Bulletin de Méthodologie Sociologique
115 53–65

DOI: 10.1177/0759106312445708
http://bms.sagepub.com

Traiter des « masses » de données prosopographiques par la numérisation d'annuaires - Espoirs et vertiges

Sylvain Laurens
Francis Marchan
Département de sociologie, MCF Université Limoges, GSPE, Gresco, Limoges

Abstract
Treating "Masses" of Prosopographical Data by Scanning Directories - Hopes and Disorientation: This note aims to provide an update on the progress made in optical character recognition (OCR) and the contribution of these techniques to the creation of prosopographical data bases in social sciences. With the example of a European investigation of European business associations, it highlights the progress made possible by OCR with the analysis of several biographical directories identifying groups of business interests. Based on this example, it is hypothesized that the development of digital technologies allow the creation of corpuses of data much larger than in the past in the framework of quantitative inquiries conducted by smaller teams. However, this article also highlights the fact that this extension of corpuses – made possible by scanning – raises new problems of method, starting with the increased time devoted to the standardization of digital data.

Résumé
Cette note se propose de faire le point sur les avancées faites en reconnaissance optique des caractères (OCR) et sur la contribution de ces techniques à la constitution de bases de données prosopographiques en sciences sociales. A partir de l'exemple d'une enquête

Corresponding Author:
Sylvain Laurens, Département de sociologie, Université de Limoges, 39E rue Camille Guérin, 87000 Limoges, France.
Email: sylvain.laurens@unilim.fr

sur les lobbys européens, il traite notamment des progrès permis par l'OCR avec l'analyse de plusieurs annuaires biographiques recensant des groupes d'intérêts patronaux. A partir de cet exemple, il émet l'hypothèse que le développement des techniques de numérisation permet la constitution de corpus de données bien plus importants dans le cadre d'enquêtes quantitatives menées par des équipes de taille réduite. Il souligne également le fait que cette extension des corpus – rendue possible par la numérisation – soulève de nouveaux problèmes de méthode et augmente le temps de travail consacré à la standardisation des données numérisées.

Keywords
Data entry, optical character recognition (OCR), prosopography, recoding data, EU lobbies

Mots clés
Saisie de données, Reconnaissance optique des caractères (OCR,), Prosopographie, Recodage des données, Lobbys européens.

Introduction

Les progrès des logiciels de reconnaissance optique des caractères (ROC ou OCR[1]) couplés avec les progrès de la numérisation standardisée, ouvrent depuis une quinzaine d'années de nouvelles perspectives pour le travail prosopographique en sciences sociales. En effet, ils rendent envisageables un codage accéléré des données qui ne peut que laisser songeurs les chercheurs qui ont dû, par le passé, compiler, une par une, un nombre important de fiches cartonnées dans le but de se représenter la morphologie sociale de telle ou telle population enquêtée. Par le simple fait que des extraits entiers d'annuaires ou de sources aujourd'hui épuisés soient ainsi rapidement convertibles dans un format compatible avec un traitement de texte, le temps de saisie et de mise en forme de données extraites de bottins mondains ou administratifs, d'annuaires biographiques ou même d'équivalents du *Who's Who* semblent pouvoir être en partie réduits.

Les premiers articles portant sur l'utilisation de ces nouvelles technologies en sciences humaines et sociales furent plutôt enthousiastes. Bien vite, l'OCR fut perçue comme le premier pas d'une méthode d'entrée des données sans clavier (« non keyboard data entry method ») ou une méthode de codage automatique et assistée par ordinateur (« computer-assisted and automatic coding » ; voir Dekker, 1994) susceptible de faire gagner beaucoup de temps aux chercheurs.

Au-delà des questions juridiques liées à la question de la propriété intellectuelle des données ainsi collectées (Isnard, 2001 ; Fréchon et al., 2010), la technique a cependant très vite soulevé des problèmes techniques.

Une des critiques les plus détaillées vis à vis de l'utilisation de l'OCR fut adressée par Mark Olsen et Alice Music McLean suite au projet ARFTL[2] visant à analyser après numérisation un grand nombre de textes classiques en littérature. Leur article paru dans *Computers et Humanities* en 1993 est sans appel : l'utilisation de l'OCR dans un projet collectif en SHS serait plus chère que d'autres méthodes et supposerait un travail de

correction post-scan très important pour corriger les fautes de reconnaissance. Celui-ci ne pourrait être accompli que par des spécialistes du sujet traité et disposant d'un haut degré de compétence. Plus encore, l'ensemble du projet pourrait même selon eux être conduit tout aussi efficacement par le recrutement de vacataires entrant les textes directement au clavier (Olsen et Music McLean, 1993).

Sans aller aussi loin dans la critique de la technologie OCR (qui a notamment connu d'importants progrès techniques ces dernières années), l'ambition de cet article est de souligner deux problèmes pratiques qui continuent de se poser à tout chercheur en sciences sociales lorsqu'il a ainsi recours à un traitement en partie automatisé de données biographiques et ce quelles que soient les sources initiales mobilisées.

Le premier de ces problèmes que nous souhaiterions traiter est celui de l'extension démesurée des corpus susceptibles d'être ainsi collectés. En effet, dans la mesure où le traitement est désormais automatisé, la tentation est grande d'intégrer à l'enquête une masse de données plus importante que par le passé. Cela est bien sûr une chance car ces corpus exhaustifs permettent d'aller plus loin dans l'analyse et notamment de procéder à des enquêtes comparatives à plus grande échelle. Mais notre expérience prouve qu'il est aussi fort probable que le temps gagné sur la saisie (grâce au traitement automatisé que permet l'OCR) soit en grande partie compensé, voire dépassé, par le temps passé à traiter des masses de données bien plus imposantes que lorsque la collecte était freinée par la perspective peu enviable d'un traitement uniquement manuel des données.

Le second problème que pose selon nous le recours à une telle technique automatisée des traitements de données est celui de la standardisation des données ainsi collectées et notamment la question de leur traitement uniforme alors même qu'elles peuvent provenir de plusieurs sources (plusieurs annuaires par exemple). L'utilisation de logiciels libres permet, nous le verrons, de résoudre en partie ces problèmes, mais quoi qu'il en soit la standardisation et l'articulation des sources entre elles restent coûteuses en temps.

Afin d'aborder ces deux problèmes pratiques, nous mobiliserons l'expérience d'une enquête quantitative sur les lobbyistes bruxellois menée à partir de plusieurs sources susceptibles de faire l'objet d'un traitement prosopographique : notamment quatre annuaires édités par la Commission européenne à compter de 1960 et jusqu'en 1986, ainsi que trois annuaires professionnels (équivalents locaux des *Who's Who*) qui recensent les différents groupes d'intérêt officiant dans la capitale belge.

Mené dans le cadre d'une enquête collective, le projet « Pressure »[3] vise à rendre compte à la fois de la structuration des intérêts économiques au niveau européen et de la nature des relations qui se nouent chaque jour entre les institutions européennes et les fédérations patronales à prétention européenne. Un travail de numérisation et de mise en résonance de différents annuaires a ainsi donné lieu à la constitution d'une base de données recensant la plupart les fédérations professionnelles européennes (ces structures que les lobbyistes désignent sous l'anglicisme d'« European business and trade associations »).

Le changement d'échelle permis par la numérisation des données

Suivant la méthode préconisée par Lemercier et Zalc (2008), la base de données « Pressure » repose sur une architecture en différentes tables structurées autour de deux tables

mères – « individus » et « organisations » – qui sont reliées à chaque nouvelle table introduite dans la base de données. En mettant en concordance ces différents annuaires, elle rend compte de la transformation historique de ces groupes d'intérêt et du profil des lobbyistes au gré de l'enracinement des institutions de l'Union Européenne.

Grâce aux possibilités offertes par l'OCR, la base de données « Pressure » confronte plusieurs sources et notamment plusieurs annuaires d'origine différentes. Plusieurs de ces annuaires sont des annuaires officiels édités par la Commission européenne depuis 1960 (sur l'histoire de ces annuaires, voir Courty et Michel, 2011) qui répertorient l'ensemble des fédérations professionnelles (lobbys européens) ayant des relations régulières avec les institutions. Sous la forme de tableaux importés sous le logiciel Access, la base « Pressure » contient notamment les données *in extenso* des annuaires suivants:

- Un annuaire de 1960 de 526 pages[4] répertoriant les noms et coordonnées de 136 lobbys européens, les noms, fonctions et nationalités de 260 dirigeants de ces structures et les noms et coordonnées de 884 structures membres (entreprises ou fédérations patronales nationales) qui financent ces lobbys européens.

- Un annuaire de 1973 de 690 pages[5] répertoriant les noms et coordonnées de 284 lobbys européens, les noms, fonctions et nationalités de 590 dirigeants de ces structures et les noms et coordonnées de 2.209 structures membres (entreprises ou fédérations patronales nationales).

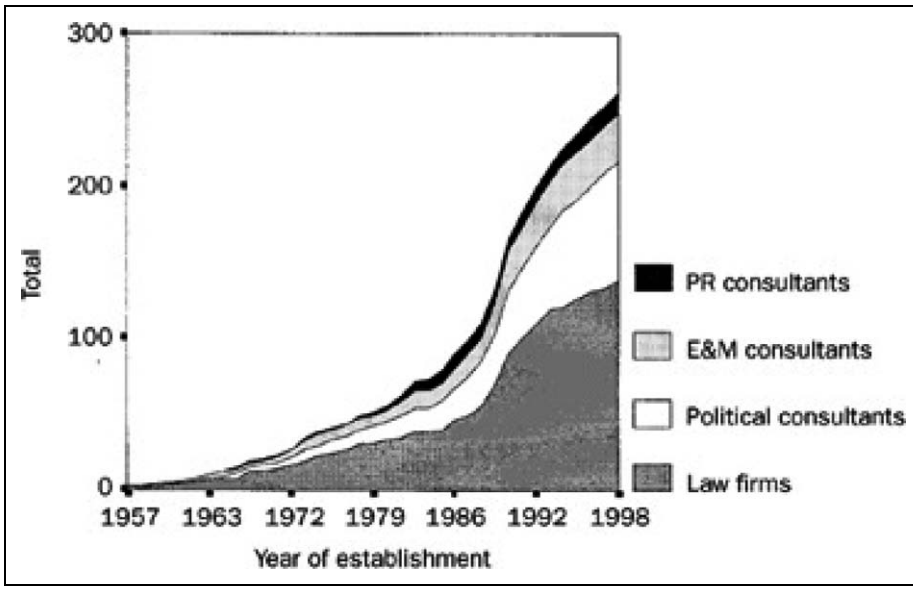
- Un annuaire de 1980 de 750 pages[6] répertoriant les noms et coordonnées de 464 lobbys européens, les noms, fonctions et nationalités de 952 dirigeants de ces structures et les noms et coordonnées de 3.505 structures membres (entreprises ou fédérations patronales nationales).

- Un annuaire de 1986 de 356 pages[7] répertoriant les noms et coordonnées de 541 lobbys européens, les noms, fonctions et nationalités de 982 dirigeants de ces structures et les noms et coordonnées de 4.330 structures membres (entreprises ou fédérations patronales nationales).

Malgré la magnitude de ces informations, grâce à la numérisation, une équipe de trois personnes a pu constituer ces bases en douze mois en travaillant à temps partiel sur ce projet. Comparé aux travaux antérieurs, ce volume important de données sur les lobbyistes européens laisse envisager, en début d'enquête, la possibilité d'apporter rapidement des réponses plus précises à des questions anciennes sur la morphologie des groupes d'intérêt européens.

Répondre en détail en brassant plusieurs sources - L'apport de l'OCR sur un sujet déjà traité

La littérature scientifique anglophone sur les groupes d'intérêt a déjà pu exploiter certaines des sources que nous mobilisons. Mais beaucoup de ces recherches exploitent ces sources de façon isolée et fragmentée. Ainsi C. Lahusen (2002) propose dans ses travaux



des courbes reprenant les catégories et les volumes de l'édition de 1999 de l'annuaire professionnel *Landmarks*. L'exploitation graphique et statistique de ce type de source laisse supposer une explosion des firmes de conseil et des lobbys entre 1986 et 1998. Mais elle ne repose en définitive que sur une seule édition d'un seul annuaire.

On perçoit d'emblée la limite de ce type de méthode : le graphique ci-dessus propose depuis une source actuelle une vision rétrospective de ce qu'a pu être le volume de création des cabinets de consulting ou des groupes d'intérêt dans le passé. Ainsi ne figurent pas sur le graphique les groupes d'intérêt créés en 1973 et qui auraient disparu ou fusionné avec d'autres entre 1974 et 1980.

Via l'OCR, la numérisation de ces annuaires *in extenso*, à différentes dates, laisse entrevoir la possibilité d'approfondir l'analyse de la représentation de l'évolution du lobbying à Bruxelles au-delà de ce que permet l'exploitation d'un annuaire contemporain. Elle permet de multiplier les « photographies » d'un même microcosme en utilisant pour chaque instantané une source contemporaine du « cliché » réalisé (au sens photographique) – un annuaire de 1960 pour décrire la situation de 1960 – et non des sources actuelles susceptibles de nous d'éclairer partiellement le passé. Elle permet également de poser la question des fondements sociaux de ces lobbys en permettant de recenser leurs adhérents nombreux et cotisant bien souvent à plusieurs structures.

Mais, on le pressent, un tel travail basé sur la numérisation massive d'annuaires épars suppose de croiser ces sources volumineuses afin de repérer quelle organisation subsiste d'un annuaire à l'autre. Il suppose la mise en résonance de ces annuaires et ouvre donc la voie à d'autres problèmes méthodologiques qui ne se posent pas lorsque l'on travaille sur une seule et même source.

Enjeux de la numérisation et mise en tableur d'annuaires

La phase de récupération de données oblige le plus souvent à « jongler » avec plusieurs logiciels de bureautique parmi les plus répandus en passant d'un navigateur Internet à un traitement de texte ou à un tableur. Le recours à tel ou tel logiciel dépend directement des caractéristiques des données sources dont nous allons évoquer les plus fréquemment utilisées. Pour parvenir à passer de documents imprimés à une base de données reprenant les informations qu'il contient, plusieurs étapes sont nécessaires: 1) leur numérisation ; 2) la création de fichiers dans un format reconnu par l'OCR (le plus souvent un pdf[8]) ; 3) la récupération des données dans un traitement de texte ; 4) leur transfert vers un tableur[9] ; 5) l'importation dans une base de données.

Les Différentes étapes de Traitement des Sources "Papier"

Techniquement, numériser un document s'apparente de plus en plus souvent aujourd'hui à en faire une simple photocopie. Il est, en effet, préférable d'exclure d'emblée le recours à un scanner ordinaire dans la mesure où le temps de copie de ces appareils est très long comparativement à celui désormais mis par les photocopieurs professionnels. Au lieu de proposer une sortie papier du document, on choisit sur ces derniers l'option de sortie "fichier pdf", un format directement exploitable par les logiciels d'OCR. Ainsi qu'Anne Permaloff et Carl Grafton (1992) le préconisaient, nous avons opté pour une qualité de numérisation optimale (au minimum 600 dpi) qui a pour seul inconvénient de proposer de très gros fichiers en sortie. Un logiciel OCR de bonne qualité optimisera cette qualité de récupération des données[10].

Les annuaires précités recensant les groupes d'intérêt européens ont donc été numérisés en suivant ces deux premières étapes. Le lancement du logiciel de reconnaissance de caractères permet de récupérer les chaînes de caractères sous plusieurs formats. Pour notre part, nous avons opté pour le traitement de texte. En effet, une constante caractérise les logiciels d'OCR performants: les chaînes de caractères reconnues sont importées sous un format « .rtf » mais – hélas – avec les principales caractéristiques de leur mise en page originale. Dans le cas de ces annuaires des groupes d'intérêt européens, l'unité de saisie de l'annuaire traité correspondait à une vignette autonome recensant une entreprise ou organisation (voir l'image ci-dessous).

Une fois reconnu par le logiciel d'OCR, le texte numérisé est mis en forme approximativement selon un modèle ressemblant et on observe les résultats suivants:

Benelux Economic Union
1000 Brussels, Belgium Email: info@benelux.be Website: www.benelux.be
Secretary-General
J P R M VAN LAARHOVEN
Deputy Secretary-General for Belgium
E BALDEWIJNS
Public Relations Officer
Karel VAN DE VELDE Tel: +32 2 519 38 30 k.vandevelde@benelux.be
Secretariat
Rue de la Régence 39
Tel: +32 2 519 38 11 Fax: +32 2 513 42 06

Benelux Economic Union
Secretariat
Rue de la Régence 39
1000 Brussels, Belgium
Email: info@benelux.be
Website: www.benelux.be
Tel: +32 2 519 38 11
Fax: +32 2 513 42 06

Secretary-General
J P R M VAN LAARHOVEN

Deputy Secretary-General for Belgium
E BALDEWIJNS

Public Relations Officer
Karel VAN DE VELDE
Tel: +32 2 519 38 30
k.vandevelde@benelux.be

Outre la bonne qualité de reconnaissance des données, on remarque plusieurs éléments: si le texte est importé au format enrichi (caractères gras, soulignement, lien html[11], majuscules/minuscules), l'ordre des informations n'est pas systématiquement respecté (lignes) et certaines lignes sont regroupées (on passe de 17 lignes dans la vignette à 11). Le temps gagné à ne pas saisir au clavier les noms des structures peut en partie être perdu par le temps passé à essayer de remettre en forme ce texte numérisé afin qu'il soit exploitable sous la forme d'un tableur. Pour ce faire, les chercheurs les plus expérimentés peuvent essayer de mettre en place des macros d'importation qui repéreront les mises en page habituelles et faciliteront leur transposition au format tableur. Mais il peut arriver d'une part que l'on ne dispose pas de ces compétences et d'autre part que les décalages ou la mise en forme produite par l'OCR ne soient pas réguliers et ne puissent être soumis à une standardisation de ce type.

Du Traitement de Texte au Tableur - Le Séquençage

A cette phase du travail, il convient souvent de prendre deux types de décisions : l'une portant sur les éléments à conserver (selon la nature des données) ; l'autre sur la manière dont les données sont structurées à la suite du processus d'OCR (selon leur structure).

Une opération délicate mais décisive va consister à découper avec un traitement de texte les champs que nous souhaitons conserver et à éliminer ceux qui ne présentent aucun intérêt. Dans notre cas, avant de pouvoir importer le produit de l'OCR sous un tableur, il a d'abord bien souvent été nécessaire d'insérer des marqueurs de tabulation (via la fonction rechercher / remplacer par le caractère « ^t »[12]).

Il est possible de découper de cette façon de grandes quantités de texte (plusieurs centaines de pages à la fois) car ces séquences de découpage ne varient pas au sein

d'un même document. Seules des erreurs de saisie initiales perturbent les requêtes (ex: « Email: » et « Email: » où un espace sépare les deux points dans le second cas). Une fois découpé, le texte peut être copié-collé dans le tableur en tenant compte du fait que les informations seront disposées en ligne et les variables en colonne. Il importe d'organiser la feuille de calcul en réservant la première ligne à l'identification des noms des champs (ex: nom de l'organisation, code postal, ville, etc.) et de copier-coller une à une chaque unité statistique (ici une vignette) dans le tableur. On obtient des informations sur plusieurs lignes et plusieurs colonnes (les tabulations présentes ou ajoutées sous Word) qu'il convient de couper-coller dans les bonnes colonnes. Cette phase d'affectation des informations dans les colonnes correspondantes du tableur reste particulièrement routinière et fastidieuse. En même temps, la régularité d'organisation et de mise en page des données dans le document source facilite des routines de transfert qui restent préférables, plus rapides et beaucoup plus fiables qu'une nouvelle saisie des informations.

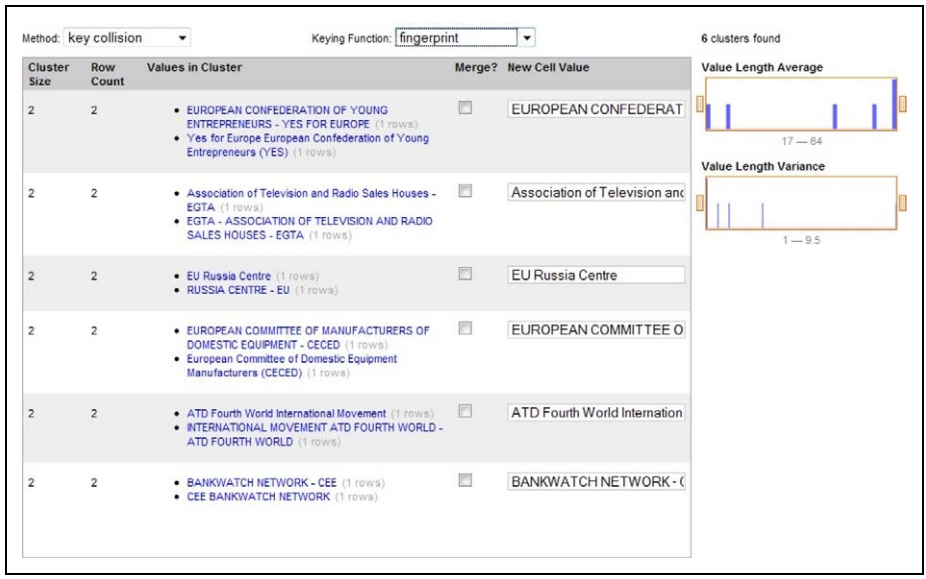
Il arrive parfois que certaines présentations exigent moins de manipulations ou des procédures différentes. Par exemple, les logiciels d'OCR respectent une présentation des données dans un document initial de plusieurs colonnes, et remplacent parfois des espaces importants dans la mise en page par des tabulations qu'il devient inutile d'insérer. Il est alors possible de copier-coller directement les textes ainsi balisés par l'OCR dans un tableur et de récupérer les informations telles qu'elles apparaissent à l'écran. Autre cas rencontré: lorsque les champs sont organisés en lignes sous Word, on sélectionne les lignes, on copie, bascule vers le tableur, option « Transposé » de la fonction « collage spécial »: les données pivotent à 90 degrés et se présentent sur une ligne. Il suffit de vérifier la correspondance entre les contenus copiés et les identifiants des colonnes.

Nous avons procédé ainsi pour cinq annuaires et identifié l'existence de plus de 6.000 organisations, 11.500 lignes cumulées pour une quarantaine de champs en moyenne. Le « rythme de croisière » de cette phase nous permettait de transférer ainsi une centaine d'unités par heure (pour une vingtaine de champs/colonnes). Le recours au copier-coller, les allers-retours entre tableur et traitement de texte, la fonction rechercher-remplacer gérés par les raccourcis claviers, pour autant laborieux qu'ils puissent paraître, protègent des inévitables erreurs de saisie en s'incorporant dans une gestuelle qui s'acquiert rapidement.

On le voit, les procédures d'extraction et de manipulation des informations que l'on souhaite récupérer en vue d'une exploitation ultérieure dépendent de leur présentation initiale et de la manière dont les logiciels d'OCR les interprètent. Il est rare qu'une importation massive des résultats d'une requête soit possible, et dans ce cas les procédures de balisage décrites ci-dessus sont nécessaires à effectuer. Si nous sommes confrontés plus souvent à des notices individuelles qu'à des listes structurées « prêtes à l'importation », seul un travail minutieux, ingrat mais indispensable permet de les organiser afin de permettre leur exploitation à grande échelle.

Faux Voisins et Uniformisation des Dénominations

Le travail de repérage des coquilles ou des erreurs de reconnaissance de caractère ne soulève pas que des enjeux formels. En effet, lorsque vient l'heure de croiser les annuaires pour rendre compte de l'évolution du nombre de groupes d'intérêts, la variation d'un seul caractère peut entraîner une double comptabilisation de la structure. Cela est inévitable



compte tenu de la taille des données si l'on procède uniquement à des juxtapositions de ces différents tableaux obtenus à travers un traitement manuel. Cela est notamment le cas dans certains travaux déjà publiés qui mobilisent plusieurs annuaires. Ainsi si on soumet le fichier source au format tableur utilisé par F. Baumgartner, C. Mahoney et J. Berkhout dans leur article paru dans *European Union Politics* (Baumgartner et al., 2010), à ce simple test de proximité sémantique sous le logiciel libre Gridworks, plusieurs cas de faux doublons apparaissent.

L'exemple ci-dessous est très parlant. Le groupe d'intérêt nommé « Association of television and radio sales houses - EGTA » figurant dans un des annuaires mobilisés par l'équipe de Baumgartner et al. est considéré dans leurs décomptes comme une organisation distincte de « EGTA – ASSOCIATION OF TELEVISION AND RADIO SALES HOUSE –EGTA » donnée dans un autre annuaire. Si chaque source donnée n'est pas mise en balance avec la source précédente et si chaque écart sémantique est enregistré comme la création d'une structure supplémentaire, on comprend alors qu'il est d'autant plus probable que les courbes historiques donnent dans bien des travaux des représentations toujours plus exponentielles du nombre de lobbys sur Bruxelles. Et ce d'autant plus que les noms figurent parfois d'un annuaire à l'autre dans une langue différente (anglais, français, néerlandais), précédé ou non de leur acronyme (sans compter les limited, ou ltd pour les entreprises anglaises, les « e.V. » pour « eingetragener Verein » qui désignent la forme associative des structures allemandes). Il arrive aussi que la simple suppression ou le simple ajout d'un article (« de la », « the », « für », ou même la typographie (absence ou présence d'un accent, de cédilles comme dans les ö, ô, ñ, ä, ó, à, ç, ô, ò pour les noms allemands, portugais, norvégiens . . .) produise de nouvelles structures non repérées comme identiques lorsque l'on croise les différentes tables dans la base de données.

Capture d'écran des résultats donnés par le logiciel Gridworks sur le fichier source mobilisé par l'équipe de Baumgartner, Mahoney et Berkhout sur les annuaires *Landmarks*. Résultats obtenus pour un seul filtre utilisé (« fingerprint »).

Le logiciel libre Gridworks (désormais « mis à jour » par Google sous le nom de Google refine) permet d'éliminer un grand nombre de ces erreurs. En repérant les chaînes de caractère, il rend possible la standardisation des données collectées via la numérisation et l'OCR. En soumettant le tableur à une série d'algorithmes, il est alors possible de repérer les erreurs de clavier possibles, les noms qui sonnent de façon proches et qui peuvent donc être potentiellement des chaînes de caractère identiques. Gridworks rend par exemple possible l'utilisation de l'algorithme metaphone[13] qui repère les cellules aux sonorités proches.

Mais il ne peut être d'aucun recours pour repérer les organisations dont le nom change sensiblement au cours du temps. Alors la standardisation des données collectées ne peut faire l'économie d'une intervention manuelle visant à repérer les continuités de structure devant les changements de dénomination. C'est là que le temps gagné via la numérisation semble subitement être compensé par le temps important de navigation dans les différents fichiers, nécessaire à la reconstitution des filiations historiques entre les différents groupes d'intérêt. Mais ce temps supplémentaire n'est pas (ou n'est plus) lié aux erreurs induites par l'utilisation de la technologie OCR, il est lié aux possibilités offertes par ces techniques qui rendent possible des comparaisons poussées dans le temps.

Les Conséquences non Prévues du Recours à l'OCR

De nombreuses fédérations patronales européennes voient, en effet, leurs noms évoluer au fil du temps. Face à ce type d'évolution, y compris Gridworks n'est d'aucun secours et il n'existe pas un nombre infini de méthodes pour pouvoir suivre l'évolution dans le temps de chacune de ces structures. Dans la base de données « Pressure », chacun des groupes d'intérêt étudié se voit attribuer un identifiant fixe dans le temps qui est, dans un second temps, relié aux occurrences nominales changeantes dans chacun des annuaires (l'organisation est comptée une seule fois si son domaine d'activité est le même, son adresse identique, ses dirigeants et surtout ses adhérents relativement stables). C'est là que l'extension démesurée des corpus rendue possible par la numérisation des annuaires a des effets sur le travail de recherche. Tout a fonctionné comme si le temps gagné en amont au moment du codage des données devait être réinvesti en aval dans le croisement de ces différentes tables.

On retrouve là la critique émise par Olsen et McLean sur la nécessité d'un travail de spécialiste afin de retraiter les données produites via l'OCR, mais à la différence près que ce travail n'est plus tant généré par les erreurs de reconnaissance de caractères (aujourd'hui rapidement repérées par Gridworks) que par la masse des données traitées.

La réduction obtenue par le croisement du logiciel Gridworks et de cette méthode de repérage est cependant significative. Comme l'illustre le Tableau 1 ci-dessous (qui ne recense que les fédérations à prétention européenne), loin d'être cette progression toujours plus exponentielle, la courbe des créations des lobbys sur Bruxelles entre 1960 et 1980 présente dès lors un autre visage.

Tableau 1. Au-delà des changements de noms - Un renouvellement partiel des organisations

	Annuaire 1960	**Annuaire 1973**	**Annuaire 1980**	**Annuaire 1986**
Fédération européenne	136	284	464	541
Nom strictement identique par rapport à annuaire précédent (après nettoyage sous Gridworks)	/	69	140	353
Léger changement nominal (article manquant par exemple non repéré par Gridworks)	/	2	6	22
Changement de nom mais même structure	/	44	81	69
Véritable création nette depuis annuaire précédent	/	169	237	97

Avec cette méthode, on n'obtient plus de courbe continue et exponentielle. Le nombre de créations de structures entre 1980 et 1986 est même moins important qu'entre 1973 et 1980. Dans la mesure où les possibilités de modifications de noms se répètent d'annuaire en annuaire, sur la durée une approche purement nominaliste et centrée sur les noms « mot à mot » des organisations ne permettrait pas de rendre compte du fait que 82,1 pour cent des fédérations professionnelles européennes recensées par la Commission en 1986 étaient déjà là en 1980, que 48,9 pour cent des organisations de 1980 étaient déjà installées en 1973... On ne pourrait rendre également compte du fait que plus d'un quart de ces groupes d'intérêt se vieillissent d'annuaires en annuaires, annonçant des dates de création en moyenne toujours plus anciennes afin de renforcer leur légitimité.

L'autre intérêt de l'OCR est qu'elle permet de comparer la liste des adhérents de ces structures et de montrer l'existence de multi-appartenances non négligeables masquées tant au niveau des individus (Couppié et Demazières, 1995) que par fois des organisations[14]. Il permet aussi, couplé à des logiciels d'analyse des séquences (Dijkstra, 1994), de repérer les carrières types opérées par les lobbyistes d'une organisation à l'autre (c'est là un moyen de mener une analyse de réseaux toujours compliquée à mettre en œuvre sur ce type de population ; voir Van Meter, 1987). Ces questions de multi-appartenance sont d'importance dans les débats sur les lobbys européens : derrière ce nombre grandissant de lobbys à Bruxelles ne retrouve-t-on pas pour autant une stabilité des intérêts représentés ? Sur ce point précis, l'OCR permet de répondre à des questions auxquelles on ne pouvait répondre précédemment qu'au prix de la mobilisation d'un collectif de recherche démesuré (Courty et al., 2011).

Conclusion

Sur ce sujet spécifique, la numérisation des annuaires et leur mise en concordance a donc permis une meilleure mise en historicité des institutions étudiées. L'OCR permet de se donner les moyens d'une analyse processuelle qui ne se limite pas à retracer le passé de ces organisations à partir de dates de création des lobbys telles qu'elles figurent dans les annuaires les plus contemporains. Elle permet de ne pas être dépendante d'une seule source et de multiplier les points de vue sur un même objet en intégrant dans l'analyse

la question des multi-appartenances des entreprises à plusieurs fédérations patronales européennes.

Si l'utilisation de l'OCR permet donc *in fine* un gain épistémologique dans le domaine de la sociologie des groupes d'intérêt européen, elle se traduit néanmoins par l'exploration de nouvelles difficultés de méthodes et celles-ci ne sont pas tant liées à la question des corrections « post-scan » comme l'avançait la littérature des années 1990 qu'au travail nécessaire pour articuler ces corpus de données entre eux et renvoie aussi – question qui mériterait d'être traitée en soi - à la difficulté d'articuler ces masses de données avec d'autres formes de méthodes d'enquêtes plus qualitatives.

En d'autres termes, si l'OCR a permis sur ce terrain particulier de la sociologie des groupes d'intérêt européen de multiplier par cinq le nombre de données comparativement à la plupart des autres enquêtes sur ce sujet, elle soulève aussi d'autres problèmes qui ont le mérite de rappeler que toute extension d'un corpus a un coût en temps non négligeable dès lors que toute opération « manuelle » s'effectue désormais sur des volumes importants. Sous cet angle, la limite des ambitions de recherche pour une petite équipe ne se situe plus tant aujourd'hui au niveau de la taille des corpus mobilisables qu'au niveau de leur articulation interne ou de leur mise en relation avec d'autres sources. Loin de nous rapprocher du mythe de l'automaticité totale du travail quantitatif, l'amélioration des techniques de reconnaissance automatiques des caractères nous renvoie alors à la nécessité d'intégrer toujours plus en amont une réflexion sur les objectifs de la recherche et sur l'architecture requise pour permettre le croisement de données toujours plus facilement mises en série sous forme numérique.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. Reconnaissance optique de caractère ou Optical character recognition.
2. http://artfl-project.uchicago.edu/
3. Projet de recherche sociologique sur les représentants d'intérêts européens (http://projetpressure.blogspot.com/)
4. *Répertoire des organismes communs créés dans le cadre de la Communauté économique européenne par les associations industrielles, artisanales et commerciales des six pays (1960) Bruxelles: Service des publications des Communautés Européennes.*
5. *Répertoire des organismes communs créés dans le cadre des Communautés européennes par les Associations industrielles, artisanales, commerciales et de service des six pays; associations de professions libérales; organisations syndicales de salariés et groupements de consommateurs (1980) Bruxelles: Service des publications des Communautés Européennes*.
6. *Répertoire des organisations professionnelles créées dans le cadre des communautés européennes* (1980), Bruxelles: Editions Delta.
7. *Répertoire des organisations professionnelles de la CEE* (1986), Luxembourg: Office des publications officielles des Communautés européennes, Luxembourg: Editions Delta.
8. Le « Portable Document Format » d'Adobe peut être aujourd'hui lu par différents logiciels gratuits, y compris sous Linux.

9. Le recours à un tableur est justifié par le fait qu'il possède des formats d'exportation directement lisibles par des logiciels dédiés à la gestion des bases de données (de type Access).
10. Nous utilisons Abbyy FineReader 9.0 Professional Edition qui permet la reconnaissance de plusieurs langues, un détail important s'agissant d'annuaires avec les coordonnées de plusieurs pays de la Communauté européenne.
11. Pour notre part, nous utilisons l'option « collage spécial » et « texte sans mise en forme » pour désactiver les liens Internet.
12. Ce caractère est dit « non imprimable », tout comme un « retour chariot » en fin de ligne.
13. Hélas Gridworks utilise l'algorithme en anglais principalement. Pour un aperçu de l'intérêt de l'algorithme metaphone et son histoire, voir http://en.wikipedia.org/wiki/Metaphone.
14. Ainsi VDMA, la puissante fédération allemande des machines-outils, cotise à près de 21 fédérations patronales européennes en 1986.

Bibliographie

Baumgartner FR, Mahoney C and et Berkhout J (2010) Measuring the Size and Scope of the EU Interest Group Population. *European Union Politics* 11: 463.

Couppié T and et Demazière D (1995) Se souvenir de son passé professionnel - Appel à la mémoire dans les enquêtes rétrospectives et construction sociale des données. *Bulletin de Méthodologie Sociologique* 49: 23-57.

Courty G, Laurens S and et Michel H (2011) Looking Inside the European Blackbox of Interest Group: Relative Stability of Interest Groups, Increasing Numbers of Organizations and Mailboxes in Brussels. Reyjavik: Interest Organizations in Europe, ECPR 2011 (Panel 173).

Courty G and et Michel H (2011) Groupes d'intérêt et lobbyistes dans l'espace politique européen - Des permanents de l'eurocratie. In Georgakakis D (ed.), *Le champ de l'Eurocratie*. Paris: Economica, 213-40.

Dekker AK (1994) Computer Methods in Population Census Data Processing. *International Statistical Review/Revue Internationale de Statistique* 62(1): 55-70.

Dijkstra W (1994) Computers/Ordinateurs: Sequence - A Program for Analysing Sequential Data. *Bulletin de Méthodologie Sociologique* 43: 134-44.

Fréchon I, Issenhuth P and et Vivier G (2010) Concilier les droits de chacun - Une éthique en dynamique. Enquête auprès de mineurs « protégés ». In Laurens S and et Neyrat F, *Enquêter - De quel droit ? Menaces sur l'enquête en sciences sociales*, Bellecombe en Bauges : Editions du Croquant, 187-209.

Isnard M (2001) Statistiques et libertés individuelles - Les apports récents de la loi. *Courrier des statistiques* 113-114 : 9.

Lahusen C (2002) Commercial Consultancies in the European Union: The Shape and Structure of Professional Interest Intermediation. *Journal of European Public Policy* 9(5): 695-714.

Lemercier C et Zalc C (2008) *Méthodes quantitatives pour l'historien*. Paris : La Découverte (coll. Repères).

Olsen M et Music McLean A (1993) Character Scanning: A Discussion of Efficiency and Politics. *Computers and the Humanities* 27(2): 121-27.

Permaloff A and et Grafton C (1992) Optical Character Recognition. *PS: Political Science and Politics* 25(3): 523-31.

Van Meter KM (1987) Ideology and Methodology: Network Analysis in the United States and France. *Connections* 10(2): 106-10.