demeure le moyen et le reflet de l'activité de l'état, et aujourd'hui comme hier son interprétation suppose la conscience de ce que l'utiliser revient à emprunter les lunettes des contemporains, dont les catégories de perception ne sont pas toujours les nôtres, de s'interroger sur les conditions de sa réalisation qui en font souvent un sous produit de l'activité des administrations, structurées par ses catégories et son activité. Les statistiques criminelles ainsi sont notoirement difficiles à utiliser parce qu'elles sont à la fois le reflet des variations de la criminalité et de celles de l'activité des services de police et tribunaux. La donnée statistique donc se lit, se comprend, s'interprète, en un mot appelle la critique de l'historien avant toute manipulation ou utilisation.



# Chapitre 4

## Voir c'est penser

Les données que l'on peut extraire des sources ou de la documentation existante sont souvent re-élaborées avant d'être intégrées dans le discours de l'historien, au moyen de traitement statistiques parfois, nous y reviendrons, à l'aide aussi de dispositifs permettant une visualisation des données. La présentation graphique des résultats, ou des matériaux utilisés est de plus en plus fréquente. Nous ferons aujourd'hui un point rapide sur les possibilités offertes par les outils existant, sans oublier que l'utilisation pertinente de nombre de ceux-ci suppose que l'on ait une conscience minimale des transformations qui sont opérées et de leurs limites, donc qu'avant de se saisir d'un outil on prenne le temps de comprendre et d'évaluer celui-ci, sans oublier les principes de bases qui définissent l'usage rigoureux des dispositifs graphiques. Pour ceux que ce domaine intéresse ou qui en auront par nécessité l'usage, je ne peux recommander en français que l'ouvrage pourtant ancien de Jacques Bertin, critiqué aujourd'hui en certains de ses aspects, mais qui demeure la seule introduction synthétique à ce domaine accessible [?]. Visualiser : tracer et dessiner la donnée. Vous trouverez cependant de nombreux éléments de cours sur internet, du plus simple au plus compliqué, par exemple à [http://www.ummtodz/IMG/pdf/Representation\\_Graphique\\_des\\_donnees\\_cle4d85d6.pdf](http://www.ummtodz/IMG/pdf/Representation_Graphique_des_donnees_cle4d85d6.pdf) pour une présentation des graphiques usuels.

Le domaine est en pleine expansion car l'existence sous forme numérique de données accumulées en cours de recherche offre la possibilité de traitements permettant de visualiser, assez facilement tout ou partie de celles-ci.

Cette possibilité est surtout utilisée de nos jours par les chercheurs

ayant recours à la quantification et se fait donc souvent dans le cadre de l'usage de logiciels dédiés, qui incorporent un des possibilités graphiques, c'est le cas par exemple du logiciel R, actuellement le plus utilisé par les chercheurs ayant recours à ces techniques parce qu'il offre, par le biais d'une multitude de bibliothèques, l'accès à de nombreuses possibilités. Si l'outil est puissant, son apprentissage exige une compétence et un investissement en temps que ne justifient pas nécessairement des besoins ponctuels (hormis le cas d'étudiants possédant une solide culture statistique et habitués l'usage de l'informatique). Si le besoin est ponctuel et les graphiques à réaliser simples, il est possible d'utiliser un tableau.

Il est cependant fréquent que l'on ait dans le cadre de recherches qui ne font pas une large place à la quantification, quelques besoin de manipuler ou de produire quelques tableaux, quelques cartes, ou de réaliser quelques schémas. Ceux-ci peuvent dans ce cadre avoir des fonctions très variées, il s'agit cependant la plupart du temps soit d'explorer une information de façon à susciter des hypothèses, soit de produire un document permettant de résumer une information abondante et complexe soit de produire un document illustratif destiné à servir de support à une présentation orale du travail en cours (ou achevé).

Il est aujourd'hui souvent possible d'utiliser de petits logiciels (dont un nombre croissant ont une interface web), peu coûteux ou gratuits dont la prise en main est simple permettant d'obtenir des sorties graphiques réutilisables dans le cadre d'un mémoire ou d'une présentation.

La plupart de ces outils, rançon de leur simplicité d'usage, n'accomplissent chacun qu'une tâche très limitée et ne permettent de traiter que certains types de données. Certains (il en existe bien d'autres et sans doute parfois de plus efficaces) sont signalés ci-dessous, rangés en fonction du type de matériau qu'ils permettent de traiter.

## 4.1 Voir ses textes

**Les tagsclouds** Ce sont des briques logicielles qui permettent de représenter un texte sous la forme d'un nuage de mots, généralement la taille des mots (attention ce sont en fait des chaînes de caractères), est en proportion de leur fréquence dans le texte. L'un des plus utilisés est Wordle que vous trouverez à <http://www.wordle.net/>. Vous pouvez

dans une certaine mesure paramétrier les sorties. Attention cependant si les documents obtenus sont souvent jolis, ils sont plus destinés à illustrer un propos ou un mémoire, ou à susciter des questions qu'à permettre l'administration de la preuve.

**Voir la structure d'un texte ?** Certains chercheurs, partant du même souci de permettre une première visualisation de textes étendus élaborent eux aussi des dispositifs permettant de synthétiser graphiquement de nombreuses informations sur ce texte, par exemple non seulement la fréquence des chaînes de caractères mais les liens entre elles. C'est la technique des arbres arborés, vous en trouverez un exemple à <http://www2.lirmm.fr/~gambette/treecloud/>. Deux indicateurs sont combinés, un comptage des occurrences et une mesure de la distance entre termes. L'efficacité heuristique et illustrative est indéniable, le dispositif donne littéralement à voir la structure d'un texte que l'on décrit. Il convient de garder une certaine prudence dans l'interprétation : la notion de distance ou de proximité entre des mots qui sont en fait des chaînes de caractère n'est pas simple.

## 4.2 Visualiser ses données

Les tableurs permettent de construire des représentations graphiques de tableaux de données. Leur prise en main n'est cependant pas toujours plus facile que celle d'un grapheur ou d'un logiciel dédié. On trouve aujourd'hui en ligne des briques logicielles permettant de répondre à des besoins ponctuels et très spécifiques. Là encore les exemples proposés sont illustratifs et ne peuvent prétendre à l'exhaustivité. L'un des plus aboutis est le projet manyeyes <http://www-958.ibm.com/software/analytics/maneyes/>

**Données dans l'espace** Les données que traitent l'historien sont souvent situés dans l'espace et la dimension géographique des phénomènes étudiés est souvent pertinente (comme tout règle elle connaît des exceptions). De ce fait un mémoire d'histoire a souvent besoin d'une cartographie minimale, qu'il s'agisse de permettre au lecteur de localiser ce que l'on étudie ou de visualiser la distribution de données dans l'espace. Nous ne saurions trop recommander d'ailleurs, chaque fois que vous étudiez un phénomène qui se déploie dans l'espace de proposer à vos lecteurs une carte permettant la localisation des places mentionnées

par vous.

Il peut également être utile, même si cela ne donne pas lieu à la confection d'un document qui sera ensuite inséré dans le mémoire d'indexer à une position géographique les informations que l'on a accumulées ou d'explorer la distribution géographique de données.

Celles-ci peuvent être des données agrégées fournies par une agence statistique, ou bien confectionnées par le chercheur. Lorsque la dimension géographique est centrale pour l'étude entreprise, la solution passe par l'usage d'un outil dédié, un système d'information géographique. Ces logiciels cependant, qui couvrent à peu près toute la gamme des besoins sont très complexes et leur prise en main est difficile. Dans la majorité des cas elle ne se justifie pas pour l'historien qui utilisera, ponctuellement des modules logiciels aux fonctions moins puissantes mais à l'usage plus aisés. Là encore la liste n'a qu'une simple valeur d'exemple, vous incitant à prendre connaissance de ressources existantes qui peuvent toutes être utiles mais ne le sont pas pour tous.

- Atlas interactifs

C'est surtout utile pour les périodes récentes, mais les grands instituts statistiques nationaux et européens ont mis en ligne sous forme d'atlas interactifs une partie des informations dont ils disposent, et offrant parfois des interfaces permettant de contrôler le choix des variables, leur discrétisation, ainsi que dans certains cas les échelles utilisées.

- L'Insee à <http://www.insee.fr/fr/bases-de-donnees/>
- Eurostat à <http://epp.eurostat.ec.europa.eu/tgm>
- D'autres instituts nationaux, par exemple, <http://nationalatlas.gov/mapmaker>
- Cartes de localisation

Il est parfois difficile pour l'historien de trouver des cartes de localisation adaptés à ses besoins, la plupart du temps il lui faut produire un document à partir d'une carte moderne. Il existe aujourd'hui de nombreux dispositifs permettant d'ajouter du contenu sur des fonds physiques ou toponymiques existants, parmi eux, là encore à titre d'exemple [+http://www.click2map.com/](http://www.click2map.com/)

- Dessiner une carte

Les outils usuels de dessin sont tout à fait adaptés à la production d'une carte, il peut, en particulier si la précision géographique est une nécessité, être utile d'utiliser un outil dédié. Il existe là en-

core des outils en ligne, permettant de piloter une application logicielle depuis l'interface d'un navigateur, parmi elle Magic Map, que vous trouvez à <http://www.natureonthemap.naturalengland.org.uk/MagicMap.aspx>, et qui permet, entre autres fonctions, de choisir le système de projection que l'on souhaite utiliser.

- Cartographie thématique

Il est assez fréquent de disposer de données décrivant la distribution d'une variable pour une liste d'unités administratives (taux de scolarisation par département, vote par canton, pauvres par paroisses etc ...). On peut souhaiter alors générer une carte représentant la distribution des valeurs de la variable selon l'unité utilisée. Les solutions logicielles permettant de le faire à moindre coût sont paradoxalement moins nombreuses qu'il y a quelques années, l'une des meilleures est aujourd'hui Cartes et donées.

### 4.3 Visualiser sa réflexion

L'usage d'outils logiciels permettant de visualiser une argumentation ou un ensemble d'information s'est depuis peu de temps beaucoup développé, en particulier dans le monde de l'enseignement secondaire, où ces outils connaissent une véritable vogue. Les outils dédiés à cela sont généralement regroupé sous l'étiquette du mindmapping ou encore cartes heuristiques. Si l'efficacité heuristique de ce type de techniques est fort anciennement éprouvée, il n'est pas toujours certain que l'usage de la machine soit ici toujours plus efficace que celui du papier et du crayon. La principale justification du passage par l'informatique est ici la possibilité qu'elle offre d'un partage et d'un travail collaboratif et la qualité formelle des sorties.

Vous trouverez une liste d'outils gratuits à , <http://socialcompare.com/fr/comparison/>

### 4.4 Donner à voir

La généralisation de l'usage de l'informatique et la diminution du coût des matériels de vidéo-projection ont conduit à imposer dans le monde scientifique comme quasi-norme la production de documents d'accompagnements pouvant être visualisés sur écran lors d'une présentation orale. Dans la majorité des cas cela se traduit par la pro-

duction de diapositives sur lesquelles s'affichent à mesure le texte dit par l'orateur, ce qui permet à l'assistance de finir de répondre à son courrier puisqu'il sera possible de récupérer le texte intégral de la communication.

Si ce type de dispositif peut être adapté lorsque l'on s'adresse à un public, non francophone, maîtrisant mal l'écrit, ou doté de faibles capacités de concentration ou de peu de compétences cognitives, il a peu d'intérêt lorsqu'il s'agit de présenter les résultats d'une recherche.

Il est alors souvent plus intéressant de pouvoir présenter des objets graphiques que l'on commentera ou qui illustreront le propos d'autant que les outils permettant une mise en image élégante sont aujourd'hui nombreux et divers. Les présentations sous forme de diaporama sont les plus courantes, souvent produites à partir de logiciels commerciaux. Des solutions libres existent cependant (beamer par exemple), mais il est possible aussi d'utiliser n'importe quel outil logiciel avec une fonction diaporama.

Se développent également des outils permettant des présentations plus élaborées, qui permettent de générer des documents complexes, souvent séduisants (prezi est le plus connu, que vous trouverez à <http://prezi.com/index/9/>), qui permettent une navigation plus aisée (effets de zoom, de profondeur). Le temps de préparation est cependant plus long, et pas toujours justifié s'il s'agit d'une présentation unique.

# **Chapitre 5**

## **Comptages d'historiens**

Pour une première approche (en français et gratuitement disponible sur internt) [?], Pour une initiation aux outils statistiques destinée aux historiens on verra [?], pour un panorama des méthodes statistiques employées par les historiens [?]

L'historien n'est pas seulement un utilisateur de données statistiques, il peut aussi produire ses propres données à partir du matériau qu'il accumule, parfois sous la forme de simples comptages, permettant de résumer une information abondante, parfois sous des formes beaucoup plus élaborées quand son matériau ou les questions qu'il se pose l'exige.

### **5.1 Que fait l'historien qui compte**

#### **5.1.1 La mise en série, définir des objets et des propriétés**

Pour compter l'historien, comme tout autre opérateur d'ailleurs a besoin, avant de connaissances mathématiques ou de compétences informatiques, de notions, d'une réflexion sur les catégories qu'il utilise. J'ai besoin donc de définitions qui définissent pour les besoins de l'analyse des objets considérés équivalents. Je ne peux pas mesurer le chômage, si je n'ai pas une définition du chômeur, et si je ne considère pas que les chômeurs, qui sont tous différents de multiples manières (l'âge, le métier, le sexe), ne sont pas, sous un certain rapport, identiques ou équivalents.

Donc le travail de l'historien qui compte c'est d'abord un travail de définition.

C'est ensuite un travail de codification ou d'identification. Il faut définir des critères qui conduiront à ranger ou pas dans une catégorie, qui est une catégorie abstraite, un phénomène ou un individu singulier à partir des informations disponibles. Vouloir compter les bourgeois parisiens du dix-neuvième et étudier leur répartition dans l'espace est ainsi un projet très complexe, parce qu'il faut non seulement définir ce qu'est un bourgeois parisien, mais encore adopter des conventions quant aux caractères qui permettent de les reconnaître dans les sources. Ce double processus implique toujours des incertitudes et une perte d'information, parce qu'elle consiste à simplifier l'information disponible afin d'en favoriser la synthèse. Quels que soient les critères choisis, il existera forcément des cas limites, que l'on peut ou non décider d'inclure.

### **5.1.2 Hypothèses explicatives**

Et mes bourgeois parisiens je ne vais pas vouloir seulement les compter et les identifier, je vais vouloir les décrire, éventuellement montrer que leurs comportements changent dans le temps, je vais vouloir montrer qu'il y en a plusieurs sortes.

Dans une situation de ce genre, j'ai abondance d'information sur chaque cas, je ne vais pas pouvoir tout conserver. Je vais définir des indicateurs, que je considère pertinents, à partir d'hypothèses ou de ce que je sais de la population que j'étudie. Je vais opérer des choix, dont il faudra d'ailleurs que je rende compte dans le texte qui va rendre compte de mon travail, et je ne peux faire ceux-ci, et les justifier qu'à partir d'une connaissance préalable de mon objet et de mon matériau.

### **5.1.3 Opérations mathématiques**

Ensuite je vais compter, et puis parfois me demander si il y a un lien entre deux indicateurs, la probabilité du succès d'une grève par exemple et la conjoncture, s'il y a assez de différences entre deux sous groupes, deux ensemble de tessons par exemple trouvés dans une fouille, pour que je considère qu'ils proviennent de deux lieux différents. Je vais en somme mettre en oeuvre des procédures statistiques dont il faudra ensuite interpréter les résultats en faisant appel aux autres informations disponibles. Si les procédures statistiques me montrent que la composition des deux amas de tesson est différente,

qu'est ce que je sais des conditions de fouilles, de l'usage des lieux, des ateliers en activité à l'époque.

## 5.2 Comment pouvons le lire

L'Historien qui produit des données, même d'assez simples comptages se livre en fait au même type d'opérations que les spécialistes de la statistique publique évoqués dans la section précédente, donc son travail est susceptible d'une lecture de même type, qui est une lecture critique, et qui sont autant de questions qu'il convient d'adresser à ses propres pratiques.

### 5.2.1 La critique par les outils

Les procédures statistiques sont des outils souvent conçus afin de répondre à des problèmes adhoc et qui de ce fait ne peuvent être utilisés de façon efficace que dans des conditions assez précisément spécifiées. Il faut donc s'assurer que les outils utilisés peuvent l'être dans le contexte.

### 5.2.2 Critique par les sources

La validité des résultats et des conclusions dépend dans une large mesure de la qualité des infos utilisées. Ce n'est pas différent de ce qui se passe dans le cas d'une étude classique, mais parfois oublié devant les tableaux de sources. Jean-Luc Pinol, utilisant les listes électorales de la région lyonnaise l'exprime clairement.

« Les limites des listes électorales sont évidentes : anti-jeunes, misogynes, xénophobes (...) Ces liste de surcroît ne concernent pas tous les Français adultes et de sexe masculin ; « certains sont rayés des listes par décision de justice (condamnés faillis), d'autres tardent à transférer leur droit électoral après changement de résidence ... et si la révision annuelle élimine les électeurs décédés, elle n'enregistre pas forcément les départs »

Jean Luc Pinol, *Espace social et espace politique. Lyon à l'époque du front populaire*, PUL, 1980, page 8

### **5.2.3 La critique par les objets**

Il est enfin souvent essentiel pour apprécier un texte s'interroger sur la manière dont sont construites variables et catégories.

## **5.3 Les outils disponibles**

Les dispositifs logiciels nécessaires à la réalisation d'opérations statistiques sont très nombreux, mais les historiens n'en utilisent qu'un petit nombre. Le choix en fait se résume à deux questions, la première étant faut-il ou non utiliser un logiciel de base de données afin de gérer le matériau utilisé. La réponse dépend à la fois de la complexité (plus que de la taille) du matériau, de l'importance de l'élément statistique dans votre propos (l'exploration statistique est elle centrale, ou bien au contraire marginale), ainsi que de votre plus ou moins grande appétence pour les outils mathématiques.

Pour un usage ponctuel, ou l'exploitation de petits volumes de données, l'usage d'un tableur multifonction (Excel ou le tableur de la suite NeoOffice ou OpenOffice est suffisant. Si vos données sont complexes (des dossiers de procédures judiciaires par exemple, ou des dossiers de carrière), il faudra structurer l'information au moyen d'une base de données, et donc s'approprier un logiciel ou un langage adapté.

Si vous envisagez des traitements statistiques complexes (analyse multivariées), il est souhaitable d'utiliser un logiciel adapté. Le plus généralement utilisé est R. (pour une première approche en français on verra [?]. Il a l'avantage d'être gratuit, très puissant, mais suppose un apprentissage qui ne se fait pas en quelques heures.

# Chapitre 6

## Éléments de statistique descriptive

Une fois l'outil choisi, il reste à déterminer les traitements à appliquer. Dans les pages qui suivent nous allons présenter de manière un peu détaillée quelques notions de base, indispensables à tous, avant d'évoquer des dispositifs plus élaborés, dont nous signalerons simplement l'existence et les conditions d'usage, renvoyant à une littérature spécialisée pour d'éventuels approfondissements.

### 6.1 Résumer une variable

#### 6.1.1 Un peu de vocabulaire

L'ensemble des objets que vous étudiez se nomme une population, chaque élément de celle-ci est un individu, chaque individu est décrit par un certains nombre de variables. Celles-ci peuvent être de plusieurs types. La distinction essentielle pour nous oppose les variables quantitatives, qui peuvent être exprimées par une valeur numérique (l'âge par exemple), et les variables qualitatives (l'orientation politique, la catégorie socio-professionnelle).

Une variable qualitative se décrit en donnant le nombre de réalisations de chacune des valeurs possibles, que l'on rapporte au nombre total d'individus. Ce rapport est appelé la fréquence.

Exemple. Soit un sondage, trois réponses à une question sont possibles : oui (rencontré 3 fois), non (rencontré 2 fois), sans opinion (rencontré une fois). La variable est décrite en associant à chaque valeur possible sa fréquence [(oui ; 3/6=0,5) ; (non ; 2/6=0,33), (abs-

tention ;  $1/6=0,17$ ], ce que l'on présente généralement sous la forme d'un tableau.

### 6.1.2 variables quantitatives

**Valeurs centrales** Les valeurs centrales sont des indicateurs permettant de résumer, de manière synthétique des informations relatives à une série de nombre et plus précisément de déterminer les valeurs autour desquelles celle-ci s'organise. La plus connue est la Moyenne arithmétique : Somme des valeurs/nombre des valeurs  
Si nous posons  $X_i$ , valeur  $i$ ème de  $X$  (la série de nombres dont nous cherchons a moyenne),  $N$  nombre de ceux-ci, nous avons

$$M = \frac{\sum_{i=1}^n X_i}{N} = \bar{X} \quad (6.1)$$

La signification de la moyenne n'est pas d'évidence, la pertinence de son usage dans le contexte d'une analyse pas plus. Un détour, sur les pas d'Alain Desrosières, par l'histoire de son usage permet de le comprendre.

«L'idée de moyenne agglutine deux idées. Historiquement, elle correspond à deux contextes. Le premier est celui des mesures d'astronomie et de physique. Pour connaître la hauteur d'une étoile dans le ciel, l'astronome du XVIII<sup>e</sup> siècle jugeant ses instruments imparfaits, fait plusieurs fois de suite la même mesure, trouve des résultats différents et calcule la moyenne. S'il a réalisé 100 mesures, il les additionne puis divise le tout par 100. Mais dans ce cas là, personne ne doute que l'étoile existe vraiment : il y a une vraie hauteur, un réalité. Cette réalité, on l'atteint du mieux possible en calculant une moyenne. Dans un autre contexte, lorsqu'on dispose d'objets différents les uns des autres, on peut souhaiter les remplacer par un seul objet, fictif cette-ci et représentant des caractéristiques moyennes [...]. Quêtelet, créateur de la statistique au sens moderne, produit un raisonnement justifiant le passage d'un type de moyenne à l'autre. Selon son idée, il était justifié de calculer une moyenne si la distribution des objets correspondait à la fameuse courbe en "chapeau de gendarme" ou courbe de Gauss. Quêtelet fit le raisonnement suivant : pour 100 mesures d'un même objet une étoile, il y a 100 nombres différents. Ils ne sont pas répartis au hasard. Leur histogramme a la forme de cette fameuse courbe de Gauss dont la loi résulte de l'addition d'un grand nombre de petites causes indépendants les unes des autres. Dans ce cas là, on peut montrer que la meilleure approximation de la hauteur de l'étoile c'est la moyenne : c'est parmi tous les nombres, celui qui est tel que la somme des carrés des écarts entre chacun des nombres et ce nombre là est le plus petit possible. Il passe ensuite à l'autre cas de figure : un grand nombre d'objets, intrinsèquement différents cette fois-ci, par exemple 100 conscrits d'un régiment. Ces conscrits sont tous différents les uns des autres. Mais lorsqu'on les mesure, on s'aperçoit que leurs tailles sont distribuées selon une courbe en chapeau de gendarme. D'où son hypothèse : de même que les mesures de la hauteur de l'étoile sont des approximations de la hauteur vraie, chaque homme est une approximation d'un "homme moyen", qui a une réalité supérieure. Quêtelet a donc inventé une sorte d'objet, qui fictif au début, devient réel sous sa plume [...]. Ce raisonnement est d'une grande importance. Il justifie en effet l'usage des grands nombres transformés par une moyenne.»

Alain Desrosières, *L'État et ses nombres*

La moyenne est presque toujours calculable

Elle n'a pas toujours de sens

Elle est de plus sensible aux valeurs extrêmes, d'où la nécessité souvent de compléter son emploi par celui d'autres indicateurs, parmi lesquels on trouvera :

Le mode Le mode est la valeur la plus fréquemment représentée dans la série

La médiane La médiane partage en deux sous ensembles, S1, et S2, de même cardinal l'ensemble des valeurs de la série. Toutes les valeurs de S1 sont inférieures à la médiane, toutes les valeurs de S2 sont supérieures à la médiane.

Beaucoup des indicateurs usuels dérivent de la moyenne

**Indicateurs de dispersion** Lorsque l'on examine des valeurs distribuée selon une gaussienne ou une quasi-gaussienne, le calcul des indicateurs centraux permet de déterminer les valeurs autour desquelles s'organise la série, mais ne dit rien de la façon dont elles sont distribuées autour de celles-ci.

Exemple : deux élèves ont obtenu les notes suivantes en mathématiques au second trimestre A=(1,19,18,2) et B=(10,9,11,10).

Les deux distributions ont une même moyenne et ni leur mode ni leur médiane ne permettent de les distinguer, elles ont peu de chance cependant d'être interprétées de même manière par les utilisateurs. En d'autres termes elles n'ont pas le même sens parce les valeurs ne sont pas distribuées de manière similaire autour des valeurs centrales ce que le calcul de paramètres de dispersion permet de préciser.

Variance et écart type

$$V = \frac{\sum_{i=1}^n (Xi - M)^2}{N} \quad (6.2)$$

La variance est donc une moyenne, celle du carré des écarts à la moyenne des valeurs de la série de nombres étudiée. Elle est directement fonction de la moyenne. On utilise généralement l'écart type plutôt que la variance avec

$$\sigma = \sqrt{V} \quad (6.3)$$

L'écart type permet d'évaluer la forme prise par la distribution des valeurs autour des valeurs centrales, dans le cas d'une distribution gaussienne il permet de la décrire avec précision.

Coefficient de variation Le coefficient de variation est utilisé parfois

afin de comparer la densité de deux distributions qui ne sont pas exprimées dans la même unité, ou ne ressortent pas du même ordre de grandeur. Avec

$$CV = \frac{\sigma}{M} \quad (6.4)$$

Dans la plupart des cas la description d'une variable unique n'a guère de sens ou d'intérêt. Il s'agit généralement de comparer les valeurs prises en différents lieux, en différents moments, ou pour différentes populations. Seul l'écart fait sens, encore faut il qu'il soit assez important pour justifier une explication. Nul besoin (défaut fréquent dans les mémoires), d'épiloguer longuement sur des écarts minimes.

## 6.2 Lien entre deux variables

Il est rare que l'historien, ou quelque spécialiste de sciences sociales que ce soit, n'ait besoin de décrire qu'une seule série. Généralement il s'attache aux relations entre des séries de valeurs.. La forme la plus simple prise par cette configuration est la recherche d'une covariation de deux séries, ou variables. Les techniques utilisées pour les mettre en valeur ne sont pas de même nature dans le cas de variables quantitatives et de variables qualitatives.

### 6.2.1 Lien entre deux variables quantitatives

Dire que deux variables quantitatives sont corrélées revient à dire que les variables co-varient. Si nous prenons l'exemple de deux variables X1 et X2, alors X1 et X2 covarient si lorsque X1 prend des valeurs élevées, X2 prend des valeurs élevées et X1 prend des valeurs basses quand X2 prend des valeurs basses. La corrélation peut également constituer en une covariation inverse. En ce cas X1 prend des valeurs basses quand X2 prend des valeurs élevées et inversement. Le calcul d'un certain nombre d'indicateurs permet de se prononcer sur l'existence d'une covariation statistiquement significative, il ne donne pas d'indications sur le sens de celle-ci (il ne permet pas de décider si  $X1=f(X2)$  ou  $X2=f(X1)$ , non plus que sur sa signification. Certaines corrélations ne sont pas pertinentes (non interprétables) au regard de l'historien.

Nous devons l'une des plus jolies illustrations de cette impossibilité de conclure de la co-variation ou de la co-occurrence à la causalité à Pierre Bayle, écrivant : "Ainsi les témoignages des Historiens se

réduisent à prouver uniquement qu'il a paru des Comètes et qu'en suite il y a bien eu des désordres dans le monde; ce qui est bien éloigné de prouver que l'une de ces deux choses est la cause ou le pronostic de l'autre, à moins qu'on ne veuille qu'il soit permis à une femme qui ne met jamais la tête à sa fenêtre, à la rue Saint Honoré, sans voir passer des Carrosses, de s'imaginer qu'elle est la cause pourquoi ces Carrosses passent, ou du moins qu'elle doit être un présage à tout le quartier, en se montrant à sa fenêtre, qu'il passera bien tôt des Carrosses. (Pierre Bayle, *Pensées diverses sur la comète*, 1680).

En plus de faire sens, le lien entre les deux variables doit être statistiquement significatif, ce que l'on peut estimer à l'aide du coefficient de corrélation

**Coefficient de corrélation** Le calcul du coefficient de corrélation entre deux variables (X et Y ici), suppose au préalable le calcul de la covariance. Nous avons, avec les mêmes notation que plus haut

$$C(x, y) = \frac{\sum_{i=1}^n (Xi - \bar{x})(Yi - \bar{y})}{N} \quad (6.5)$$

Le coefficient de corrélation, noté r vaut lui

$$r(x, y) = \frac{C(x, y)}{\rho_x \rho_y} \quad (6.6)$$

Il correspond donc à la covariance rapportée au produit des deux écarts types. La logique est la même que pour le coefficient de variation, il s'agit de se débarrasser des écarts d'unité et d'ordre de grandeur afin de n'exprimer que l'intensité de la liaison entre les deux variables. Le coefficient de variation varie entre -1 et 1. Plus sa valeur absolue est proche de 1, plus la liaison constatée est statistiquement solide, si elle est de signe positif, les deux séries varient de conserve (prennent en même temps des valeurs élevées, en même temps des valeurs basses), si elle est de signe négatif il s'agit d'une co-variation inverse (lorsque l'une prend des valeurs élevées l'autre prend des valeurs basses).

**Droite de régression** Lorsqu'une corrélation linéaire existe entre deux variables, il est possible de représenter graphiquement celle-ci en traçant une droite, nommée droite de régression, qui ajuste au mieux le nuage de points constitués des couples de valeur (Xi, Yj)

Cette droite a pour équation :

$$y = \frac{C(x, y)}{\rho^2 x} (x - \bar{x}) + \bar{y} \quad (6.7)$$

Attention là encore, il est presque toujours possible de calculer l'équation d'une droite de régression, ce qui ne veut pas dire que cela ait forcément un intérêt.

### 6.2.2 Liens entre deux variables qualitatives

L'étude des associations entre deux variables qualitatives obéit à des logique un peu différentes. Il ne s'agit pas ici de savoir si deux valeurs varient conjointement mais si sont souvent associées certaines des propriétés des individus statistiques étudiés, si en d'autres termes existent des attractions fortes entre modalités de deux variables. Cette exploration est souvent menée de manière intuitive à partir de la lecture des tableaux de fréquences associés à la table des données donnant la distribution des deux variables. Ceux-ci sont produits par transformations, par le calcul et l'écriture depuis le tableau de données.

Celle-ci se présente sous la forme d'une matrice dont chaque ligne correspond à un individu et les colonnes aux variables.

<i>Gudule</i>	<i>F</i>	<i>latin</i>
<i>José</i>	<i>M</i>	<i>Paslatin</i>
<i>Frank</i>	<i>M</i>	<i>latin</i>
<i>Maria</i>	<i>F</i>	<i>latin</i>
....	...	...

Le tableau décrivant la distribution conjointe des deux variables XY est produit par dénombrement des ensembles d'individus possédant conjointement les modalités Xi de X et Yj de Y. Le cardinal de chacun des ensembles ainsi défini est reporté dans un tableau dont les lignes correspondent aux modalités possibles de X (resp Y) les colonnes aux modalités possibles de Y (resp X).

	<i>M</i>	<i>F</i>
<i>Latin</i>	<i>effectiflatinetM</i>	<i>effectiflatinetF</i>
<i>paslatin</i>	<i>effectiflatinetM</i>	<i>effectifpaslatinetF</i>

Il est possible à partir de cette distribution de définir trois tableaux de fréquence un tableau des destinées, un tableau des provenances

(recrutement dans le cas des tables de mobilité sociale), un tableau enfin donnant le poids de chaque groupe définir par chaque couple de modalité. Ceux ci sont calculés en divisant les valeurs du premier tableau par les totaux en ligne et en colonne du premier tableau, par l'effectif total pour le dernier. Ils offrent des informations non redondantes (on ne peut déduire un tableau de fréquence d'un autre connu).

La lecture intuitive de tableaux de ce type consiste à les comparer implicitement à un tableau théorique, celui que l'on obtiendrait si n'existe aucun lien entre les deux variables examinées. Ce tableau théorique peut être calculé à partir des marges de la table décrivant la co-distribution des deux variables. Cela conduit certains utilisateurs à calculer un tableau des écarts, qui porte dans chaque case la différence entre les effectifs du tableau observé et du tableau théorique.

Il est possible aussi d'utiliser des techniques de constructions graphiques permettant de visualiser les écarts à l'indépendance [?]

Si cela permet de repérer les écarts entre les plus importants, pour ce qui est des effectifs, cela ne permet pas de se prononcer sur la validité statistique de ceux-ci, sont-ils assez forts pour ne pas résulter du hasard ? Le recours à test statistique permet de s'en assurer. Deux sont utilisés pour ce type de situation, les tests du chi<sup>2</sup>, qui supposent de travailler sur des populations assez grandes (toutes les cases du tableau théorique doivent avoir des effectifs >5), ou bien pour des effectifs plus petits les tests de Fisher

# **Chapitre 7**

## **description de grands tableaux**

Dans bien des cas l'historien ne cherche pas à déterminer les liens entre deux variables, mais à explorer un tableau d'enquête de grande taille décrivant de nombreux individus (statistiques), au moyen de multiples critères, ou, ce qui revient un peu au même, à se prononcer sur les interactions entre de multiples facteurs au cours du temps. Il s'agit alors pour lui d'explorer les interactions entre les variables décrites par le tableau, de se prononcer sur l'existence au sein de sa population de types différents. La démarche est ici pour une bonne part une démarche exploratoire dont l'enjeu est de permettre à l'opérateur d'acquérir une connaissance plus fine de son jeu de données, plus que de se prononcer sur l'existence de relations causales. Plusieurs techniques statistiques permettent de répondre à ces objectifs.

### **7.1 analyse descriptive multivariée**

L'analyse factorielle, qui existe sous de multiples formes, le choix entre elles étant déterminé par la nature des variables étudiées, est une méthode d'analyse développée initialement par Jean Paul Benzecri et ses élèves[?] afin de permettre le dépouillement de volumineux tableaux de données sans introduire, préalablement à l'exploitation des données de modèle ou d'éléments d'interprétation. Si nous sommes moins certains aujourd'hui du fait que l'analyse factorielle nous permet d'atteindre l'ordre du réel - le choix des variables introduites dans l'analyse, les décisions de classification, qui incorporent une théorie de l'objet déterminent en effet en partie les résultats -

l'analyse factorielle demeure un outil très performant de dépouillement de grands tableaux, parce qu'elle permet un repérage systématique des liens entre variables et modalités de celles-ci, la mise en évidence de configurations fréquemment présentes dans le jeu de données et des proximités entre individus. Elle permet en fait une exploration fine du jeu de données tout en livrant des représentations graphiques permettant de mettre en évidence les structures organisant celui-ci.

Les AF peuvent être réalisées sous R, ou bien à l'aide des modules dédiés d'un outil disponible en ligne comme Analyse. <http://analyse.univ-paris1.fr/> Les manuels les plus clairs, les plus accessibles aussi aux historiens permettant la construction et l'interprétation de ces traitements demeurent ceux de Philippe Cibois (par exemple [?])

## 7.2 Modes de classification automatique et semi automatiques

### 7.2.1 Classifications post-factorielles

L'un des intérêts de l'Analyse factorielle est de permettre, au moyen de calculs à partir des résultats fournis par celle-ci, une division de la population étudiée en classes, ou en profils. Ce type de technique est très utilisé dans le monde de l'entreprise (marketing, suivi clientèle, études d'opinions) beaucoup moins par les spécialistes de SHS. Il est pourtant un outil très efficace de présentation des résultats (il revient en fait à regrouper des individus se rapprochant le plus possible d'idéaux-types) et à réduire la complexité du jeu de données en permettant de résumer par une variable nouvelle (l'appartenance aux classes distinguées) une partie de l'information fourni par celui-ci avant de recourir à de nouveaux traitements.

Recourir à cette technique n'est cependant pas sans risques. Plusieurs algorithmes coexistent (plusieurs modes de calcul), qui ne produisent pas toujours des résultats stables et convergents (les différents modes de calcul peuvent aboutir à proposer des regroupements différents, et le même calcul répété plusieurs fois ne produit pas toujours les mêmes regroupements d'individus).

### 7.2.2 Techniques neuronales

Les techniques de classification qui réutilisent les données produites par une analyse factorielle tendent à dissocier (c'est à dire à répartir dans des groupes différents), les individus qui ne se ressemblent pas. Elles ont en somme pour point commun et priorité première d'éviter les "mauvais" classements. Dans un certain nombre de cas, il peut être, pour les besoins de l'analyse (ou pour compléter et corriger les résultats proposés par les analyses post-factorielles) de se donner comme priorité de regrouper les individus (statistiques), qui se ressemblent le plus. L'une des techniques les plus efficaces est alors le recours aux algorithmes de Kohonen, dont les résultats sont représentés sous forme d'un maillage (les cartes de Kohonen). Ces méthodes, qui suppose le recours aux propriétés des réseaux neuronaux, sont cependant, pour qui ne maîtrise pas les fondements mathématiques de ces outils, difficiles à mettre en oeuvre et délicates à interpréter. De façon plus générale il est préférable d'avoir recours à des outils dont on connaît assez bien les propriétés, ou éventuellement d'y avoir recours avec l'aide de personnes compétentes pouvant valider les procédures utilisées et les résultats obtenus, sous peine de parvenir rapidement à des erreurs statistiques et des monstruosités interprétatives.

## 7.3 Régressions, analyse de la variance

Le pont commun entre les méthodes de classification présentées ici (il en existe beaucoup d'autres, généralement moins usitées et/ou plus complexes à mettre en oeuvre), est de ne pas poser a priori d'hypothèses sur les relations entre les variables et les formes de celles-ci (ces méthodes sont dites sans modèles). Elles sont donc pour l'essentiel des méthodes exploratoires destinées à révéler l'existence d'attractions entre variables, individus ou modalités qui ne sont pas immédiatement perceptibles du fait du volume et de la complexité des données manipulées. Elles sont de ce fait particulièrement bien adaptées aux pratiques de l'historien.

Il existe d'autres techniques, qui ont selon les cas une visée prédictive ou explicative qui permettent de vérifier (mesurer), l'ajustement des données étudiées à un ensemble de règles décrivant les liens entre variables. Rarement développées dans le contexte des recherches en

sciences sociales, elles sont cependant parfois utilisées par les spécialistes de ceux-ci et peuvent, en certains cas se révéler utiles. L'interprétation des résultats obtenus est cependant souvent particulièrement difficile. Je signale ici l'existence de quelques unes, les plus utilisées actuellement, ou bien les plus faciles d'emploi, et vous invite, si vous avez l'usage de ce type d'outils à vous reporter à une documentation spécialisée.

### **7.3.1 Régressions linéaires multiples**

### **7.3.2 Régressions logistiques**

### **7.3.3 L'arbre de décisions**

Un contexte d'usage particulier : une variable dichotomique (c'est à dire ne proposant que deux valeurs possibles) "à expliquer", un ensemble de variables entre lesquelles on n'arbitre pas a priori  
Un intérêt, la simplicité du dispositif et la lisibilité des résultats  
Des limites, outre un contexte d'usage très particulier, la stabilité des résultats peut-être dans certain cas modestes.

# Chapitre 8

## Objets textuels et leurs traitement

Le cas particulier des données textuelles. La référence la plus pratique (mais d'une lecture parfois ardue est ici [?], on peut aussi utiliser[?]

### 8.1 Le corpus et sa constitution

Les pratiques d'analyse textuelle se développent depuis quelques années, facilitées par la mise à disposition de grands amas de textes. Là encore, utiliser les techniques de ce type pour un document unique n'a pas d'intérêt, il s'agit toujours de mettre en évidence des écarts, des différences. La notion de corpus (un ensemble de textes appartenant à un même ensemble), est donc ici fondamentale. Le corpus doit présenter plusieurs caractéristiques.

1. Raisonné (je peux rendre compte de sa construction)
2. Homogène (des textes partageant des caractéristiques communes)
3. Segmentable : je peux distinguer au sein du corpus des sections possédant des propriétés distinctes (je peux les rattacher à différents locuteurs/scripteurs ou bien à différentes périodes )

### 8.2 Le mot et la chaîne de caractères

Historiquement l'analyse textuelle s'est développée en permettant le comptage des formes.

La forme : une chaîne de caractères entre deux séparateurs (ponctuations espaces)

vers la fenêtre je fis un vers j'allais aller ou j'irai  
 Les progrès de la puissance de calcul des machines, ont permis peu à peu de chercher à atteindre les mots, quelles que soient leurs variations de formes (toutes les formes possibles d'un verbe par exemple). Cela suppose une opération préalable qui se nomme la lematisation. De violents conflits ont opposés les spécialistes divisés longtemps quant à la pertinence de cette nouvelle approche. Les deux semblent considérées aujourd'hui plus complémentaires que concurrentes.

## 8.3 Les outils

Lexico3. Longtemps le plus utilisé, il garde ses partisans, car il est sans doute le plus facilement utilisable par des novices, moins puissant que ses successeurs, il n'est pas non plus toujours parfaitement stable

Coocs. Un prolongement de Lexico (utilise en entrée les sorties statistiques de Lexico 3), permet la représentation du texte sous forme d'un chaînage grâce à la notion de cooccurrence multiple. Sur de gros corpus le traitement peut être long voire échouer

Hyperbase

; Alceste. Un logiciel issu du monde scientifique (d'un laboratoire du CNRS), mais devenu un produit commercial

TXM le plus récent et de loin le plus puissant actuellement disponible

## 8.4 opérations

### 8.4.1 Définir le corpus

### 8.4.2 Préparer le document

Les logiciels d'analyse textuelle sont pour la plupart des dispositifs savants élaborés au sein de laboratoires dont l'écriture logicielle n'est pas l'activité première et qui ne disposent pas de ressources comparables à celles d'un grand éditeur commercial. Il s'ensuit que l'ergonomie et l'élégance ne sont pas toujours leurs qualités premières, non plus souvent que la stabilité, même si les plantages sont aujourd'hui plus rares qu'autrefois. Cela nécessite de sauvegarder fréquemment son travail et en particulier, dès qu'ils sont obtenus les résultats à partir desquels on veut mener l'analyse. De plus, la gestion des innombrables formats et modes d'encodage rencontrés dans le monde

informatique, ainsi que les contraintes propres à l'analyse textuelle impliquent une préparation du corpus qui permette au logiciel de l'accepter comme intrant, qui peut être un moment particulièrement frustrant. Le problème des majuscules

Formats et encodages

La segmentation du corpus

Langue et champs discursifs nous apparaissent aujourd’hui comme des systèmes de différentiation. Ce ne sont pas tant lorsque nous étudions l’ensemble des scripteurs intervenant dans une controverse ou évoquant un même thème les caractéristiques de l’ensemble du corpus qui font sens pour nous mais la façon dont chacun des locuteurs se différencie de tous les autres, créant par là une position. Il faut alors, pour appréhender ainsi un objet, pouvoir distinguer les uns des autres des discours dont nous considérons pourtant qu’ils appartiennent au même univers. Cela revient à découper au sein de l’ensemble étudié des segments que l’on pourra distinguer en opposant leurs particularités à celles des autres segments. Il est très souvent possible, et cela rend l’analyse plus riche, de procéder à plusieurs segmentations sur un même corpus.

Cela se fait dans la plupart des cas par un balisage des textes, c.a.d. l’introduction dans le document contenant l’ensemble des discours de signaux indiquant le début (parfois la fin) d’un nouveau segment. La difficulté provient alors souvent de ce que les systèmes de balisage acceptés par les différents logiciels ne sont pas toujours compatibles entre eux, ce qui oblige, si l’on veut disposer d’une fonctionnalité présente uniquement sur l’un de ceux-ci, à une nouvelle préparation du corpus.

Lemmatiser (ou pas)

Historiquement la statistique textuelle, pour des raisons qui tenaient à la mémoire disponible et à la puissance de calcul nécessaire a d’abord pris en compte non pas ce que spontanément nous appelons un mot, mais des formes (ou chaînes de caractère). A la fois source d’ambiguité et, selon les détracteurs de cette approche faisant disparaître le mot et la langue cette pratique a été vivement critiquée, non sans conserver des adeptes. Beaucoup de spécialistes d’analyse textuelle recommandent aujourd’hui de lemmatiser les corpus, c’est à dire de remplacer les formes de celui-ci par la forme canonique (ou

lemme) du terme, qui est alors décrite par plusieurs catégories grammaticales (genre, nombre, nature du mot, etc ...).

L'historien étant par la force des choses et sa propre ignorance, généralement tout à fait ignorant des théories linguistiques ne peut que constater qu'une simple analyse des formes lui permet souvent d'obtenir des indices utiles cependant que si la lemmatisation offre de plus riches possibilités.

Elle est cependant plus lourde à mettre en oeuvre, particulièrement lorsque l'on utilise de très grands corpus et/ou des corpus contenant encore de nombreuses scories (utilisation massive de produits d'OCR non corrigées par exemple). En d'autres circonstances et si l'analyse textuelle du corpus touche aux enjeux essentiels du travail de recherche (analyse de discours, évolution de faits de langue), sa pratique ne peut qu'être recommandée d'autant qu'existent aujourd'hui pour le français contemporain des logiciels de lemmatiation qui automatisent une bonne part d'un travail autrefois particulièrement long et fastidieux.

Il existe ainsi une bibliothèque du logiciel R, dévolue à l'analyse textuelle, qui permet d'appeler une fonction de lemmatisation [?].

Le logiciel treetagger, développé par l'université de Stuttgart est utilisé par beaucoup de praticiens. Il offre l'avantage, par le biais de fichiers de paramétrage de permettre de traiter des corpus en différentes langues européennes (une dizaine à ce jour, dont le français). L'usage de ces outils suppose cependant un apprentissage et les choix opérés entre différentes méthodes et outils ne sont pas neutres (il existe plusieurs méthodes de lemmatisation, les spécialistes discutant, parfois vivement, de l'intérêt ou de la supériorité de telle ou telle). L'usage de ces outils est utile aux étudiants engagés dans des projets de recherches pour lesquels l'analyse d'un corpus textuel est centrale et les questions linguistiques importantes.

### 8.4.3 Explorer le lexique

#### Le dictionnaire :

La liste des chaînes de caractères (si le corpus n'est pas lemmatisé), présentes dans le corpus avec pour chacune le nombre d'occurrences (c.a.d. le nombre d'apparitions de celle-ci au sein du corpus). Permet un repérage (à grands traits) des univers lexicaux présents, particulièrement si l'on s'attache aux termes signifiants.

Deux remarques : en l'absence de lemmatisation ou de groupements de formes les résultats obtenus doivent être maniés avec une grande

prudence (un verbe par exemple peut-être très présent mais difficile à repérer parce qu'apparaissant sous de multiples formes, une forme très présente peut l'être parce que renvoyant à plusieurs mots assez fréquents). D'où l'importance alors de revenir vers le contexte afin de vérifier à quoi correspondent les formes rencontrées dont la fréquence nous intrigue.

Il est souvent tentant de se contenter de commenter la liste des quelques formes les plus fréquentes. Or, lorsque les corpus sont assez grands, la distribution des formes au sein de ceux-ci suit une loi de Zipf, soit une loi de puissance inverse. Il s'en déduit que les formes les plus fréquentes, peu nombreuses, ne regroupent qu'une petite partie de l'information statistique délivrée par le document.

**La liste des contextes** La plupart des logiciels permettent en effet d'obtenir la liste des contextes d'apparition d'une forme. L'opération, outre qu'elle est souvent indispensable à l'interprétation constitue aussi une puissante technique de lecture, particulièrement dans le cas de très grands corpus (plusieurs millions de mots) permettant un repérage rapide des secteurs d'un corpus abritant un mot-clé.

**Tables d'occurrences** Il est généralement possible également d'obtenir une mesure, et une représentation graphique, de la distribution d'une forme ou d'un groupe de formes au sein d'un corpus segmenté, voire dans le cas de certains corpus (et de certains logiciels) ordonné selon un axe temporel (cas d'une revue, d'un journal, d'une collection de discours) au fil du corpus.

**Les cooccurrences** Les linguistes ont depuis longtemps remarqué qu'au sein d'un corpus ou d'un texte un certain nombre de mots sont très fréquemment associés, que d'autres au contraires se rencontrent peu ou pas ensemble. La notion de cooccurrence systématisé ce constat en mettant en évidence la probabilité que deux termes apparaissent souvent au sein d'un corpus à peu de distance l'un de l'autre. Les difficultés commencent quand il s'agit de s'entendre sur ce que "à peu de distance veut dire". Est-ce dans la même phrase, le même paragraphe, la même page ? Nous dirons, par analogie, que tout calcul de distance suppose l'établissement d'une métrique, or plusieurs sont

possibles et il n'existe pas à l'heure actuelle de raisons fortes permettant d'en choisir une plutôt qu'une autre. Les logiciels disponibles offrent de ce fait souvent la possibilité de paramétriser les analyses de co-occurrences et il est souvent utile de tester plusieurs réglages, non seulement le choix de l'unité pertinente peut ne pas être le même selon le texte étudié, mais encore, il arrive que différentes explorations délivrent des informations également pertinentes et non redondantes.

**Les spécificités** La recherche des spécificités revient à explorer le corpus, non en partant de son lexique, mais en partant de sa segmentation en parties. Il s'agit alors de se demander si chacune (ou toute combinaison possible des segments du texte), peut être caractérisée par un nombre particulièrement élevé d'apparition de certaines formes, ou à l'inverse leur rareté relative. On raisonne là à partir de la notion de fréquence et la validité statistique des observations est garantie par des textes qu'il est parfois possible de paramétriser.

#### 8.4.4 Voir la langue

Raisonnner uniquement à partir des chaînes de caractères, ou même des formes d'un même mot, revient à travailler un objet linguistique en faisant abstraction de la grammaire (choix des temps, des modes, des personnes), et de la syntaxe (longueur des phrases, ordre des propositions), dont nous savons pourtant qu'elles constituent des propriétés discriminantes des discours. La lemmatization, qui permet de référer chaque forme rencontrée à des catégories grammaticales permet de retrouver cette dimension et la possibilité, au moyen d'outils statistiques très proches de ceux utilisés pour l'analyse du lexique de différencier les segments d'un corpus en fonction de celles-ci.

#### 8.4.5 Traitements statistiques

**Analyses factorielles** Les praticiens de l'analyse textuelles utilisent très souvent les techniques de l'analyse factorielle en vue tant d'explorer le jeu de données que constitue le corpus, que de proposer une vue synthétique des écarts mis en lumières entre les segments de celui-ci. La plupart des logiciels offrent des modules d'analyse factorielle (généralement une Acp), dont les résultats peuvent ensuite être transmis si besoin à des logiciels statistiques qui en permettent divers retraitement. Les règles de construction et d'interprétation sont les

mêmes que pour toutes analyse factorielle, la particularité étant ici que l'enjeu essentiel de l'analyse est souvent de s'interroger sur les similarités/dissimilités des différentes parties du corpus.

**Lois de puissances inverses** Nous avons indiqué plus haut que la distribution des formes au sein d'un corpus textuel de grande taille obéissait à une loi de Zipf, soit une loi appartenant à la famille des lois de puissance, définies par l'équation ci-dessous

$$1/x^\alpha \quad (8.1)$$

Cette propriété outre qu'elle disqualifie les pratiques qui consistent à caractériser une collection d'énoncés ou de discours en ne prenant en compte que quelques termes très fréquents ouvre la possibilité de définir des stratégies de lecture de ce type de données attentives aux formes concrètes des distributions observées [?]



# Chapitre 9

## Sondages et échantillon

Nous terminerons ce cours par quelques mots consacrés aux techniques de sondage et d'échantillon, qui suscitent souvent des questions de la part des historiens.

### 9.1 Une pratique sociale

#### 9.1.1 Définitions

La technique du sondage revient à étudier une population non pas à partir de l'observation de chacun de ses membres, mais en constituant un sous groupe de taille réduite, dont on va examiner les propriétés. Ce sous groupe est appelé un échantillon. L'ensemble des techniques qui ont permis sa constitution est le plan de sondage

#### 9.1.2 Une pratique assez récente

Les techniques qui permettent de le faire avec un certaine précision sont, au regard de l'historien relativement récentes. Le débat sur la possibilité de connaître les propriétés d'un groupe à partir de l'observation d'une partie de ses membres anime les cercles de statisticiens à la fin du dix-neuvième siècle. L'accord sur les possibilités théoriques et sur les conditions à réunir est atteint dans années 1920. Le congrès international de statistique de 1925 consacre à la question d'assez longues sessions, posant que :

La fraction retenue comme spécimen de l'ensemble doit être représentative (c'est à dire en tendance constituer une miniature de la population totale)

Deux possibilités s'offrent pour cela.

- Le tirage au hasard d'individus sur une liste (cela implique de posséder la liste ce qui est parfois compliqué)
- procéder par choix judicieux

Il convient pour que l'exercice soit rigoureux de comparer autant que possible plusieurs échantillons et d'accompagner les résultats d'une fiche détaillant les procédures suivies.

Nous n'en sommes encore là qu'au stade d'un accord entre spécialistes. Pour que échantillonnages et sondages soient utilisés par les acteurs sociaux et politique il faut attendre les années 30.

La naissance du sondage politique, et la révélation pour le grand public de son existence se produit en 1936 aux Etats-Unis quand Gallup, un universitaire qui fondera la compagnie qui porte son nom, parvient à prédire l'élection de Roosevelt que les analystes politiques donnaient perdant.

Il faut attendre les lendemains de la seconde guerre mondiale pour que sociologues et politistes français s'y intéressent

### 9.1.3 Les usages

Au lendemain de la guerre est fondé la Fondation IFOP par Jean Stoetzel de retour des Etats-Unis. Le nouvel institut public (Insee), lance lui de grande enquêtes par échantillonnage et sondage dans des domaines très divers :

Depuis quelques temps même l'Insee ne procède plus à dénombrement exhaustif de la population française, mais l'étudie par le biais de sondages à grande échelle

Les enquêtes par sondage sont aujourd'hui une pratique très courante dans de très nombreux domaines, en particulier dans le monde scientifique (épidémiologie, par exemple)

Les plus célèbres, les plus connues du grand public demeurent cependant bien sûr les enquêtes d'opinion et les sondages à la veille d'élections

### 9.1.4 Pq un tel succès ?

La réponse est assez simple :

Cette technique permet une diminution coût des enquêtes, coût de collecte des informations, mais aussi coût du traitement, particuliè-

rement avant la généralisation d'ordinateurs puissants  
 De tous s'accordent sur la fiabilité de la théorie mathématique sous-jacente qui est aujourd'hui solidement établie. Le prestige du chiffre et avantage donné par la caution scientifique dans le cadre de la civilisation contemporaine qui confère à celui qui peut avancer données chiffrées *a priori* fiable une forte légitimité y contribue également.

## 9.2 Une théorie mathématique

### 9.2.1 Estimation et intervalle de confiance

Le sondage s'appuie sur la théorie des probabilités et en particulier sur loi grands nombres qui pose que quand on répète un grand nombre de fois une même expérience, les valeurs obtenues au cours de celle-ci varient faiblement autour d'une valeur centrale.

L'exemple classique est la pièce de monnaie, si je joue un petit million de fois j'ai une chance raisonnable d'obtenir à peu près une moitié de pile et de face sauf si la pièce est truquée

Nous savons déterminer assez bien les formes de cette oscillation au moyen d'un notion mathématique qui s'appelle l'espérance

Du coup si on assimile les résultats de l'observation d'un individu à une expérience, voire chaque échantillon à une expérience (tirage). Nous pouvons déterminer, avec une assez grande précision, la plage de valeur à l'intérieur de laquelle se trouve le caractère étudié pour la population entière que nous voulons connaître.

Cette plage est exprimée de la façon suivante [valeur observée - ic/2, valeur observée + IC/2] ic étant l'intervalle de confiance associé à notre mesure, en fonction d'un facteur de risque accepté fixé à l'avance (cela signifie que la valeur recherché a 95/100 de chances de se trouver à l'intérieur de l'intervalle)

#### Cas d'une moyenne

Comment calcule-t-on cet intervalle Posons une population d'effectif N observée au moyen d'un échantillon de taille n Le taux de sondage (c'est-à-dire) la probabilité d'appartenir à l'échantillon est pour chaque individu est  $f = N/n$ .

Soit m la moyenne prise par une valeur Y dans notre échantillon Pour un échantillon assez grand nous aurons ;

$$IC = M \pm 1,96\sqrt{VZ} \quad (9.1)$$

Avec

$$VZ = \frac{(1-f)\rho^2}{N} \quad (9.2)$$

$$\rho^2 = \frac{\sum_{i=1}^n (Y_i - m)^2}{n-1} \quad (9.3)$$

En pratique on a le plus souvent quand n est grand (on le montre)

$$VZ = \frac{\rho^2}{n} \quad (9.4)$$

On obtient en simplifiant quelque peu  $IC = m \pm 1.96 \sqrt{\frac{variance Y}{n}}$

On constate que la valeur dépend très peu du taux de sondage (quand la population est grande), mais très fortement de n avec une courbe taille intervalle f(n) qui a une allure très particulière ( y=racine (1/x)

### Cas d'un pourcentage

On a dans ce cas là  $ro2=p(1-p)$  (p valeur observée) Ce qui nous donne  $IC= pplus ou moins 1.96 racine (1-f)p(1-p)/n$

Si on applique cela veut dire que pour une valeur p proche de 50/100 pour un échantillon comprenant 500 individus il y a 95/100 de chances pour que P (valeur pour ensemble de la popu) soit compris entre 500  $p-5 < P < p+5$  1000  $p-3.1 < P < p+3.1$  2000 2.3 5000 1.4 10 000 1

La prudence est de mise quand on a affaire à une enquête isolée Ajoutons que IC ne mesure que l'effet possible du tirage de tel ou tel échantillon, bien d'autres facteurs peuvent conduire à la formation de biais

### 9.2.2 Choix raisonnés

La volonté d'obtenir des données les plus précises possibles sans pour autant faire exploser les coûts, la difficulté pratique parfois à opérer un tirage aléatoire conduisent les spécialistes des sondages à multiplier les opérations censées améliorer la qualité de leur produit

#### échantillon stratifié

Le technique la plus courante est la production d'un échantillon dont on contrôle au préalable le fait qu'il respecte la distribution d'un certain nombre de variables de contrôles dont les valeurs sont connues pour la population dans son ensemble (ex le plus classique

age/sex/csp) Plusieurs techniques : la plus courante distribuer aux enquêteurs une feuille à remplir sur laquelle sont précisées leurs cibles (un homme + 50 ans, csp + parisien). On leur demande alors de contacter les personnes correspondant à leurs cibles jusqu'à ce que l'une d'elles réponde.

Inconvénient majeur pour certaines enquêtes : le taux non réponses peuvent varier fortement en fonction catégories de population, on peut se retrouver avec comme répondants des individus tout à fait atypiques au regard de la catégorie qu'ils sont censés représenter.

### échantillon par grappe

On tire au sort des groupes d'individus dont on va observer les membres (passager d'un train pour les usagers de la Sncf), passager d'une compagnie aérienne et on va tirer les avions. C'est aussi le principe des sondages post-électoraux (tirage des bureaux de vote) Là encore des difficultés sont possibles, liée à l'existence de possibles effets de grappe (les passagers d'une même ligne peuvent varier selon heures, jours, période années), ou selon des variables pas toujours connues et donc difficiles à contrôler.

### 9.2.3 Redressement de variables

Dans un certain nombre cas, il est possible d'exercer un contrôle a posteriori sur les données brutes, quand la situation d'enquête fait soupçonner la possibilité d'un biais systématique. L'exemple typique est la sous déclaration certaines valeurs (par exemple le vote front national, longtemps peu avouable). Si est connue une variable seconde, fortement liée à la première, on peut tenter de calculer ampleur du biais. (Exemple : connaissance le nombre de votants FN en 2002, je demande aux gens que j'interroge combien ont voté pour FN en 2002, j'ai un estimateur de sous déclaration parce que je connais le nombre de gens qui ont effectivement voté FN en 2002, pour arriver à mes fins, ou bien on peut utiliser des variables d'opinions liées entre elles (favorable à la peine mort, trop d'immigrés en France, image positive du régime de Vichy).

Nous avons vu cependant récemment qu'une marge d'incertitude subsistait

### **9.3 Lectures critiques**

#### **La critique des sondages**

Incertitudes et imprécisions nourrissent une critique des sondages parfois virulente qui est parfois simple ignorance de ce qu'est le sondage. Une estimation, une mesure, donc une indication relative à une valeur probable, mais jamais, parce que c'est impossible une « vraie valeur ».

D'autres critiques plus virulentes et plus pertinentes portent plus particulièrement sur un type particulier de sondage, le sondage d'opinion et de valeurs. Patrick Champagne dont je vais ici reprendre l'argumentaire [?] est le meilleur représentant de ce courant.

Il distingue 3 type enquête

- comportements et pratiques politiques (vote effectué)
- intention de vote ou sondage sortis des urnes qui sont des intentions de comportement
- opinions et valeurs sur lesquels portent ses critiques

Selon lui ils ne décrivent pas des opinions ou des valeurs mais des déclarations produites en réponse à un stimulus exercé dans circonsances particulières qui n'ont pas d'équivalent dans la vie réelle

On peut considérer que la pratique est :

- sans enjeu pour locuteur qui peut avoir un usage dégagé voire ludique
- peut aboutir à obtenir des gens des réponses à des questions qu'ils ne se posent pas voire qu'ils ne comprennent pas.
- Très fortes variations des résultats selon libellé des questions d'une part, formulation des réponses préformatées de l'œuvre

Prenons quelques exemples. A la question posée en 1987 « Pensez vous qu'aujourd'hui l'état s'oriente réellement vers un changement de politique concernant les économies d'énergie » Le pourcentage d'individus répondant oui varie de 2 à 66/100 selon que les réponses possibles sont Oui/non ou bien (oui bien ; oui très sérieusement ; oui mais prudemment ; oui mais de façon ponctuelle ; oui mais de façon incohérente ; non ; ne sait pas)

De même façon quand on demande après une question portant sur les atteintes à l'environnement aux enquêtes ce qu'ils entendent spontanément par environnement : 24/100 évoquent la pollution, 12/100 la protection nature, 36 le cadre urbain et social 24 le cadre de vie, 26 la nature

De même dans une enquête de 1087, quand on interroge les enquêtés pour savoir si ils connaissent la réglementation en matière d'environnement, 23/100 répondent oui, mais 95/100 ont une opinion sur qualité de celle-ci.

En résumé le sondage opinion revient à demander à des gens qui ne s'en préoccupent pas leur avis sur une question et à obtenir d'eux une réponse même si ils ignorent tout du sujet, voire ne comprennent pas la question

Le sondage donne bien accès à quelque chose, mais certainement pas à ce que pensent les gens.

Certains sondages, et les sondages d'opinion sont souvent de ce type, surtout quand ils se présentent isolés, révèlent surtout les discours et les stratégies de communication des commanditaires de l'enquête, qui tentent d'imposer comme sujet de débat voire comme argumentaire, un thème un problème, ce qui va déterminer les questions posées

Cette critique du sondage d'opinion, même si on pas obligé de partager tous ses attendus a une vertu. Elle rappelle que le sondage ne peut se lire, s'utiliser, se comprendre vraiment, qu'en référence au contexte de sa production, et en connaissant les opérations, dont certaines mettent en œuvre des techniques mathématiques. L'activité historique, en ce sens, quelle que soit son instrumentation et son usage ou non du nombre, demeure une activité fondamentalement critique, dont l'originalité première est sans doute une attention maniaque aux conditions de production des matériaux qu'elle utilise et aux transformations que ceux-ci subissent lors de leur traitement par l'historien, une critique donc, particulièrement scrupuleuse quand elle s'intéresse aux opérations de l'historien lui-même. Que les outils soient aujourd'hui souvent numériques et que la description du matériau utilisé suppose sans doute plus fréquemment qu'autrefois aux statistiques, ou à diverses branches des mathématiques ne change rien à l'affaire.



# Bibliographie

- [Ben82] Jean-Paul Benzecri. *Histoire et préhistoire de l'analyse de données*. Dunod, Paris, 1982.
- [Ber99] Jacques Bertin. *Sémiologie graphique : Les diagrammes, les réseaux, les cartes*. Presses de l'Ehess, 1999.
- [Ber08] Gérard Berry. *Pourquoi et comment le monde devient numérique*. Fayard, 2008.
- [BH97] André Salem Benoît Habert, Adeline Nazarenko. *Les linguistiques de corpus*. Armand Colin, 1997.
- [Cha90] Patrick Champagne. *Faire l'opinion. Le nouveau jeu politique*. Éditions de minuit, Paris, 1990.
- [Cib83] Philippe Cibois. *Les méthodes d'analyse d'enquête*. PUF, 1983.
- [Cib84] Philippe Cibois. *L'analyse des données en sociologie*. PUF, 1984.
- [CL08] Claire Zalc Claire Lemercier. *Méthodes quantitatives pour l'historien*. La découverte, Paris, 2008.
- [CN13] Frédéric Clavert and Serge Noiret, editors. *L'histoire contemporaine à l'ère numérique*. Peter Lang, 2013.
- [Cor10] Pierre-André Cornillon. *Statistiques avec R*. Presses universitaires de Rennes, Rennes, 2010.
- [eSP12] Thierry Lafouge et Stéphanie Pouchot. *Statistiques de l'intellect. Lois puissances inverses en sciences humaines et sociales*. Presses de l'Ensib, Villeurbanne, 2012.
- [Gue04] Alain Guerreau. *Statistique pour historien*. École des chartes, 2004.
- [Her07] Clarisse Herrenschmidt. *Les trois écritures. Langue , nombre, code*. Gallimard, Paris, 2007.

- [JPG11] Andrea Zorzi Jean Philippe Genet, editor. *Les historiens et l'informatique : un métier à réinventer*, volume 44 of *Collection de l'École française de Rome*. École française de Rome, Rome, 1 edition, 2011.
- [LL95] Marie Piron Ludovic Lebart, Alain Morineau. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 1995.
- [Mar00] Nicolas Markey. *Tame the BeaST. BibTEX de B à X*. Ens Cachan, Cachan, 2000.
- [Mül04] Jean-Pierre Müller. ttdd - une librairie r pour l'analyse de données textuelles. In *JADT 2004. Septièmes Journées internationales d'Analyse statistique des Données Textuelles*, 2004.
- [Noi11] Serge Noiret. La digital history : histoire et mémoire à la portée de tous. *Ricerche storiche*, XLI(1) :111–149, avril 2011.
- [Oet11] Tobias Oetiker. *Une courte (?) introduction à Latex2e*. Laas/Cnrs, 2011.
- [Ryg10] Philippe Rygiel. L'inchiesta storica in epoca digitale. *Memoria e ricerca. rivista di storia contemporanea*, 35, 2010.
- [SG05] Frédéric Saly-Giocanti. *Utiliser les statistiques en histoire*. Armand Colin, Paris, 2005.