NUMERO III : PROGRES ET CRISES

Revue Crises et Société

Le recours aux humanités numérique en histoire du droit – Chronique Culture

Par Léo BRUN et Pauline VERDIER

Pour citer cet article : Léo BRUN et Pauline VERDIER, « Le recours aux humanités numérique en histoire du droit – Chronique Culture [en ligne], *Revue Crises et Société*, 3 (2024), disponible sur : https://www.crisesetsociete.com

---

Le recours aux humanités numériques en histoire du droit

En histoire du droit, se développent, notamment dans le cadre de l'histoire des idées[1], de nouvelles recherches qui viennent parfois remettre en cause certains paradigmes en place depuis longtemps. Pour ce faire, les historiens puisent dans les moyens à leur disposition et mobilisent notamment les humanités numériques, qui sont des outils utiles pour leurs travaux[2]. Dans ce cadre, il nous paraît important de signaler que cet article est une simple présentation de la méthode que nous tentons d'utiliser, afin de dresser un panorama – non exhaustif – de ce qui peut être fait à l'aide des humanités numériques.

Dans un premier temps, une question semble s'imposer : celle qui a vocation à déterminer pourquoi avoir recours aux humanités numériques, ainsi qu'à des représentations graphiques de leurs résultats. Il nous semble que ce recours ne doive pas être forcé : il survient essentiellement d'un besoin, de questions que l'on se pose et auxquelles les humanités numériques nous apparaissent les plus aptes à répondre. Le problème d'une lecture seule, aussi fine soit-elle, est que l'on risque de passer à côté d'éléments déterminants d'analyse des textes. L'outil informatique apparaît dès lors, sinon absolument indispensable, du moins très utile. Tout ceci même si, n'étant pas lui-même infaillible, il suppose notamment une vérification manuelle des résultats (pour le chercheur qui constitue lui-même sa base de données). Le chercheur ne fait donc que s'appuyer sur l'outil informatique.

À titre d'exemple, et afin d'être plus précis, nous allons nous centrer principalement sur les *Archives de philosophie du droit et de sociologie juridique*, revue juridique publiée durant l'entre-deux-guerres. Elle constitue en effet une bonne illustration de nos propos, et a déjà fait l'objet d'une étude opérée par nos soins[3]. Nous nous bornerons à faire figurer quelques interrogations relatives à cette publication et à montrer l'intérêt des humanités numériques pour y répondre. L'étude d'une revue telle que celle-ci peut donner lieu à de nombreuses interrogations, notamment sur les sources de réflexion des auteurs, leurs lectures, les thématiques abordées, etc. Les humanités numériques permettent de répondre relativement vite à ce type de questions.

Ainsi, pour mettre en place une méthode d'analyse des données, il faut décomposer notre travail en plusieurs étapes, que nous allons détailler. Il s'agit de la création et du tri des données, puis de leur représentation, et enfin de l'exploitation des représentations graphiques. Il nous apparaît plus clair de les présenter en deux étapes, du fait de l'imbrication des différentes phases les unes dans les autres.

La phase préliminaire : vers la création des bases de données

Dans ce cadre, la réalisation d'analyses de données demande un important travail en amont comme en aval de l'utilisation de l'outil informatique. Nous pensons que cela ne demande ni plus, ni moins de travail à l'historien : seulement une mise en œuvre différente de ses capacités, par l'utilisation des machines. On ne collecte évidemment pas tout et n'importe quoi, mais on se demande en amont ce qui serait utile pour répondre à notre besoin. Ainsi, l'historien, dans sa subjectivité, apparaît à travers les choix qu'il opère. Par conséquent, il ne s'agit pas pour nous de prôner l'objectivité des humanités numériques, mais d'en affirmer la plus grande fiabilité, lorsqu'elles sont mises en œuvre avec soin. D'autant plus qu'il est

particulièrement important de joindre à l'étude quantitative, une étude qualitative des résultats obtenus. Par exemple, si étudier les auteurs cités dans un ouvrage, voire dans toute l'œuvre d'un juriste peut être révélateur, s'intéresser à l'utilisation qu'il fait de ses sources (éloge, critique, ou simplement mention sans prise de position) l'est d'autant plus. Cet exemple nous permet de mettre en avant un nouvel aspect de la contribution de l'outil informatique à l'analyse du chercheur.

Pour expliciter cela, arrêtons-nous ensemble quelques instants sur ce processus de « création des bases de données ». En amont, il suppose – comme tout travail de recherche en histoire – la sélection d'un corpus de travail, de sources, de textes à étudier. Il peut ainsi s'agir simplement d'un document d'archives, d'un imprimé, d'une revue : en somme de tout texte ancien ou récent. Une fois l'objet d'étude défini, commence la phase de création des bases de données et avec elle, l'utilisation d'outils informatisés qui vont accélérer, voire même rendre possible notre travail d'analyse de données.

Pour commencer ce dernier, il nous faut récupérer le texte qui fait l'objet de notre étude sur ordinateur. Deux options s'offrent alors à nous en fonction du support originel du document : papier ou numérique. Évidemment, le processus de création des bases de données est beaucoup moins laborieux pour le chercheur qui dispose d'une source préalablement numérisée. Il faut noter à ce stade que la qualité de l'image numérisée est importante. La netteté des lettres et chiffres, le caractère explicite du découpage des paragraphes du document, la rectitude des lignes, le contraste entre les caractères et le fond de l'image sont des éléments susceptibles de moduler la qualité du travail mobilisant les humanités numériques. Ces précisions nous semblent importantes car le scan – qui est sans doute l'outil de numérisation le plus répandu – n'est pas toujours possible, en fonction du document source : une reliure trop rigide, ou encore un ouvrage trop fragile sont autant d'obstacles à sa mise en œuvre. Il est essentiel d'adapter la technique de capture de l'image au support : si elle peut s'opérer par différents moyens, ce n'est pas sans incidence sur la qualité finale du fichier numérisé[4].

Il est par conséquent important que le travail de numérisation soit mis en place avec les bons outils et par des personnes conscientes des éléments permettant l'obtention de documents de bonne qualité (les chercheurs formés à ces questions sont ainsi certainement parmi les plus aptes). Ce travail est parfois complexe, car il peut demander certaines capacités techniques[8], mais son utilité est essentielle au bon fonctionnement du processus d'OCR[5]. Cette assertion est à mesurer puisque la plupart des programmes de reconnaissance de caractères fiables fournissent des moyens d'augmenter la qualité de l'image. Ainsi, un grand nombre de documents peuvent aujourd'hui être convertis au format texte sans pré-traitement manuel des images qui les composent. Il faut toutefois nuancer notre propos puisque l'expérimentation nous a montré que les filtres destinés à améliorer l'image ont parfois un effet contraire à celui escompté. Les variations de luminosité, ou une capture d'image trop sombre ou trop peu contrastée peuvent largement s'empirer sous l'effet de tels filtres (alors que les images restent très aisément lisibles sans cela). La conséquence de ceci est l'obtention d'une image peu lisible pour l'ordinateur, rendant impossible un travail de qualité mobilisant les humanités numériques. Cela montre l'importance capitale d'une bonne image.

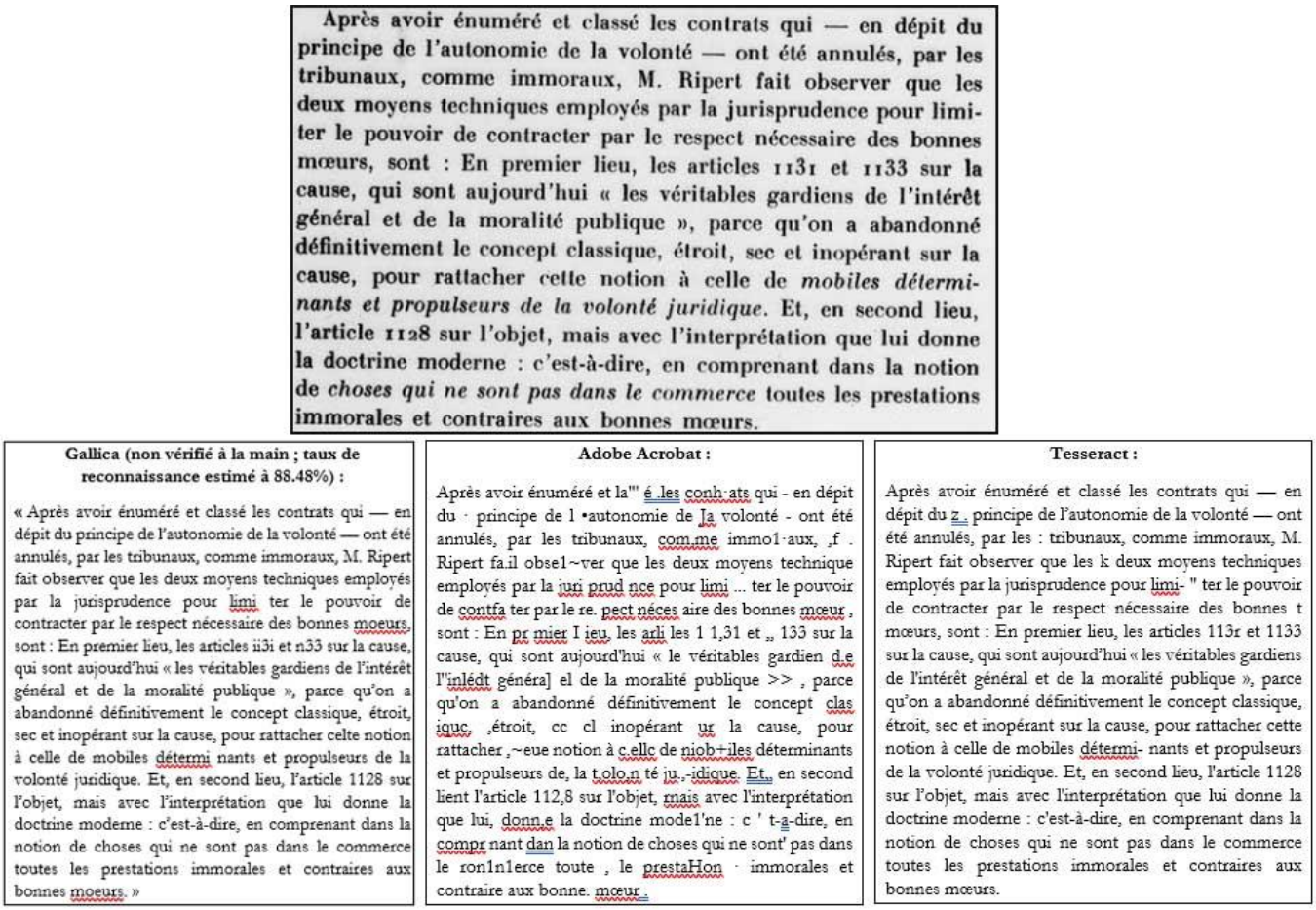
La numérisation est donc l'un des enjeux actuels de la recherche, car elle permet de rendre accessibles des documents qui ne l'auraient pas été auparavant (ou du moins difficilement[6]). L'intérêt et les lacunes des bases de données en ligne ont déjà été relevés dans notre discipline[7], mais ce constat vaut pour toute matière historique. Celles-ci sont nombreuses (notons par exemple Gallica, Persée, Dalloz, RetroNews ou encore Cairn) et en expansion ces dernières

années. Nous sommes pourtant encore bien loin d'un monde dans lequel le chercheur aurait accès en quelques clics à tout document nécessaire à son travail. Ainsi, dans la majorité des situations, c'est au chercheur de numériser les textes qu'il entend analyser.

Ce n'est que lorsque notre source est identifiée et mise sous format numérique que commence véritablement le travail de création des bases de données. Il repose alors principalement sur l'OCR (Optical Character Recognition) – selon la terminologie généralement consacrée – ou en français ROC (Reconnaissance Optique de Caractères) qui correspond au processus de transcription d'une image en texte. L'exploitation des résultats du processus d'OCR peut alors nous servir pour rechercher la présence de mots clés au sein des documents envisagés, afin de procéder à différentes études statistiques. Il peut également être utilisé afin d'identifier les textes susceptibles d'être utiles à nos travaux. C'est d'ailleurs ce que font aujourd'hui tous les chercheurs – qu'ils soient conscients ou non d'utiliser les résultats de document convertis en texte par OCR pour ce faire – lorsqu'ils tapent des mots clés sur les différents moteurs de recherches (Google Scholar, Gallica, Persée, etc.). L'utilisation de cette Reconnaissance Optique de Caractères au sein des bases de données est de plus en plus développée : en France avec Gallica ou Persée, à l'internationale sur Jstor, Google, Wikisource, ou encore l'effort de numérisation de la NDL (National Diet Library qui est la bibliothèque de la chambre basse japonaise)[8].

Les programmes d'OCR sont donc nombreux et très divers dans leurs modalités. Certains sont très faciles à utiliser, même pour les néophytes, puisqu'ils sont inclus dans des programmes bien connus (c'est notamment le cas de la fonction OCR d'Adobe Acrobat DC). D'autres sont plus difficiles d'accès, car ils demandent certaines connaissances techniques pour être utilisés[9]. C'est notamment le cas de l'OCR de la NDL et de celui de Google (Pytesseract[10]) qui supposent tous deux l'utilisation de Python. Ces solutions sont parmi les plus efficaces que nous avons identifiées, tout en restant accessibles gratuitement.

Le fonctionnement et l'utilité concrets de l'OCR restant à démontrer, il est opportun de présenter ce dont elle est capable. Pour ce faire, nous avons utilisé plusieurs programmes d'OCR sur un même texte[11] :



Après avoir énuméré et classé les contrats qui — en dépit du principe de l'autonomie de la volonté — ont été annulés, par les tribunaux, comme immoraux, M. Ripert fait observer que les deux moyens techniques employés par la jurisprudence pour limiter le pouvoir de contracter par le respect nécessaire des bonnes mœurs, sont : En premier lieu, les articles 1131 et 1133 sur la cause, qui sont aujourd'hui « les véritables gardiens de l'intérêt général et de la moralité publique », parce qu'on a abandonné définitivement le concept classique, étroit, sec et inopérant sur la cause, pour rattacher cette notion à celle de *mobiles déterminants et propulseurs de la volonté juridique*. Et, en second lieu, l'article 1128 sur l'objet, mais avec l'interprétation que lui donne la doctrine moderne : c'est-à-dire, en comprenant dans la notion de *choses qui ne sont pas dans le commerce* toutes les prestations immorales et contraires aux bonnes mœurs.

Gallica (non vérifié à la main ; taux de reconnaissance estimé à 88.48%) :

« Après avoir énuméré et classé les contrats qui — en dépit du principe de l'autonomie de la volonté — ont été annulés, par les tribunaux, comme immoraux, M. Ripert fait observer que les deux moyens techniques employés par la jurisprudence pour limi ter le pouvoir de contracter par le respect nécessaire des bonnes moeurs, sont : En premier lieu, les articles ii3i et n33 sur la cause, qui sont aujourd'hui « les véritables gardiens de l'intérêt général et de la moralité publique », parce qu'on a abandonné définitivement le concept classique, étroit, sec et inopérant sur la cause, pour rattacher celte notion à celle de mobiles détermi nants et propulseurs de la volonté juridique. Et, en second lieu, l'article 1128 sur l'objet, mais avec l'interprétation que lui donne la doctrine modeme : c'est-à-dire, en comprenant dans la notion de choses qui ne sont pas dans le commerce toutes les prestations immorales et contraires aux bonnes moeurs. »

Adobe Acrobat :

Après avoir énuméré et la'" é les conh·ats qui - en dépit du · principe de l •autonomie de Ja volonté - ont été annulés, par les tribunaux, com me immo1·aux, ,f . Ripert fa.il obse1~ver que les deux moyens technique employés par la juri prud nce pour limi ... ter le pouvoir de contfa ter par le re. pect néces aire des bonnes mœur , sont : En pr mier I ieu, les arli les 1 1,31 et „ 133 sur la cause, qui sont aujourd'hui « le véritables gardien d.e l''inlédt génera] el de la moralité publique >> , parce qu'on a abandonné définitivement le concept clas iquc, ,étroit, cc cl inopérant ur la cause, pour rattacher ,~eue notion à c.ellc de niob+iles déterminants et propulseurs de, la t.olo.n té ju.,-idique. Et., en second lient l'article 112,8 sur l'objet, mais avec l'interprétation que lui, donn.e la doctrine mode1'ne : c ' t-a-dire, en compr nant dan la notion de choses qui ne sont' pas dans le ron1n1erce toute , le prestaHon · immorales et contraire aux bonne. mœur .

Tesseract :

Après avoir énuméré et classé les contrats qui — en dépit du z., principe de l'autonomie de la volonté — ont été annulés, par les : tribunaux, comme immoraux, M. Ripert fait observer que les k deux moyens techniques employés par la jurisprudence pour limi- " ter le pouvoir de contracter par le respect nécessaire des bonnes t mœurs, sont : En premier lieu, les articles 113r et 1133 sur la cause, qui sont aujourd'hui « les véritables gardiens de l'intérêt général et de la moralité publique », parce qu'on a abandonné définitivement le concept classique, étroit, sec et inopérant sur la cause, pour rattacher cette notion à celle de mobiles détermi- nants et propulseurs de la volonté juridique. Et, en second lieu, l'article 1128 sur l'objet, mais avec l'interprétation que lui donne la doctrine moderne : c'est-à-dire, en comprenant dans la notion de choses qui ne sont pas dans le commerce toutes les prestations immorales et contraires aux bonnes mœurs.

Nous remarquons que dans tous les cas, le processus tente de transcrire l'image en texte. Cependant, en fonction du programme utilisé, le résultat est plus ou moins bon. Tous ne se valent donc pas et il est certain que l'utilisation des meilleurs disponibles augmente la qualité du travail de recherche (tout en limitant le contrôle par le chercheur de la qualité des résultats obtenus). On remarque dans l'exemple que le texte issu d'Adobe est bien moins fidèle que celui de Google.

Les détails techniques de ces processus d'OCR étant nombreux, nous n'avons abordé que les principaux dans le cadre de cette étude, qui se veut concise. D'autres existent pourtant, comme la langue des documents, qui peut faire varier les résultats, car les divers programmes d'OCR ne sont pas capables des mêmes prouesses pour toutes les langues. C'est notamment le cas pour celles qui utilisent des caractères non latins (japonais, russe, coréen, arabe, chinois, etc.). Ces quelques limites techniques en tête, l'OCR possède donc de nombreux avantages et son utilisation est aujourd'hui accessible à tous les chercheurs.

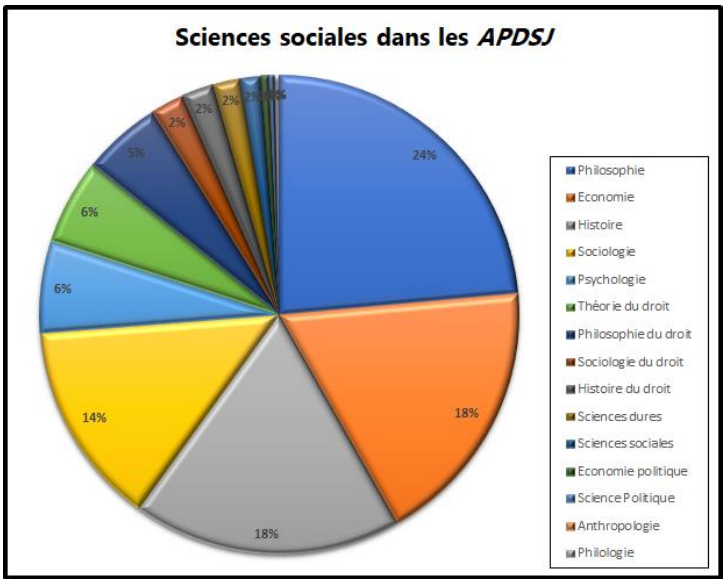
Une fois le document sélectionné, numérisé et le texte reconnu, vient l'extraction des données. Ici encore, le chercheur doit choisir les données qu'il souhaite extraire en fonction de ce qu'il entend démontrer. L'avantage de travailler avec un document en format texte est l'accroissement de l'efficacité de travail sur des analyses plus poussées. Si le chercheur veut savoir si un auteur parle plutôt de façon positive ou négative d'un concept, la recherche dans le texte lui permettra de cibler son étude sur cette expression afin de ne pas en laisser passer

certaines occurrences, et de ne pas prendre trop de temps dans la recherche de celles-ci. Concrètement, il s'agit de mettre en place une recherche par mots-clés.

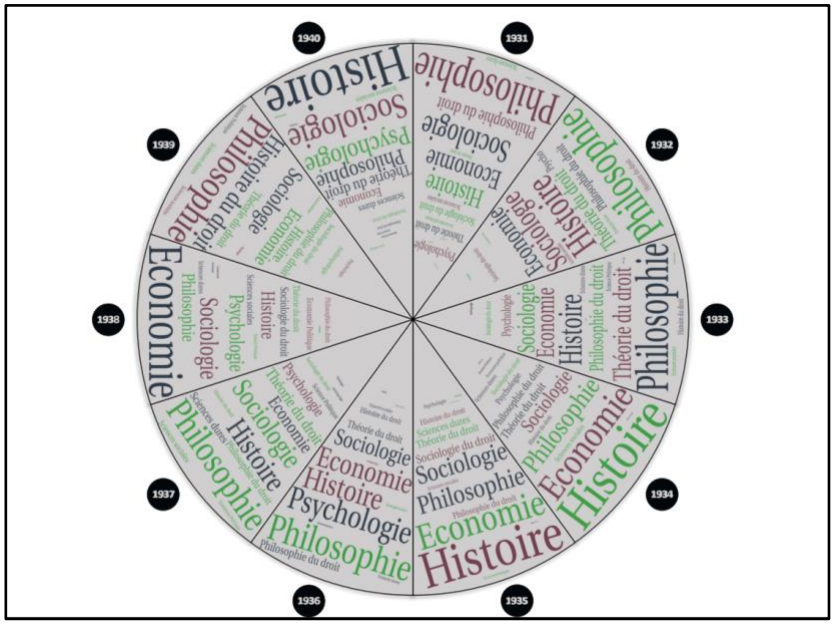
Le vif du sujet : l'exploitation des données

Ensuite, il faut choisir le graphe en fonction des données sélectionnées et de notre objectif. Même si les chiffres montrent souvent assez bien d'eux-mêmes les grandes tendances, le mieux est de créer des représentations graphiques capables de les faire parler plus expressément. Cela permet également de donner une représentation plus visuelle, donc plus accessible et compréhensible que ne le sont de simples chiffres. Pour ce faire, des classiques représentations disponibles dans Excel à la mise en forme par Photoshop, en passant par les graphes Gephi, de multiples outils sont à la disposition de l'historien du droit. Le choix de la représentation graphique est capital pour plusieurs raisons. Premièrement, parce que toutes les représentations graphiques ne se valent pas : certaines sont très mauvaises, car les explications adjacentes doivent être très nombreuses pour les rendre lisibles. Cela n'est pas souhaitable, puisqu'elles perdent tout intérêt en ne rendant pas les données plus accessibles qu'auparavant.

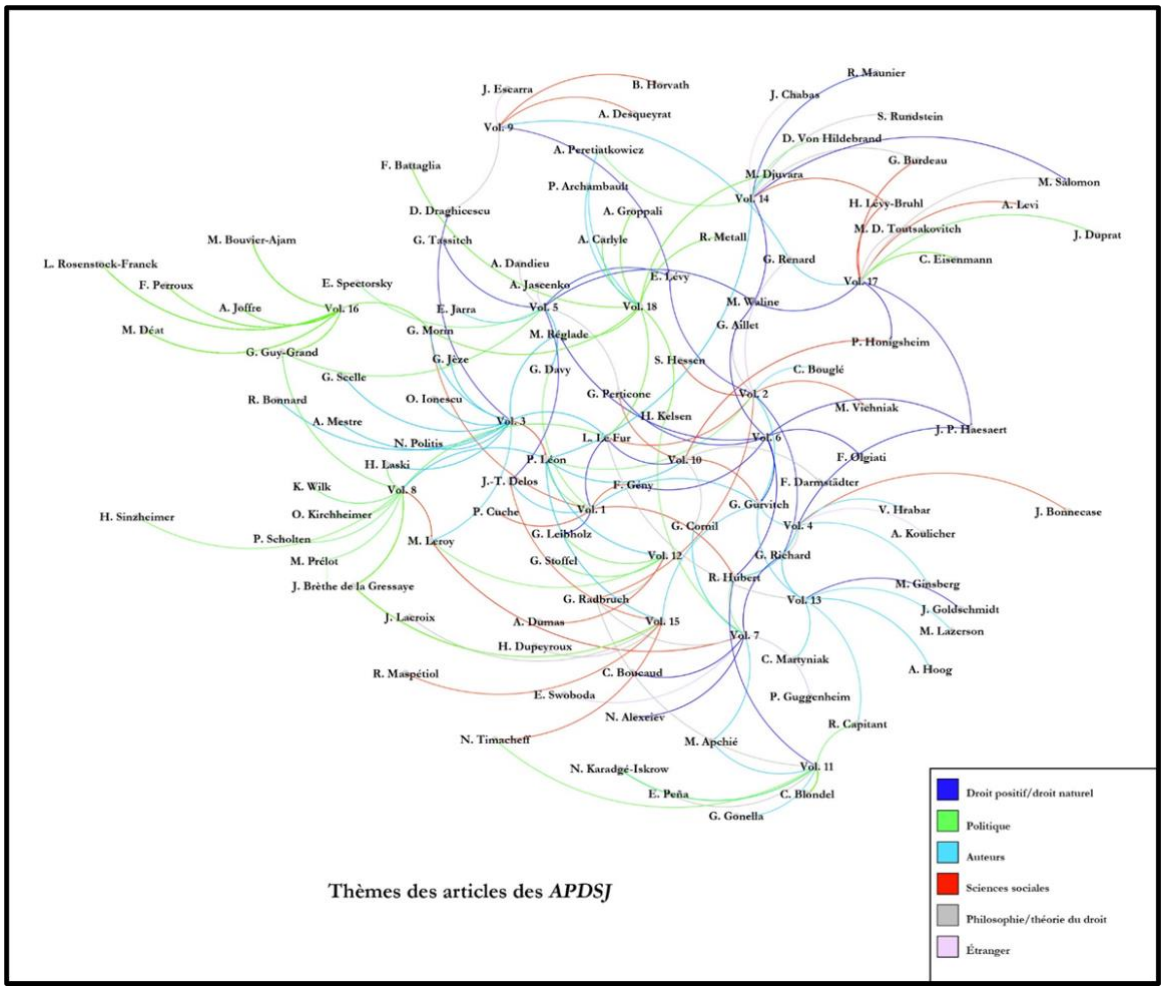
Deuxièmement, car par les représentations graphiques, on peut exploiter des chiffres identiques de différentes façons, les utiliser pour mettre divers éléments en valeur. Nous pensons qu'il existe toujours diverses manières de faire parler les chiffres. Pour reprendre l'exemple des *Archives de philosophie du droit et de sociologie juridique*, dans l'étude des différentes sciences sociales représentées, une étude d'un certain nombre d'occurrences a été opérée. Nous avons choisi de présenter quatre types de représentations auxquelles une étude comme celle-ci peut donner lieu, et d'en expliquer brièvement les supports. Cela nous permet d'en montrer l'intérêt et la complémentarité. La première, récapitulant toutes les occurrences de chaque discipline dans notre exemple, montre les thématiques les plus abordées par les auteurs en son sein. Elle présente l'avantage de donner une vue d'ensemble de la revue, ce qui ne serait pas simple sans les outils présentés (puisqu'il est difficile, voire impossible, de synthétiser d'une vue tant de données). Pour donner naissance à une telle représentation, nous nous sommes limités à l'utilisation d'Excel.



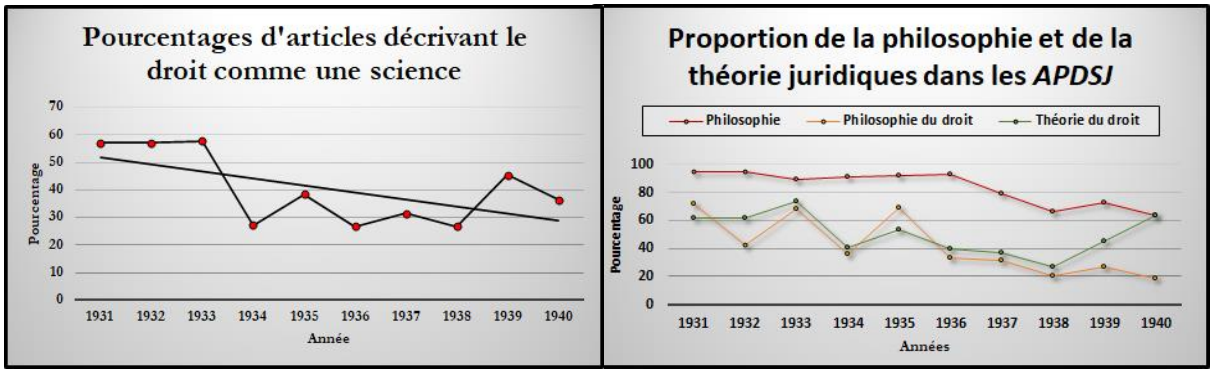
La seconde, plus précise sans doute, à partir des mêmes données, se concentre sur la présence de ces disciplines au cours des années. Elle ne permet pas de voir quelles sciences sociales sont les plus présentes dans l'absolu, puisque le graphique n'est pas – en l'occurrence – à l'échelle d'une année sur l'autre[12], mais de mettre en lumière les variations annuelles des préoccupations des contributeurs. Ainsi, à partir des mêmes données, toutes issues du travail sur le texte issu de l'OCR, on peut étudier différents aspects des documents étudiés. Les informations données par la première représentation se trouvent ainsi complétées, puisque celle-ci donne simplement une idée générale de l'importance des disciplines dans l'ensemble de la revue. Cela permet d'entrevoir deux aspects de la revue : le premier tend à la percevoir comme un bloc uni et unique, alors que le second la décompose, dans un découpage que l'on pourrait imaginer différent, mais qui reste éclairant.



On pourrait arguer qu'un défaut de la représentation ci-dessus est encore en un sens le manque de précision, surtout au vu de l'exemple choisi : puisqu'il s'agit en l'occurrence d'une revue, il est possible de ne pas traiter chaque volume comme un ensemble, mais de prendre en compte les spécificités de ses contributeurs. Dès lors, pour étudier une revue, on peut, dans un cas, donner une image plus complète de la diversité des sous-thèmes abordés (comme les deux graphes ci-dessus) ou avoir une approche plus générale de ce point, mais donner à voir un paysage plus précis des auteurs impliqués (comme celui ci-dessous). Ainsi, en fonction de ce qu'il nous paraît intéressant d'observer, nous avons plusieurs types de représentations à notre disposition pour mettre en valeur nos données.



Il est également possible, par exemple, d'étudier des tendances, au sein d'une revue comme d'un ensemble d'ouvrages du même auteur (cf. graphe de gauche). Cela permet d'étudier les modifications d'orientation du périodique ou les évolutions de la pensée de l'auteur en question. On peut même imaginer une comparaison de tendances entre deux expressions concurrentes (cf. graphe de droite).



Les représentations permettent donc plusieurs degrés de précision, de fabriquer plusieurs catégories, grilles de lecture, mais également sont plus ou moins faciles à comprendre d'un regard. Il nous semble, de façon assez intuitive, que plus une représentation est précise, moins elle est claire, et inversement. Il nous semble important de trouver une, ou des représentations qui soient suffisamment élaborées pour être précises et suffisamment simples pour être

comprises sans trop d'explications complémentaires : les commentaires permettent (et doivent permettre, selon nous) la plupart du temps seulement de tirer des conclusions, et d'aller au-delà des données brutes.

Les graphiques que nous venons de présenter permettent ainsi de mettre en lumière des éléments qui auraient été négligés par des études ne s'intéressant qu'aux titres et à des articles en particulier de la revue étudiée. Ici, comme ailleurs, une telle étude permet d'accéder à des informations nouvelles en évitant le piège de se limiter aux paradigmes existants. Au contraire, en invitant à les repenser, on comprend bien l'apport concret que peuvent représenter les humanités numériques en histoire du droit.

[1] Cf. notamment : Tom Clark, Benjamin Lauderdale, « The Genealogy of Law », *Political Analysis*, n° 20, 2012, p. 329-350.

[2] Cf. notamment : Nader Hakim, Annamaria Monti, « Histoire de la pensée juridique et analyse bibliométrique : l'exemple de la circulation des idées entre la France et l'Italie à la Belle Époque », *Clio@Themis*, n° 14, 2018 ; Pierre Bonin, « L'historiographie de l'histoire du droit, tendance récente et prochains territoires », *in* Jacques Krynen, Bernard d'Alteroche (dir.), *L'Histoire du droit en France. Nouvelles tendances, nouveaux territoires*, Paris, Garnier, 2014, p. 552-553.

[3] Pauline Verdier, *Sauver la science du droit : l'exemple des Archives de Philosophie du droit et de Sociologie juridique (1931-1940)*, Mémoire Droit Bordeaux, 2023.

[4] Par exemple, la photo peut être prise avec des filtres destinés à aligner le texte ou encore à en augmenter le contraste avec le fond de l'image. Pour ce faire, il existe des applications gratuites qui permettent de reproduire l'effet d'un scan, notamment pour les ouvrages qu'il est difficile de manipuler.

[5] L'OCR (Optical Character Recognition) – selon la terminologie généralement consacrée – ou en français ROC (Reconnaissance Optique de Caractères) correspond au processus de transcription d'une image en texte.

[6] C'est notamment le cas des documents conservés uniquement à l'étranger, car leur emprunt est parfois difficile et souvent couteux.

[7] Cf. Nader Hakim, Annamaria Monti, « Histoire de la pensée juridique et analyse bibliométrique… », *op.cit.*, § 26.

[8] L'OCR utilisée est d'ailleurs à la disposition de tous, puisqu'elle est fournie en Open Source directement sur internet : https://github.com/ndl-lab/ndlocr_cli

[9] Certains programmes gratuits et à l'accès assez simple sont également disponible, à l'instar d'OCR4All par exemple, qui produit des résultats satisfaisants.

[10] Il est possible d'y accéder gratuitement et librement en ligne : https://github.com/tesseract-ocr/tesseract

[11] Extrait : Zuleta, « La règle morale dans la législation colombienne sur les obligations », *Bulletin de la Société de législation comparée*, n° 64, 1935, p. 420.

[12] Les nuages de mots ont été générés séparément pour chaque année : ils ne sont pas à l'échelle les uns par rapport aux autres. Ils ont été regroupés par Photoshop afin de permettre une comparaison plus aisée entre les années.