La mise en place d'un répertoire d'entreprises en Palestine

P. Brion, N. Gharbi et JR. Suesser*

L'office statistique palestinien s'est vu confier en 2008 la présidence d'un comité national chargé de mettre en place un répertoire administratif d'entreprises. Cet article détaille les aspects techniques et organisationnels qui ont conduit à la mise en place de ce répertoire administratif partagé qui, bien qu'il présente des différences avec un répertoire statistique d'entreprises, pourra servir de base à un tel outil. Le projet coordonné par Expertise France a bénéficié d'un soutien financier accordé par la Direction générale du Trésor français et de l'expertise technique de l'Institut national de la statistique et des études économiques (Insee).

Contexte

La question de l'enregistrement des unités économiques par les administrations et de la tenue à jour de leur répertoire associé est une préoccupation commune à tous les pays. La Palestine présente des difficultés spécifiques dans la mesure où l'Autorité palestinienne n'a pas tous les attributs de la puissance publique sur l'ensemble du territoire, en particulier dans des zones contrôlées par Israël et sur lesquelles elle ne peut pleinement intervenir.

En 2008, le PCBS[1], Office statistique palestinien, s'est vu confier par une décision gouvernementale la présidence d'un Comité national chargé de mettre en place un répertoire administratif d'entreprises (ABR[2]) alimenté par les différentes institutions palestiniennes (administrations, chambres consulaires, etc.) et destiné à être partagé entre elles. L'objectif recherché avec ce répertoire était d'aider chacune de ces institutions à mieux enregistrer les unités concernées dans leurs champs de compétence. Ce mandat était particulièrement pertinent dans un contexte où les différentes institutions ont des difficultés à procéder à leur enregistrement et où le secteur informel est significatif.

Pour aider le PCBS à avancer dans la réalisation de ce mandat, les autorités palestiniennes ont sollicité, au cours de l'année 2011, l'assistance de la France pour un soutien d'ordre méthodologique[3], cela en raison de la notoriété de l'expérience de l'Insee dans le domaine des répertoires inter-administratifs (avec le répertoire Sirene). Plus précisément, c'est la Direction Générale du Trésor qui a mis à disposition du projet un financement de 0,7M€ dont la gestion a été confiée à Expertise France. Ce projet s'est achevé en décembre 2018.

Il ne s'agissait pas de chercher à transposer un modèle comme le modèle français, basé sur un contexte très centralisé (où un identifiant unique est attribué à toute unité qui se crée, puis transmis aux différents partenaires institutionnels et partagé avec eux), mais de s'adapter à la réalité du terrain, tout en s'inspirant de certaines pratiques françaises. L'objectif était d'avoir un produit qui aide chacune des administrations dans leur travail d'actualisation de la connaissance des unités qu'elles devraient connaître, et aussi de la qualité des valeurs des variables qu'elles enregistrent pour savoir quand, comment et où avoir les relations nécessaires avec les unités qu'elles suivent.

Démarche suivie

Une première étape a consisté à identifier, outre le PCBS, quatre types d'institutions palestiniennes principales ayant le plus d'éléments utiles dans leur répertoire pour la mise en place d'un premier répertoire partagé à partir des informations qu'elles enregistraient. Les contacts ont été noués avec le ministère de l'Économie nationale et celui des Finances et du Plan, ainsi qu'avec la Chambre de Commerce nationale et une dizaine de grandes municipalités[4].

[1] *Palestinian Central Bureau of Statistics.*

[2] *Par la suite le sigle ABR sera utilisé, comme Administrative Business Register.*

[3] *Pour la partie française, le projet a été coordonné par Expertise France et l'Insee a apporté son expertise technique.*

[4] *Il n'y a pas de système de sécurité sociale à l'heure actuelle en Palestine. S'il avait existé, l'organisme gestionnaire aurait fait partie des partenaires du projet.*

* Philippe Brion a été l'expert référent de l'Insee pour le projet d'assistance au PCBS philippe.brion55@gmail.com, Jan Robert Suesser a été le responsable du projet jrsues@wanadoo.fr, et Nadra Gharbi a été l'experte principale pour le projet g_nadra@yahoo.com

Chacun de ces partenaires disposait d'une liste d'unités établie pour ses propres besoins. Ces listes n'étaient pas identiques, ni en termes de champ couvert, ni de variables suivies, ni de qualité de la collecte.

Le PCBS s'appuyait pour sa part sur ses recensements quinquennaux des établissements[5], et s'apprêtait alors à lancer un recensement en 2017.

Dans le cadre de sa responsabilité pour l'élaboration d'un répertoire partagé, le PCBS a réuni l'ensemble des administrations et autres institutions pour présenter le projet. Chacune a été sollicitée pour faire le point sur ses acquis et sur ses difficultés à tenir son propre répertoire. L'objectif était de connaître les caractéristiques de chaque fichier avec les spécificités résultant des prérogatives propres de chaque administration. Cette confrontation des réalités a permis de soulever des questions de champ (quelles unités sont suivies par chaque institution, quelles variables sont importantes pour l'accomplissement de sa mission), et des questions liées aux actualisations des informations. Le partage de ces connaissances devait permettre d'adapter le contenu et le processus d'actualisation pour que chaque partenaire contribue et bénéficie du répertoire à construire dans une stratégie gagnant-gagnant pour l'ensemble des parties prenantes.

Répertoire inter-administratif versus répertoire statistique

Si l'on se place du point de vue d'un office statistique national, la mise en place d'un répertoire statistique d'entreprises est une thématique « classique », qui a donné lieu à de nombreuses publications (on peut par exemple se référer au manuel des Nations unies de 2015 : *Guidelines on Statistical Business Registers – United Nations Economic Commission for Europe*). Cette mise en place est en grande partie liée au besoin d'avoir une base de sondage pour tirer des échantillons d'entreprises afin de suivre l'activité économique, et une référence à la population globale qui permette de faire les redressements et extrapolations pour la production d'estimations statistiques. Le PCBS est dans ce cas, confronté, entre deux recensements économiques qu'il réalise, à des problèmes de mise à jour de la base issue du recensement pour tirer des échantillons représentatifs destinés à suivre l'activité économique avec un jeu d'enquêtes réalisées par sondage.

Avec le projet, on se situe dans une optique différente, même s'il existe de nombreux points de convergence. L'idée est de fabriquer un fichier d'unités caractérisées par quelques variables d'identifications pertinentes pour l'ensemble des utilisateurs du répertoire partagé ; même si les problèmes conceptuels posés par la mise en place de ce répertoire sont en grande partie semblables à ceux retenus pour les répertoires statistiques d'entreprises (problèmes d'unités, de couverture, de nomenclatures utilisées pour les différentes variables, entre autres), ainsi que les techniques de fabrication de ce fichier (techniques qui sont présentées plus loin), les règles d'inclusion d'une unité dans le répertoire devront être plus restrictives que dans le cas d'un répertoire statistique, au moins pendant la période initiale. Mais au final, le répertoire inter administratif mis en place constituera une base majeure pour un répertoire statistique.

Description des répertoires utilisés par les différents partenaires

Chaque institution partenaire du projet dispose d'un répertoire propre, constitué pour des besoins spécifiques qui se traduisent par des caractéristiques différentes : champ couvert par chacune, variables d'intérêt liées aux missions à réaliser, qualité de l'information et suivi de la mise à jour.

Le ministère de l'Économie (MoNE) enregistre toutes les unités légales, nommées « compagnies » qui sont concernées par des opérations d'importation ou d'exportation. Il leur fournit un identifiant spécifique (commençant par 56). Ce numéro est exigé par Israël pour toute marchandise transitant sur son territoire. Le champ du fichier ainsi constitué le rend quasi exhaustif pour les grandes entreprises. Cependant, ce fichier n'enregistre pas d'établissements (au sens localisations physiques de la « compagnie »). Souvent, l'adresse à laquelle la « compagnie » a été enregistrée n'est pas l'endroit où se situe l'activité de l'entreprise, mais celle d'un avocat ou d'une autre personne ayant fait la démarche d'enregistrement. De plus, la cessation d'activité d'une « compagnie » n'est pas suivie dans ce fichier puisqu'elle n'a pas de conséquence en termes de gestion pour le ministère. Le fichier mis à disposition par le MoNE contenait un peu plus de 25 000 enregistrements. Depuis 2007, l'Autorité palestinienne n'administrant plus Gaza, il n'est alimenté que pour la Cisjordanie. Un fichier de près de 3000 enregistrements pour Gaza a pu être exploité par ailleurs.

Le ministère des Finances dispose de plusieurs fichiers de contribuables, non rapprochés. L'objectif étant de collecter et de contrôler l'impôt, l'unité enregistrée est le contribuable, qui n'est pas nécessairement une entreprise ou un établissement (au sens de la localisation géographique de l'entreprise). Pour le projet, c'est le répertoire de collecte de la TVA qui a été choisi. Là non-plus, ce ne sont pas des entreprises ou des établissements qui sont l'unité suivie dans le fichier. Par exemple, un contribuable devant s'acquitter de la TVA pour plusieurs localisations a le choix entre apparaître dans un ou dans plusieurs enregistrements du fichier. Pour les « compagnies », ce fichier enregistre l'identifiant donné par le ministère de l'Économie. Pour les autres contribuables, il enregistre l'identifiant personnel (de la carte d'identité) du contribuable concerné, mais aussi un autre identifiant lorsqu'un contribuable-individu paye plusieurs TVA. La qualité des informations d'adresses des lieux d'activité des entreprises et établissements n'est pas assurée, et des duplications adviennent sans information permettant de les mettre en relation. Le fichier exploité pour la constitution du répertoire contenait près de

[5] *Lesquels constituent la base de sondage pour les enquêtes statistiques.*

30 000 enregistrements concernant la seule Cisjordanie. Un second fichier complémentaire d'environ 25 000 enregistrements n'a pas pu être traité dans le temps du projet d'assistance, mais constitue bien évidemment un matériau que le PCBS a exploité depuis la fin du projet.

Le fichier des chambres de commerce est pris en compte pour le projet : il est national, ce qui a l'avantage de ne pas introduire de doubles comptes ; en revanche, il n'enregistre ni les établissements, ni les différentes activités des entreprises et il n'est pas exhaustif. Le fichier mis à disposition contenait 23 500 enregistrements.

Les municipalités sont en relation pour de multiples raisons avec les établissements actifs sur leur territoire. En particulier, elles encaissent un impôt, annuellement. Dans les faits, la qualité de leur fichier est très hétérogène. Celles qui mènent un travail de terrain poussé possèdent une information très intéressante pour la constitution d'un ABR et son actualisation. Les adresses des établissements qu'elles suivent peuvent être considérées comme de bonne qualité. Parmi les variables qu'elles sont censées enregistrer, il y a l'identifiant du MoNE pour les « compagnies », et l'identifiant personnel (celui de la carte d'identité) du propriétaire, mais les municipalités ne l'ont pas de manière systématique. L'impôt collecté étant lié à l'activité exercée, un même établissement peut être dupliqué dans le fichier s'il exerce des activités dont les taux d'imposition sont différents. Si une entreprise dispose d'établissements situés dans différentes municipalités, on ne dispose pas dans les fichiers des municipalités d'information permettant de reconstituer le lien entre ces établissements.

Enfin, le PCBS dispose du fichier du recensement des établissements, qui est exhaustif au moment de la collecte (formel et informel, et recensé par un quadrillage du territoire). En raison du secret statistique, son utilisation pour la constitution du répertoire partagé nécessite de gérer des problèmes de confidentialité. Par exemple, une unité n'apparaissant que dans ce seul fichier ne pourra donner lieu à la création d'un enregistrement dans l'ABR. Cependant, ce fichier dispose d'atouts importants, entre autres pour choisir les valeurs des variables lorsqu'il faut choisir entre des valeurs différentes présentes dans les autres fichiers, et aussi pour les liens existants entre différents établissements qui peuvent ainsi être rattachés à une même entreprise (cette information a été collectée lors du recensement de 2017). Pour la Cisjordanie, le fichier contient près de 100 000 enregistrements ; et plus de 50 000 pour Gaza.

Création d'une version V1 de l'ABR

Comme indiqué *supra*, les partenaires n'immatriculent pas les mêmes unités (pour certaines des institutions partenaires, l'enregistrement « unitaire » peut être par exemple une « activité » réalisée par une unité économique dans un lieu donné), et il n'existe pas d'identifiant partagé par tous pour apparier aisément les enregistrements des différents fichiers disponibles[6].

Une première phase de recherche a porté sur les variables à utiliser pour mener des appariements entre les unités présentes dans les fichiers des partenaires, en passant en revue des variables classiques « référençant » une entreprise (nom, adresse, numéros de téléphone). La mauvaise qualité de la collecte de ces variables a conduit à considérer que les liens établis avec celles-ci étaient trop partiels pour aboutir à une couverture et une qualité acceptables d'un répertoire construit sur leur partage. Après avoir acquis une meilleure connaissance des contenus des fichiers, une autre optique a pu être adoptée, s'appuyant d'abord sur l'identifiant donné par le ministère de l'Économie pour les « compagnies » et sur l'identifiant personnel du propriétaire pour les autres unités économiques. Avec ce type de variable, l'appariement est fait ou pas (aux erreurs de saisie près), sans que les faiblesses de saisie des autres variables ne soient un problème. Par ailleurs, tous les partenaires sont intéressés par la collecte de ces variables, leur saisie est simple, et leur enregistrement progresse régulièrement.

Le constat fait sur les différents fichiers disponibles chez les partenaires a également conduit à travailler sur deux niveaux d'unités pour fabriquer un répertoire partagé :

- Un niveau « entreprise » (pertinent pour le ministère de l'Économie et plutôt pertinent pour celui des Finances et du Plan) ;
- Un niveau « établissement » qui constitue la déclinaison locale de l'entreprise (plutôt pertinent pour les municipalités). Une entreprise peut avoir un ou plusieurs établissements.

Le PCBS et, dans une moindre mesure, les chambres de commerce gèrent des informations utiles pour repérer les deux niveaux et les liens entre eux. Le lien entre les établissements d'une entreprise est un défi que l'ABR a commencé à relever avec succès.

La phase technique : nettoyage des fichiers en amont, algorithme (DUKE), validation manuelle

Pour créer le répertoire partagé, un processus de traitement des données a été mis en place pour permettre d'apparier les enregistrements des partenaires de façon à créer une ligne pour chaque unité active à intégrer dans le répertoire, cela sans duplication, et de créer les liens entre les enregistrements du répertoire partagé et ceux des fichiers des partenaires.

Une première phase de nettoyage des fichiers a principalement concerné un objectif : limiter les facteurs qui « artificiellement » accentuent les différences entre les valeurs d'une variable que l'on cherche à comparer.

Par exemple, des numéros de téléphone enregistrés dans des formats différents peuvent être considérés comme

[6] *Ceci est tautologique puisque, justement, le répertoire inter-administratif n'existe pas à ce stade.*

différents alors que, comparés selon un même format, ils seraient identiques. Le nettoyage a, entre autres, consisté à incorporer les indicatifs géographiques à tous les numéros de téléphone enregistrés dans les fichiers partenaires.

Un autre type de traitement a consisté à nettoyer les noms des propriétaires ou les noms des « compagnies ». Les signes de ponctuations ont été enlevés avant de comparer les données, les abréviations usuelles ont été homogénéisées, etc. Plusieurs centaines de traitements ont été mis en œuvre.

Après ces traitements, les chaînes de caractères comparés avaient été largement débarrassées de différences qui n'avaient pas de sens en termes de « différences signifiantes ».

La phase suivante, décrite ci-après, a considérablement gagné en pertinence après ces nettoyages.

Un autre objectif de nettoyage, poursuivi en amont, consistait en le repérage des duplications d'une même unité au sein d'un même fichier et, pour ce faire, en réalisant des comparaisons entre enregistrements des différents fichiers sans les polluer avec des appariements redondants. Les techniques utilisées ont été adaptées aux différents motifs qui produisent des duplications dans les processus même de gestion de l'information nécessaires à chaque institution partenaire.

Ensuite, comme indiqué *supra*, il est apparu *in fine* que deux variables aisément comparables étaient assez régulièrement prises en compte par les partenaires : le numéro d'identification donné par le ministère de l'Économie nationale pour les unités économiques qui importent ou exportent (l'essentiel des grandes unités – plus de 10 employés – ont un tel numéro) et le numéro de la carte d'identité du propriétaire que les partenaires essayaient d'avoir dans leurs fichiers. Ces deux variables sont aisément utilisables. Cependant, elles ne sont pas suffisantes pour repérer de façon unique une entreprise et un établissement : soit parce qu'elles ne sont pas toujours collectées, soit parce que la même valeur peut apparaître dans plusieurs enregistrements d'un même fichier ou pour des unités différentes présentes dans les fichiers de plusieurs partenaires.

Étant donné le volume de données des différents fichiers partenaires, il était d'emblée clair que le travail d'appariement devait combiner des outils de rapprochement automatisés et des outils de validation manuelle.

Cette approche en deux temps résulte du fait qu'une procédure de rapprochement des unités identiques ne pouvait pas se baser sur une comparaison exacte des données. En effet, les informations présentes dans les différents champs (pas uniquement les deux identifiants mentionnés ci-dessus) ne respectaient pas forcément une même orthographe (voire la même langue, avec l'utilisation de l'anglais en sus de l'arabe). En outre, des informations telles que les numéros de téléphone ou les identifiants ne sont pas exemptes d'erreurs de saisie et/ou de collecte et/ou de données manquantes (pour le téléphone des titulaires de ligne aux fonctions particulières), qui conduisent à des différences de saisie entre les fichiers des partenaires. De plus, un fichier pouvait contenir une même unité dupliquée (plusieurs établissements d'une même entreprise, mais aussi une même entreprise ou un même établissement ayant été enregistré successivement pour chacune de ses activités).

Pour la phase algorithmique, il y a donc eu recours à des approches de calcul de distances entre chaînes de caractères qui, au lieu de comparer les chaînes de caractères d'une façon exacte, calculent un indice de rapprochement de deux chaînes distinctes.

Une des méthodes de rapprochement choisie pour notre projet, est la méthode de Levenshtein qui calcule une distance entre deux chaînes de caractères et donne ainsi un degré d'appariement entre ces deux chaînes (voir en annexes).

Après avoir utilisé l'approche algorithmique sur les différents fichiers, un travail de validation manuelle a été appliqué aux résultats obtenus. Ce travail, destiné à s'assurer de la qualité de toute unité introduite dans l'ABR, a été long et a nécessité la mise en place d'une équipe dédiée pendant plusieurs mois. Pour le faciliter, une adaptation des résultats du logiciel Duke a été réalisée ; elle est décrite en annexes.

La fabrication de la première mouture de l'ABR

Une fois ces travaux d'appariement et de validation manuelle réalisés, plusieurs règles ont été appliquées pour l'introduction d'une unité dans le répertoire partagé, pour le lien entre cette unité et les enregistrements des fichiers des partenaires et pour le choix des valeurs des variables d'identification de cette unité.

En effet, lors de l'introduction d'une unité dans le répertoire partagé (entreprise ou établissement), plusieurs impératifs doivent être respectés, avec deux écueils principaux à éviter :

- Il ne faut pas créer de duplication d'un enregistrement existant. Autrement dit, ne pas introduire une unité déjà présente dans le répertoire partagé à partir d'un appariement antérieur fait avec des enregistrements dans d'autres fichiers, cela au prétexte que les deux appariements générateurs n'ont pas pu eux-mêmes être appariés ;
- Il ne faut pas introduire une unité qui n'est plus active.

Dans la phase de construction de la version initiale de l'ABR, c'est-à-dire celle réalisée à partir des fichiers des partenaires constitués depuis des années, il a été décidé que l'inclusion d'une unité dans l'ABR ne serait actée qu'à partir du moment où cette unité (entreprise ou établissement) est présente dans au moins deux fichiers différents. Cela diminue fortement le risque d'enregistrement d'une entreprise qui aurait été introduite via un autre fichier, et aussi le risque d'introduire une entreprise non active. Cependant, une fois la « double origine » constatée, les deux enregistrements peuvent correspondre à deux établissements différents, chacun dans le fichier d'un des deux partenaires. Ce point est expertisé

lors de la validation manuelle, d'abord, mais pas exclusivement, à partir des adresses disponibles.

Aussi, deux types d'unités ont été créés qui différencient le niveau « entreprise » et le niveau « établissement ». Un identifiant « ABR » est affecté à chaque unité créée, commençant par une lettre (L pour les établissements et E pour les entreprises) suivie de huit chiffres générés au hasard (sans signification, donc). Les liens entre les deux niveaux sont gérés dans le répertoire. Pour l'essentiel, c'est l'information collectée lors du recensement des établissements mené par le PCBS qui a pu être utilisée de façon fiable pour articuler les niveaux établissements-entreprises, les autres partenaires n'ayant pas d'information (ou pas d'information exhaustive) sur ces liens.

Afin de pouvoir plus aisément faire les futures mises à jour (nouvelles entreprises, établissements additionnels, unités jusqu'à présent non enregistrées, unités enregistrées mais dont le lien avec l'unité du répertoire partagé n'avait pu encore être établie), qui pourront être générées par le traitement des fichiers actualisés des partenaires, il a été demandé à ces derniers d'introduire un des deux identifiants « ABR » possibles dans leur fichier courant pour leurs enregistrements qui sont déjà connectés au répertoire partagé.

La première version produite du répertoire partagé pour la Cisjordanie contient 40 000 établissements[7], alors que le recensement qui couvre un champ assez comparable en contient près de 100 000. Ce nombre d'unités de l'ABR, sensiblement inférieur à celui du recensement, vient de la règle des deux origines qui a été appliquée pour initier l'ABR. Celle-ci a en effet conduit à ne pas introduire dans le répertoire partagé initial des unités qui seraient actives mais enregistrées par une seule institution partenaire sans avoir pu être appariées avec un autre fichier.

Les tableaux 1 et 2 donnent la répartition des entreprises (« compagnies » et « non compagnies »[8]) selon leur présence dans les fichiers des institutions partenaires à partir desquelles elles ont pu être appariées : force est de constater l'importance des manques pour chaque fichier des partenaires, que ce soit pour les « compagnies » ou les entreprises qui ne le sont pas, ceci soulignant l'apport de l'ABR pour la connaissance du domaine.

Par ailleurs, lors de contrôles de la qualité, on a pu évaluer à partir des données du recensement qu'à ce stade plus de 90 % des établissements de plus de 10 salariés ont été introduits[9]. Cela montre que si le répertoire administratif est loin de couvrir l'ensemble des établissements recensés, sa couverture des grandes unités est de bonne qualité.

Tableau 1

Nombre d'entreprises « compagnies » en Cisjordanie dans l'ABR V1 (fin 2018) selon leur présence dans les différents fichiers partenaires

Fichier du ministère de l'Économie	x	x	x	x	x	x	x	x	**14 041**
Fichier TVA (ministère des Finances)	x		x	x	x				**11 840**
Chambres de commerce	x	x		x		x			**4 206**
Municipalités	x	x	x				x		**3 707**
TOTAL ABR V1	**1 356**	**213**	**1 616**	**1 957**	**6 911**	**680**	**522**	**786**	**14 041**

Lecture : Dans l'ABR, 213 « compagnies » sont enregistrées à la fois dans le fichier du ministère de l'Économie, dans celui des chambres de commerce et dans un fichier d'une municipalité. Au total, 14 041 « compagnies » sont enregistrées dans l'ABR. Remarque : dans le cas où l'entreprise est indiquée comme présente dans le seul fichier du ministère de l'Économie dans le tableau ci-dessus, c'est la présence de l'entreprise dans le fichier du recensement du PCBS de 2017 qui a justifié son inclusion dans l'ABR.

Tableau 2

Nombre d'entreprises « non compagnies » dans l'ABR V1 (fin 2018) selon leur présence dans les différents fichiers partenaires

Fichier TVA (ministère des Finances)	x		x	x	x			**9 881**
Chambres de commerce	x	x		x		x		**2 363**
Municipalités	x	x	x				x	**11 355**
TOTAL	**1 002**	**4**	**3 429**	**1 319**	**4 131**	**38**	**6 920**	**16 843**

Remarque : Dans les cas où l'entreprise est indiquée comme présente dans un seul fichier, dans le tableau ci-dessus, c'est la présence de l'entreprise dans le fichier du recensement du PCBS de 2017 qui a justifié son inclusion dans l'ABR.

[7] Soit environ 31 000 entreprises.

[8] Nous appelons dans le texte "non compagnies" les entreprises qui n'ont pas le statut de "compagnies".

[9] Cependant, la variable taille n'est pas dans le répertoire partagé à l'heure actuelle puisqu'elle est uniquement présente dans l'opération statistique du recensement (et donc couverte par le secret).

Les variables contenues dans la V1 de l'ABR

Dix variables ont été retenues pour être incluses dans le répertoire partagé[10] : noms et identifiants, variables d'adresse, lieu du siège, année de création. Plus précisément :

- Nom commercial pour les « compagnies » (tel qu'enregistré par le ministère de l'Économie)
- Numéro d'identification des « compagnies » attribué par le ministère de l'Économie
- Nom du propriétaire pour les « non-compagnies » (tel qu'enregistré par tel ou tel partenaire)
- Numéro d'identité du propriétaire attribué par le ministère de l'Intérieur
- Gouvernorat (26 en Cisjordanie)
- Localité (plusieurs centaines en Cisjordanie)
- Rue
- Coordonnées GPS (telles qu'enregistrées par tel ou tel partenaire)
- Siège ou pas siège (cette information n'a pas de valeur légale ; elle correspond à des déclarations faites à tel ou tel partenaire)
- Année de création (pour les « non compagnies », l'information retenue correspond à celle reçue par tel ou tel partenaire – en général la plus ancienne ; pour les « compagnies », c'est la date d'enregistrement dans le fichier du ministère de l'Économie)

Ces dix variables permettent à chaque administration de prendre contact avec une unité de l'ABR qui n'est pas connectée à son propre fichier et vérifier si elle la connaît déjà (sans qu'elle ait pu être appariée) ou pas, et si elle relève de son champ d'activité. Le partenaire peut aussi vérifier des écarts d'information entre ce qui figure dans son propre fichier et l'information du fichier partagé. Ces dix variables ne font pas partie de celles pour lesquelles des règles de confidentialité s'imposent aux différents partenaires.

La mise en place d'un système pérenne

Le travail réalisé pour construire la version initiale du répertoire partagé a permis de franchir une étape importante de mise en commun, par les institutions partenaires, de leurs informations d'identification. D'une part, ces différentes institutions ont mis à disposition leurs fichiers de gestion en explicitant leur contenu (champ, unités, variables, mises à jour, etc.), donnant ainsi une vision opérationnelle du contexte dans lequel elles fonctionnent et de la qualité des données qu'elles utilisent. D'autre part, elles ont intégré l'importance de collecter de façon systématique les variables importantes pour l'appariement de leurs enregistrements, au premier rang desquelles les deux identifiants du ministère de l'Économie nationale et l'ID personnel de la carte d'identité.

Il faut noter que les deux identifiants spécifiques définis dans « ABR » (pour les établissements L et pour les entreprises E) n'ont pas vocation, en aucune façon, à remplacer l'identifiant propre de gestion de chaque institution partenaire, et cela d'une part parce que l'ensemble des unités que ces institutions couvrent ne se trouve pas forcément dans l'ABR et, d'autre part, parce que leurs unités ne sont pas toujours des entreprises ou des établissements (mais, par exemple, des contribuables pour le ministère des Finances).

En revanche, l'introduction de l'identifiant ABR dans les fichiers des institutions partenaires constitue une avancée réelle, dont toutes les parties prenantes de l'ABR pourront profiter pour les mises à jour de leur répertoire, tant pour les unités qu'elles n'enregistraient pas encore que pour avoir accès aux valeurs à jour des variables suivies dans l'ABR pour les unités qu'elles ont déjà.

La question est désormais de faire vivre l'ABR. Cela concerne l'introduction de nouvelles unités, la mise à jour des valeurs des variables suivies, et la prise en compte des informations sur une fin d'activité.

Pour ce faire, des règles d'actualisation doivent être définies, en particulier celles relatives aux échanges entre les parties prenantes : quel support utiliser (il est proposé dans un premier temps d'utiliser des solutions robustes, même si elles n'intègrent pas des procédures très automatisées), quel protocole d'échanges (avec les questions de sécurisation afférentes) et quelle fréquence[11] ? Mais plus que les échanges, ce sont les règles d'arbitrage à appliquer en cas de remontée d'informations contradictoires qui mériteront toute l'attention des équipes de travail. Le PCBS devrait tenir le rôle d'administrateur de ce répertoire inter administratif.

En parallèle, des améliorations au niveau de la couverture de l'ABR sont envisageables : d'une part, pour des unités non encore introduites, les travaux d'appariement peuvent aider à augmenter le nombre d'unités pouvant être retrouvées et connectées ; d'autre part, la sensibilisation faite sur l'importance d'une collecte de qualité des données pour les nouvelles unités enregistrées par les partenaires devrait faciliter à l'avenir les appariements automatiques et la validation manuelle, avec *in fine* des décisions fiables d'introduction dans le répertoire partagé. Il est envisageable, si les identifiants du MoNE et personnels sont systématiquement enregistrés, de pouvoir adapter l'actuelle règle des deux sources en la ramenant, pour plusieurs populations, à une seule source nécessaire.

Des discussions pourront également avoir lieu pour décider s'il est possible, ou pas, d'introduire de nouvelles variables

[10] On parle ici des variables partagées par l'ensemble des partenaires ; en revanche, il existe d'autres variables transmises par les partenaires uniquement destinées à la gestion de l'ABR (et donc à usage de son administrateur), comme par exemple les identifiants utilisés par chaque partenaire, qui n'ont pas vocation à être partagées à l'ensemble des partenaires.

[11] Sur ce sujet, si, comme il a été indiqué précédemment, il n'est pas question de transposer à l'identique le modèle français, la visite en France d'une délégation rassemblant des représentants de différentes institutions palestiniennes a cependant permis à ceux-ci de prendre connaissance des processus d'échanges concernant le répertoire Sirene.

dans l'ABR : ceci nécessite de se mettre d'accord sur le contenu de ces variables (par exemple, si on s'intéresse à une variable sur l'emploi, les types de personnes à prendre en compte) ainsi que sur la nomenclature à utiliser (si l'on veut par exemple introduire un code d'activité). Et enfin, surtout, sur la détermination du partenaire qui sera responsable de cette collecte et des outils pour son actualisation.

Par ailleurs, certaines catégories non prises en compte actuellement pourraient l'être dans le futur, par exemple des unités « non localisables » comme les taxis.

Enfin, ce projet donne au PCBS les bases de constitution d'un répertoire statistique d'entreprises qui pourrait avantageusement compléter le système existant qui, basé sur le fichier du recensement d'établissements, était jusqu'à présent difficile à mettre à jour en continu.

Références bibliographiques

United Nations Economic Commission for Europe (2015), "Guidelines on Statistical Business Registers".

Willenborg, L. and Heerschap, H. (2012), "Matching", Method Series 12, Statistics Netherlands.

Zeidan H. (2018), "Matching techniques and administrative data records linkage", *Statistical Journal of the IAOS*, No 34, pp. 599–603.

Les résultats du projet sont à mettre au crédit des collègues palestiniens qui se sont impliqués de multiples façons autour de la Présidente du PCBS, Ola Awad, et tout particulièrement de Saleh Al Khafri, Haleema Saeed, Esraa Abu Salameh, Haitham Zeidan, Musab Abu Baker, Osaid Alardah, Adham Dwikat, Amjad Jwabra, Husam Khalifa, Ayman Qanir, Rasha Sarrawi, Basheer Shbitah, Maher Tannous, Thaer Zagal. Ils doivent évidemment beaucoup à l'apport des collègues de l'Insee, en activité ou retraités, Jean-Marc Béguin, Franck Cotton, Hugues Picard, Éric Sigaud, Karim Tachfint, Constance Torelli ainsi qu'à une étudiante de l'Ensai, Chaïmae Baghdadi, et à l'implication d'Expertise France (Anaïs Abou-Hassira, Clara Delmon, Beata Suszterova).

Annexes

La distance de Levenshtein et le logiciel Duke

La distance de Levenshtein est une distance, au sens mathématique du terme, donnant une mesure de la différence entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer, pour passer d'une chaîne à l'autre [voir par exemple Willenborg et Heerschap, 2012]. Une fois choisie cette méthode de calcul de distance entre deux chaînes de caractères, a été déterminée une méthodologie d'appariement des différentes lignes des fichiers des partenaires entre elles. Cet appariement s'est appuyé sur la prise en compte de variables (données/colonnes des fichiers) qui permettent d'identifier un établissement/entreprise. La méthode la plus directe était de comparer pour chaque ligne d'un fichier considéré comme « base », les valeurs prises par les variables sélectionnées avec celles des autres fichiers. Ceci conduit à *(n1 x n2 x … x ni)* itérations pour chaque process d'appariement, avec *ni* le nombre de lignes pour le fichier partenaire *i*. Ce qui peut entraîner des temps de calcul élevés.

En pratique, il a fallu trouver un moyen efficace et rapide pour l'exécution du process d'appariement. Une des méthodes utilisées a consisté à indexer toutes les lignes des fichiers à apparier en fonction des variables sélectionnées pour faire la comparaison. L'appariement des lignes se base donc sur une recherche préalable des groupes de lignes ressemblant à la ligne du fichier pris comme base pour la comparaison et, ensuite sur l'application du calcul de la distance de Levenshtein. Des seuils fixés permettent de décider si les lignes peuvent être considérées comme ayant une bonne probabilité de correspondre à une même unité. De cette façon, nous passons d'une jointure itérative à une jointure par hachage dont la complexité est de *O(n)* où *n* est le nombre des lignes du fichier de référence.

Duke

Pour faire ce travail, le logiciel Duke a été utilisé. Duke est un outil Open Source d'appariement / dédoublement codé en Java et qui implémente toutes les fonctionnalités nécessaires (voir référence en bibliographie). Duke permet de réaliser : (i) L'appariement des données en se basant sur divers algorithmes (dont la distance de Levenshtein) (ii) L'indexation des données via la librairie Lucene déjà utilisée par plusieurs projets d'indexation et de recherche lexicale et/ou textuelle. (iii) La prise en compte de plusieurs formats de données y compris le CSV. Pour ce projet, c'était important puisque c'était un format utilisable pour les fichiers transmis par les partenaires.

Duke calcule des probabilités pour déterminer si deux enregistrements pourraient correspondre ou pas à une même unité, cela en se basant sur les informations des variables définies pour la comparaison. Duke part de l'hypothèse que deux lignes ont une chance sur deux pour coïncider (soit une probabilité de 0.5). Ensuite, il commence à affiner les probabilités pour chaque variable en prenant en considération des bornes inférieure et supérieure configurées par l'utilisateur. Finalement, Duke agrège ces probabilités en appliquant la formule de Bayes. La probabilité ainsi obtenue est ensuite comparée à un seuil défini dans la configuration faite par l'utilisateur. Lorsqu'elle est supérieure ou égale à ce dernier, les deux lignes sont considérées comme potentiellement « identiques » par Duke.

Pour illustrer cela, prenons le cas des variables des deux lignes suivantes prises dans deux fichiers :

Fichier	MONE_ID	COMMERCIAL_NAME	ADDRESS
1	569401147	Ettisalat falastine	Ramallah
2	569401147	Ettisalat watanya	Ramalah

L'algorithme de comparaison appliqué par Duke est le suivant :

- Comparaison des adresses

'Ramallah' ~ 'Ramalah' : les deux chaînes de caractères ne sont pas complètement identiques – peut-être une erreur de frappe – ce qui nous donne une valeur de la distance de Levenshtein de *0.875*.

Pour calculer la probabilité de correspondance, on applique la formule suivante :

*((ss – 0.5) * sim²) + 0.5, avec ss : le seuil supérieur et sim : la distance de Levenshtein*

Pour les adresses, nous prenons un seuil supérieur de 0.65 qui signifie que si les adresses coïncident parfaitement, on a une probabilité de 0.65 pour que les deux lignes coïncident à leur tour. La probabilité de correspondance pour les adresses vaut 0.6148.

Ceci nous donne une probabilité que les deux enregistrements correspondent à la même unité également de *0.6148,* après application de la formule de Bayes en partant de 0.5 comme probabilité initiale [12].

- Comparaison des noms commerciaux

Nous prenons 0.85 comme seuil supérieur pour les noms commerciaux.

'Ettisalat falastine' ~ ' Ettisalat watanya' : la distance de Levensthein valant 0.6315, la probabilité de correspondance vaut 0.6396.

Ce qui nous permet de passer d'une probabilité globale (à savoir la probabilité que les deux enregistrements correspondent à la même unité) de 0.6148 à 0.7391 [13].

- Comparaison des identifiants

Nous prenons 0.95 comme seuil supérieur pour les identificateurs.

'569401147' ~ '569401147' : 1 ; la probabilité de correspondance vaut 0.95.

Ce qui nous permet de passer d'une probabilité globale de 0.7391 à 0.9818.

Ainsi nous obtenons une « bonne » valeur de la probabilité, au-dessus du seuil défini, ce qui conduit à la conclusion que les deux lignes coïncident.

Une présentation adaptée des résultats de l'appariement pour la validation manuelle des liens proposés

L'un des enjeux majeurs pour que le process de travail soit adapté à la prise de décision sur les liens entre enregistrements résidait dans la présentation des données issues du traitement Duke, et en l'occurrence la disposition des lignes appariées. En effet, Duke produit un modèle plat de résultats ; en d'autres termes, une ligne dans le fichier résultat correspond à deux lignes appariées.

Cette présentation est pratique pour un appariement (1 – 1). Mais le travail de validation concerne ici des relations (1 - N) — où une ligne du premier fichier peut trouver plusieurs homologues dans le deuxième fichier — et des cas (N – M) — où une ligne du premier fichier trouve plusieurs homologues dans le deuxième fichier et une ligne de ce dernier peut s'apparier avec plusieurs lignes du premier fichier.

Exemple :

Prenons le cas d'un fichier à n lignes dont les lignes sont notées L1.1, …, L1.n. Et un deuxième fichier à m lignes dont les lignes sont notées L2.1, …, L2.m.

---

[12] *On applique la formule* $P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|\bar{B})P(\bar{B})}$

Où B est l'événement « même unité » et A correspond à l'information obtenue sur les deux adresses : P(B) est fixée au départ à 0.5, et P(A|B) est la probabilité qu'on ait ce résultat sur les adresses sachant que les deux enregistrements sont la même unité ; au vu des calculs réalisés, celle-ci vaut O.6148 ; $P(A|\bar{B})$ *est la probabilité que l'on obtienne ce résultat sachant que l'on a deux unités différentes, soit (1-0.6148).*

[13] *Même application de la formule de Bayes que précédemment, mais cette fois on part d'une valeur de P(B) qui a été réévaluée à 0.6148.*

Dans le cas où toutes les lignes du premier fichier sont en relation 1-1 avec le deuxième, nous aurons le résultat suivant dans Duke :

L1.1 ; L2.4

L1.2 ; L2.45 …

Dans le cas 1-N, nous pouvons avoir le résultat suivant (l'ordre des lignes résultat est aléatoire) :

L1.1 ; L2.4

L1.1 ; L2.2

L1.2 ; L2.8

L1.1 ; L2.7

L1.2 ; L2.1 …

Dans le cas N-M, nous aurons le résultat suivant :

L1.1 ; L2.4

L1.1 ; L2.2

L1.2 ; L2.8

L1.1 ; L2.7

L1.2 ; L2.4 …

Pour la tâche de « validation manuelle » des appariements proposés par Duke, il était impératif de trouver une présentation des résultats qui serait conviviale et facilitant le travail complexe de validation. Une structure en arborescence s'imposait. Pour cela, il a fallu intervenir sur le code source de Duke pour ajouter ce nouveau modèle de représentation des résultats de l'appariement qui devaient être validés.

À cet effet, il a fallu ajouter une notion de « profondeur » dans les résultats et une hiérarchie qui pourrait, lorsque cela se posait, différencier plus facilement les entreprises et les établissements.

Si nous reprenons l'exemple précédent, nous arriverons au résultat suivant :

Cas	1-1	1-N	N-M
Résultat	*L1.1* |_ *L2.4* *L1.2* |_ *L2.45*	*L1.1* |_ *L2.4* |_ *L2.2* |_ *L2.7* *L1.2* |_ *L2.8* |_ *L2.1*	*L1.1* |_ *L2.4* |_ *L1.2* |_ *L2.8* |_ *L2.2* |_ *L2.7*

Grace à cette représentation des appariements retenus dans la phase automatique, on a constaté une nette amélioration du temps de validation manuelle par les agents du PCBS, passant en moyenne de 10 établissements par heure par agent à 100 établissements par heure.