FAIRE FACE AUX DÉFIS DE LA DOCUMENTATION ET DE LA MISE À DISPOSITION DES ENQUÊTES INTERNATIONALES ET LONGITUDINALES : LE CAS DU GENERATIONS AND GENDER PROGRAMME

Arianna Caporali[1]

[1] *Institut national d'études démographiques (Ined), 133, boulevard Davout - 75020 Paris,* arianna.caporali@ined.fr

Résumé. Les métadonnées sont essentielles dans la mise à disposition des enquêtes internationales et longitudinales pour fournir aux utilisateurs des informations sur la comparabilité des données de différents pays et de différentes vagues d'enquête. Une documentation des données bien planifiée est par conséquent nécessaire pour le bon management de ces enquêtes. Cet article traite de la documentation des données dans le *Generations and Gender Programme (GGP)*. Il s'agit d'une enquête démographique longitudinale auprès des personnes âgées de 18 à 79 ans en Europe et au-delà. Les instruments d'enquête ont été adaptés aux différents contextes nationaux, ou bien incorporés dans des enquêtes existantes. Les défis de la documentation des enquêtes GGP consistent en la nécessité de combiner les informations sur les spécificités de chaque pays avec les informations concernant l'harmonisation *post hoc* des données, et ce à travers différentes vagues. Pour relever ces défis, la documentation des données suit une procédure structurée. Les métadonnées sont fournies conformément à *la Data Documentation Initiative* (DDI), une norme internationale de documentation des données, et elles sont mises à disposition avec les jeux de données en ligne via le logiciel Nesstar. L'article aborde les limites de la documentation des enquêtes GGP, ainsi que la manière dont ils seront résolus à l'avenir. Basé sur un plus grand suivi de la collecte par l'équipe de coordination centrale, le nouveau cycle de GGP à partir de 2020 permettra d'optimiser la documentation et la mise à disposition des données.

Mots-clés. Documentation, mise à disposition, enquêtes internationales, métadonnées, DDI

Abstract. Metadata are essential in the activities of providing access to international and longitudinal surveys to give users information on the comparability of data from different countries and different survey waves. A well-planned documentation is therefore necessary for a good management of these surveys. This paper presents the ways in which metadata are provided in the Generations and Gender Programme (GGP). This is a longitudinal demographic survey of 18-79 year olds in Europe and beyond. Survey instruments have been adapted to different national contexts, or incorporated into existing surveys. The challenges of GGP survey documentation are the need to combine information on the specificities of each country with information on *post hoc* data harmonization processes across different waves. To face these challenges, data documentation follows a structured procedure. Metadata are provided in compliance with the Data Documentation Initiative (DDI), an international standard for documenting social science data, and are made available online together with the datasets through the software Nesstar. The article discusses the limitations of GGP survey documentation, as well as how they will be addressed in the future. Based on greater monitoring of the data collection by the central coordination team, the new round of GGP from 2020 will optimize the documentation and the activities of providing access to data.

Keywords. Data documentation, providing access to data, international survey, metadata, DDI

1. Introduction : les métadonnées dans les enquêtes internationales

Les données d'enquête peuvent être utilisées uniquement en combinaison avec des métadonnées complètes. Ce sont des « données sur les données » indispensables à l'analyse quantitative des données et à leur interprétation. Sans les métadonnées, le partage des données et les analyses secondaires ne seraient pas possibles, et les chercheurs ne pourraient pas tester et répliquer les études des autres (King 1995). Dans les programmes d'enquêtes internationales, les métadonnées sont primordiales pour documenter la qualité des données de différents pays et évaluer les problèmes d'équivalence entre les méthodologies et les données de différents pays (Mohler et al. 2010). Les métadonnées sont encore plus importantes dans le cadre d'enquêtes longitudinales, où les vagues d'enquêtes subséquentes doivent également être décrites.

La préparation des métadonnées concerne les producteurs de données, ainsi que les centres d'archives de données d'enquêtes chargés du travail de mise à disposition des données (Caporali, Morisset et Legleye 2015). Les producteurs de données créent des fichiers cohérents et préparent la documentation de l'enquête. Les archives de données d'enquête doivent garantir leur réutilisation même au-delà des frontières nationales. À cette fin, elles créent des fichiers de métadonnées exhaustifs et structurés selon les normes internationales. Aujourd'hui, la norme appelée *Data Documentation Initiative* (DDI)[1] est largement utilisée en sciences sociales, à la fois pour des enquêtes nationales et internationales (Mohler et al. 2010). Elle permet de décrire les enquêtes en général (méthodologie, terrain, résumé, etc.), ainsi que chaque variable dans le détail (texte de la question, méthode de calcul des variables dérivées, etc.). DDI est souvent implémentée en combinaison avec des logiciels qui permettent de publier et d'explorer les données et les métadonnées en ligne, comme par exemple le logiciel Nesstar[2].

Qu'est-ce que le travail de documentation de la comparabilité des enquêtes internationales et longitudinales ? Quels sont les problèmes de documentation posés par ces enquêtes? Comment peuvent-ils être surmontés? Ces questions sont importantes parce qu'une bonne documentation des données joue un rôle central dans les enquêtes internationales (Mohler et al. 2010). Cet article présente la manière dont les métadonnées sont préparées et mises à disposition dans le *Generations and Gender Programme* (GGP). Le défi spécifique que pose la documentation de données dans ce projet est la nécessité de combiner des informations provenant d'enquêtes déjà collectées avec des informations sur les processus d'harmonisation *post hoc* (c'est-à-dire après la collecte)[3] des données

[1] DDI (http://www.ddialliance.org/) a été lancé au milieu des années 1990. Il se compose d'un ensemble d'éléments normalisés en XML (eXtensible Markup Language) pour la description générale des enquêtes jusqu'au niveau de chaque variable d'un ensemble de données. Deux types de DDI existent. Le premier est le *DDI-Codebook* (DDI-C ou DDI 2), qui est conçu pour documenter des données d'enquête simples et se concentre sur les éléments d'un dictionnaire de codes traditionnel. Ce type de DDI peut être utilisé en combinaison avec le logiciel Nesstar (Vardigan, Heus et Thomas 2008). Le deuxième type de DDI est le *DDI-Lifecycle* (DDI-L ou DDI 3) qui documente les enquêtes tout au long de leur cycle de vie (Kramer et al. 2011). Ce type de DDI est particulièrement adapté aux enquêtes internationales et longitudinales, car il permet des comparaisons entre des éléments d'ensembles de données différents (Mohler et al. 2010). Une troisième version de DDI (DDI 4) vise à répondre aux besoins des nouveaux utilisateurs et de nouvelles technologies.

[2] Nesstar (http://www.nesstar.com/) a été développé dans la seconde moitié des années 1990 (Marker 2013) et il est géré par le Centre norvégien de données de recherche (NSD). C'est un logiciel de gestion des métadonnées au format *DDI-Codebook* qui offre un outil analytique en ligne. Les internautes peuvent visualiser les distributions des variables à l'aide de tableaux et de graphiques, ainsi qu'effectuer des analyses statistiques de base (p. ex., des tableaux croisés, des régressions et des corrélations). Il est également possible d'appliquer des poids, d'effectuer des analyses sur des sous-ensembles de répondants, de télécharger les données et les métadonnées dans divers formats de fichiers (y compris DDI), ainsi que les résultats d'analyse.

[3] Voir Granda, Wolf et Hadorn (2010) sur les différentes procédures d'harmonisation de données d'enquête et sur leur documentation.

et ce pour chaque vague de l'enquête. L'article discute des limites de la procédure mise en place et de la manière dont ces limites seront palliées à l'avenir.

2. Le cas du Generations and Gender Programme (GGP)

2.1 Caractéristiques principales et méthodologie

GGP est un programme d'enquêtes par panel sur les personnes âgées de 18 à 79 ans en Europe et au-delà[4], lancée en 2000 par la Commission économique des Nations Unies pour l'Europe (UNECE), pour étudier les changements démographiques (Vikat et al. 2007). Il est géré par un consortium d'institutions de recherche. L'équipe de coordination centrale de GGP est basée à l'Institut interdisciplinaire de démographie des Pays-Bas (NIDI). De nombreuses autres institutions contribuent à son développement, dont l'Institut nationales d'études démographiques (Ined). GGP est la suite de certains programmes internationaux d'enquêtes sur la fécondité menés depuis les années 1960, tels que *Family and Fertility Surveys* (FFS). Par rapport à ses prédécesseurs, il présente des aspects innovants (Vikat et al. 2007). GGP est une enquête longitudinale (avec un intervalle de trois ans entre chaque vague) visant à étudier les défis sociodémographiques et économiques, tels que la faible fécondité, l'évolution de la famille et le vieillissement de la population. GGP aborde ces thèmes de façon multidisciplinaire par le biais de la démographie, la sociologie, l'économie et la psychologie[5]. Il fournit des données d'enquête avec les *Generations and Gender Surveys* (GGS) sur les relations entre les générations et le genre le long des parcours de vie des individus, et il complète les données de l'enquête au niveau micro avec une base de données contextuelles au niveau macroéconomique (Caporali et al. 2016 ; Vikat et al. 2007)[6].

GGP est basé sur un modèle de gestion relativement décentralisé. Les instruments d'enquête et les directives méthodologiques pour la collecte des données sont élaborés par l'équipe de coordination centrale (UNECE 2005, 2007). Ils comprennent des recommandations[7] et un questionnaire « standard » fourni en anglais au format .pdf. Ces recommandations peuvent être adaptées par les équipes nationales aux différents contextes nationaux. La collecte de données peut faire l'objet d'enquêtes « sur mesure » (p. ex., en Bulgarie, en France et en Allemagne), ou bien être partiellement incorporé dans des enquêtes existantes (p. ex., en Australie, en Hongrie et en Italie). Cela conduit à des différences dans les calendriers et les méthodologies de terrain entre les pays (Fokkema et al. 2016), ainsi qu'à des différences dans le niveau de « conformité » au questionnaire standard (Emery et Caporali, en préparation). Par exemple, cela se traduit par des écarts dans les années de la première vague (variant entre 2004 en Bulgarie et 2012 en Suède) et dans le nombre minimum de tentatives de contact des répondants. Le niveau de conformité au questionnaire varie entre 84% (Bulgarie) et 23% (Pays-Bas) des questions posées. Les variables et les catégories de réponses spécifiques aux pays sont souvent conservées dans les fichiers de données mis à

[4] En Mars 2018, 19 pays ont réalisé la première vague : Allemagne, Australie, Autriche, Belgique, Bulgarie, Estonie, France, Géorgie, Hongrie, Italie, Japon, Lituanie, Norvège, Pays-Bas, Pologne, République tchèque, Roumanie, Russie, et Suède. La deuxième vague est disponible pour 13 de ces pays. La troisième vague est en cours de mise à disposition.

[5] Parmi les sujets de recherche que GGP permet d'étudier il y a p. ex. : l'impact des événements précoces sur les trajectoires de vie, la dynamique des relations familiales, la relation entre les intentions de fécondité et la décision d'avoir un enfant, l'impact du rôle des aides-soignants sur leurs vies, les interactions entre les comportements individuels et les contextes socio-économiques et politiques des pays, la conciliation entre vie familiale et vie professionnelle.

[6] Pour plus d'information sur GGP et ses données voir : https://www.ggp-i.org/.

[7] Les recommandations méthodologiques incluent : réaliser trois vagues (une vague tous les trois ans), tirer un échantillon à la première vague d'environ 10 000 individus vivant dans des ménages ordinaires, avoir au moins 8 000 répondants à la troisième vague, mettre en place un échantillonnage aléatoire, avoir un nombre comparable d'hommes et de femmes, limiter les pertes entre vagues par sur-échantillonnage et suivi des personnes entre les vagues, et réaliser des enquêtes en face à face CAPI (*Computer-assisted personal interviewing*) ou PAPI (*Paper and pencil personal interviewing*).

disposition. Malgré la faible coordination du travail de terrain, GGP met en œuvre une procédure centralisée d'harmonisation et de documentation des données. Les métadonnées suivent la norme DDI et s'appuient sur les possibilités offertes par le web. Comme dans d'autres programmes d'enquêtes internationales (p. ex., *l'European Social Survey*[8] et *l'European Values Study*[9]), les données et les métadonnées sont diffusées en ligne par le biais de Nesstar.

2.2 Les défis de la documentation des données

La méthodologie de GGP pose un certain nombre de défis pour la documentation des données. Premièrement, il est nécessaire de fournir une description complète des spécificités de chaque pays. Cela implique des métadonnées détaillées sur la méthodologie de collecte des données, incluant la description du mode de collecte, les protocoles de contact, les procédures d'échantillonnage et de pondération des données. Il faut également documenter les niveaux de conformité des pays au questionnaire standard. Pour cela, il est nécessaire de décrire pour chaque variable les déviations éventuelles des pays par rapport au texte de la question et aux catégories de réponses standards[10], ainsi que d'informer sur sa disponibilité dans les fichiers des données de chaque pays et vague de l'enquête. Il est également important de fournir tous documents supplémentaires pouvant être utile pour la compréhension des spécificités des pays, comme par exemple les questionnaires nationaux et les rapports méthodologiques. Toutes les informations spécifiques au pays doivent être collectées auprès de l'équipe nationale chargée du travail sur le terrain.

Deuxièmement, les métadonnées spécifiques aux pays doivent être complétées par des métadonnées qui sont les mêmes pour toutes les enquêtes GGP. Cela concerne les métadonnées au niveau de chaque variable qui doivent inclure la description du questionnaire standard (textes des questions, catégories de réponse, etc.), ainsi que la méthode de calcul des variables dérivées. Ces métadonnées doivent également inclure une description du processus d'harmonisation *post hoc* des données. Celui-ci consiste notamment à vérifier les labels des variables et de leurs modalités, à examiner la cohérence des jeux de données à la structure du questionnaire international (ex. : contrôle des réponses valides et des valeurs manquantes), et finalement à calculer des variables dérivées.

Troisièmement, toutes ces métadonnées doivent être facilement accessibles et compréhensibles par les utilisateurs des données. Elles doivent être fournies pour chaque vague et de telle manière que les utilisateurs puissent comparer les spécificités des pays et les processus d'harmonisation au cours des vagues suivantes. Enfin, toutes les informations doivent être organisées dans Nesstar afin de permettre une exploration conviviale des données et des métadonnées en ligne, et conformément à la spécification *DDI-Codebook* supportée par Nesstar (voir[1] et [2]). Il est important de souligner que cette spécification a été conçue pour des enquêtes simples. La spécification *DDI-Lifecycle*, plus adaptée aux enquêtes longitudinales et comparatives, pourrait simplifier la documentation de la comparabilité entre pays et vagues de l'enquête (mais elle n'est pas supportée par Nesstar).

2.3 Comment les défis sont surmontés

2.3.1 Un travail de documentation nécessairement structuré et minutieux

Afin de faire face aux défis de la documentation des données GGP, le travail (effectué par l'Ined en

[8] Voir : http://nesstar.ess.nsd.uib.no/webview/.
[9] Voir : http://zacat.gesis.org/webview/index.jsp?object=http://zacat.gesis.org/obj/fCatalog/Catalog5.
[10] Ces déviations peuvent donner lieu à la création de variables spécifiques à un pays, quand la question posée diffère par rapport au questionnaire standard, tout en abordant le même sujet. Les noms de ces variables ont un code spécifique au pays comme suffixe (p. ex., pour la France le code est « 15 »). Les déviations peuvent donner lieu à des catégories de réponses spécifiques à un pays, quand la question posée est conforme au questionnaire standard, mais les catégories de réponse diffèrent, même partiellement. Ces catégories sont précédées par un code spécifique au pays. Des déviations peuvent également concerner l'univers des questions, c'est-à-dire les personnes à qui la question a été posée.

collaboration avec le NIDI, responsable de l'harmonisation des données) doit être structuré en deux procédures, l'une concernant la documentation des variables et l'autre la documentation des méthodologies de terrain. Chaque variable d'un nouveau fichier harmonisé (p. ex., un fichier de données inédit ou une nouvelle version d'un fichier déjà publié) est tabulée (dans SPSS et STATA, les deux formats dans lesquels les données GGP sont disponibles) pour vérifier s'il y a des labels et/ou des modalités non renseignés. Les codes des valeurs manquantes sont également révisés pour s'assurer qu'ils sont correctement attribués. Toutes erreurs éventuelles sont corrigées et, si nécessaire, des clarifications supplémentaires sont demandées à propos des spécificités des pays. Cela peut prendre beaucoup de temps en fonction du nombre de questions à discuter et de leur complexité. Une fois le jeu de données finalisé, il est importé dans Nesstar, à partir d'un fichier DDI pré-rempli avec les métadonnées qui sont les mêmes pour toutes les enquêtes GGP. Cela permet de renseigner automatiquement ces informations dans les items DDI où elles sont stockées. Chaque variable est ensuite révisée et, si nécessaire, ces métadonnées sont affinées « à la main » pour prendre en compte et expliquer les éventuelles spécificités des pays. C'est un travail détaillé, méticuleux et chronophage. Par exemple, pour la première vague, il doit être fait pour plus de 2000 variables (version 4.3). De plus, à ce stade, un fichier de données distinct est créé (dans SPSS) pour documenter la disponibilité des variables entre les pays et les vagues. Ce fichier de données est également importé et documenté dans Nesstar.

En ce qui concerne les métadonnées sur les terrains, ces informations sont collectées auprès des équipes nationales sur la base d'un modèle structuré en fonction des items DDI choisis pour documenter les enquêtes GGP[11]. Les informations fournies sont révisées et, si nécessaire, des clarifications sont demandées. Cela implique souvent beaucoup de travail à la fois de la part des équipes nationales et de l'Ined. Ces informations sont ensuite importées dans Nesstar, ainsi que les liens vers la documentation nationale pertinente. Lorsque tout le processus de documentation est terminé, les fichiers de données et les métadonnées sont publiés en ligne en libre accès dans le *GGP Online Codebook and Analysis* (https://www.ggp-i.org/data/browse-the-data/).

2.3.2 Une organisation conviviale des données et des métadonnées en ligne

Il est important d'assurer un accès rapide et une compréhension facile à toute cette grande quantité de données et métadonnées. Pour cela, trois fichiers de données différents sont mis à disposition dans le *GGP Online Codebook and Analysis*, basés sur les différents profils des utilisateurs de GGP. Premièrement, les *Polled Datasets* contiennent, pour chaque vague d'enquête GGP, toutes les variables de tous les pays disponibles. Ces fichiers s'adressent aux utilisateurs qui veulent avoir des aperçus des données et métadonnées pour plusieurs pays en même temps. Ils contiennent des métadonnées identiques pour toutes les enquêtes GGP. Par exemple, il y a des informations sur la manière dont les enquêtes GGP doivent être citées, un résumé et des mots-clés sur leur contenu, ainsi que l'explication du processus d'harmonisation des données. Les utilisateurs peuvent visualiser pour chaque variable les tris à plat pour plusieurs pays en même temps et accéder à des informations détaillées sur le questionnaire standard de GGP, ainsi que sur les spécificités de chaque pays. Ces métadonnées présentent par exemple : le texte de la question posée, les déviations des pays par rapport à la question et / ou les catégories de réponses du questionnaire standard, l'univers (c'est-à-dire le sous-ensemble de répondants à qui la question a été posée) et, dans le cas de variables dérivées, les explications de la méthode de calcul.

[11] Il y a deux items principaux : « Métadonnées » contient des informations sur l'enquête en général, et « Description des variables » contient la description détaillée de chaque variable. Le champ « Métadonnées » est organisé en trois sous-domaines : 1) « Description du document », avec des informations sur le fichier DDI-Nesstar, 2) « Description de l'étude », présentant des métadonnées sur les terrains des enquêtes (méthodologie de la collecte des données, méthode d'échantillonnage, taux de réponse, etc.), 3) « Description des fichiers de données » décrivant le contenu du fichier contenant les variables. On demande aux équipes nationales des informations concernant la « Description de l'étude ».

Deuxièmement, des fichiers de données et métadonnées appelés *Country Data Files* s'adressent aux utilisateurs intéressés par des pays spécifiques. Pour chaque pays, ces fichiers contiennent les données et les métadonnées des toutes les vagues disponibles. Ils permettent de tenir pleinement compte des déviations nationales dans les méthodologies du terrain, ainsi que de comparer les informations et les données concernant les différentes vagues. Ils incluent des métadonnées et des descriptions de variables qui sont les mêmes pour toutes les enquêtes GGP. En plus de cela, ils contiennent des informations spécifiques aux pays sur le terrain comme : les procédures d'échantillonnage, le mode de collecte des données, les caractéristiques des enquêteurs, le protocole de contact, les actions pour minimiser les pertes entre les vagues, les variables pour pondérer les données, le taux de réponse. Dans ces fichiers figurent également des informations sur les différences entre les versions ultérieures des fichiers de données des pays (ex. : les variables dérivées qui ont été ajoutées ou révisées). Des liens donnent accès à une documentation supplémentaire spécifique aux pays, comme par exemple le site web du projet national GGP, de documents méthodologiques et les questionnaires nationaux. Cependant, la disponibilité de ces informations spécifiques aux pays peut varier d'un pays à l'autre, et, pour certains pays, il n'a pas été possible de collecter toutes les métadonnées requises. Pour chaque variable, ces fichiers fournissent les mêmes informations que celles disponibles dans le *Polled datasets*. Toutefois, les distributions des variables incluent uniquement les cas spécifiques au pays. De plus, grâce aux fonctionnalités offertes par l'outil Nesstar, il est possible d'effectuer des tabulations et des analyses statistiques simples en fusionnant les jeux de données de différentes vagues.

Troisièmement, le fichier de données *Variable Availability* documente la disponibilité des variables entre les pays et les vagues. Ce fichier contient toutes les variables de tous les pays et vagues de l'enquête disponibles. Les noms de variables ont le préfixe « x », où la lettre « x » représente n'importe quelle vague d'enquête. Chaque variable a des observations égales aux codes GGP des jeux de données des pays où elle est présente[12]. Les informations sur la conformité des pays au questionnaire standard GGP fourni par ce fichier de données sont particulièrement intéressantes à des fins exploratoires. Par exemple, il est possible de comprendre rapidement dans quels pays il est possible d'utiliser des données GGP pour étudier un sujet donné. Il renseigne sur la couverture géographique et temporelle/longitudinale de chaque variable.

2.4 Limites de la documentation des données

Pour documenter correctement les enquêtes GGP, une grande quantité de métadonnées est nécessaire. En particulier, la préparation des métadonnées spécifiques au pays peut prendre beaucoup de temps et nécessiter beaucoup de ressources et d'efforts de la part de l'équipe de coordination centrale et des équipes nationales de GGP. C'est notamment le cas pour les enquêtes GGP qui sont incorporées dans des enquêtes nationales existantes. Dans ce cas, la documentation des déviations par rapport au questionnaire standard de GGP, ainsi que celle concernant les choix effectués pendant le processus d'harmonisation des données, est cruciale. Cette documentation peut nécessiter de renseigner un grand nombre d'informations très détaillées et précises sur un certain nombre de variables, telles que les explications des questions spécifiques au pays et les codes de réponse. Tout ce travail ralentit la mise à disposition des données. En conséquence celle-ci a souvent lieu quelques années après la fin du terrain (ex. : en Suède, le terrain de la première vague est terminé en 2012 et le fichier des données a été mis à disposition en 2015).

Puisque la documentation intervient après la collecte et le processus d'harmonisation des données, des informations peuvent s'avérer manquantes ou être peu claires, surtout en cas de changement dans les équipes nationales. Par exemple, il peut y avoir des lacunes concernant les raisons pour

[12] Par exemple, si une variable est disponible dans le jeu de données de la première vague pour la France, cette variable aura une observation égale à « *15.1 - France W1* », où « 15 » est le code GGP pour la France et « 1 » pour la vague 1.

lesquelles il a été décidé d'introduire des déviations par rapport au questionnaire standard et/ou aux recommandations méthodologiques. En outre, les questionnaires nationaux ne sont pas toujours disponibles en anglais, ce qui complexifie leur utilisation dans la documentation. De plus ils ne peuvent pas être importés directement dans Nesstar car ils sont fournis, le plus souvent, en .pdf, un format qui n'est pas compatible avec ce logiciel. Les variables et les catégories de réponse spécifiques des pays sont, par conséquent, documentées « à la main » ; ce qui ralentit la mise à disposition des données et peut être source d'erreurs. En revanche, les fonctionnalités analytiques de Nesstar sont intéressantes pour les utilisateurs tant à des fins exploratoires que pédagogiques et elles représentent un atout de la mise à disposition des données GGP.

3. Perspectives pour l'avenir : GGP 2020

Après le financement de l'UE dans le cadre du 7ème programme-cadre (2008-2012), une nouvelle ère pour GGP a commencé en 2016, quand le Forum européen stratégique pour les infrastructures de recherche (ESFRI) lui a accordé le statut de projet émergent. GGP prévoit d'élargir ses bases de données et prépare un nouveau cycle de collecte de données à partir de 2020. Ce nouveau cycle de collecte de données, appelé GGP 2020, se basera sur une méthodologie renouvelée (Gauthier et Emery 2016 ; Emery et Gauthier 2017 ; voir aussi le site internet de GGP).

Afin de réduire les coûts du terrain, la nouvelle collecte sera multi-mode, en face à face ou online. De plus, la nouvelle méthodologie visera à : accroitre la conformité des pays au questionnaire standard et la comparabilité des données, réduire les ressources pour l'harmonisation *post hoc* et la documentation des données, et permettre une publication plus rapide des jeux des données. Dans ce but, la nouvelle collecte sera basée sur des prescriptions (et non-pas des recommandations) méthodologiques ; ce qui permettra davantage d'harmonisation *ex ante* des données. La centralisation sera accrue et le travail de terrain sera suivi de près par l'équipe coordinatrice de GGP basé au NIDI. Cette nouvelle méthodologie sera possible grâce à un nouveau questionnaire[13] fourni aux équipes nationales et déjà traduit dans les langues des pays et en format CAPI[14].

La documentation des données sera mise en œuvre dès le début du terrain, avec une plus grande automatisation des procédures. Les jeux des données et les questionnaires nationaux seront importés dans Nesstar directement en format *DDI-Codebook*[15]. Les jeux des données commenceront à être révisés et documentés dans Nesstar pendant la collecte. Les spécificités des pays pourront ainsi être expliquées avec davantage de précision. Par conséquent, la mise à disposition des enquêtes sera améliorée et plus rapide. Cette méthode a été testée lors de l'enquête pilote en Biélorussie en 2017, avec des résultats prometteurs : l'enquête a été publiée dans Nesstar (en version *beta*) 4 jours après la fin du terrain. Elle sera expérimentée à nouveau en 2018 lors de l'enquête pilote qui testera le design multi-mode en Allemagne, Croatie, et Portugal. L'équipe GGP continuera également à mener une veille constante sur les opportunités offertes par les développements concernant les standards de documentation et les outils d'exploration en ligne des données. Une attention particulière sera notamment réservée aux développements concernant *DDI-Lifecycle* et DDI 4 (voir[1]) et les logiciels qui permettent de les implémenter.

4. Conclusions

Les métadonnées sont cruciales dans les enquêtes internationales et longitudinales, pour informer sur la comparabilité des données de différents pays et vagues d'enquête. Une documentation

[13] Le questionnaire sera révisé mais comparable avec celui du cycle de collecte précédent.
[14] Ceci sera possible notamment en s'appuyant sur les logiciels TMT (*Translation Management Tool*) et Blaise.
[15] Il est en effet possible de « corriger » les fichiers XML de collecte via une transformation XSL dans le but de réagencer certaines informations et de conserver le format DDI compatible.

comparative optimale doit être complète, suivre les standards de documentation des données et s'appuyer sur les possibilités offertes par le web (Mohler et al. 2010). Elle ne doit pas ralentir le processus de mise à disposition des données. Cet article décrit l'expérience concernant la documentation et la mise à disposition des enquêtes GGP. Celle-ci montre que, dans ce programme d'enquêtes internationales, basé jusqu'à présent sur un modèle de gestion relativement décentralisé, il est nécessaire de déployer beaucoup de ressources afin d'assurer une documentation complète des données. En revanche, davantage de centralisation comme celle prévue dans GGP 2020, avec notamment un suivi du terrain dans les différents pays par l'équipe coordinatrice, peut avoir un effet bénéfique sur la qualité de la documentation des données, ainsi que sur la rapidité de leur mise à disposition. À l'avenir, l'expérience d'autres programmes internationaux d'enquêtes (ex. : *Survey on Health, Ageing and Retirement in Europe*, *European Social Survey*) sera prise en compte et comparée avec celle de GGP.

Bibliographie

Caporali A., Morisset A., et Legleye S. (2015). La mise à disposition des enquêtes quantitatives en sciences sociales: l'exemple de l'Ined. *Population-F* 70(3): 567-597. DOI: 10.3917/popu.1503.0567.

Caporali, A., Klüsener, S., Neyer, G., Krapf, S., Grigorieva, O., et Kostova, D. (2016). The contextual database of the Generations and Gender Programme: Concept, content and research examples. *Demographic Research* 35(9): 229-252. DOI: 10.4054/DemRes.2016.35.9.

Kramer, S., Banks, R., Chang, V., Sieber, I., Vardigan, M., et Zenk-Möltgen, W. (2011). Presenting longitudinal studies to end users effectively using DDI metadata. *DDI Working Paper Series* – Longitudinal Best Practice, 4. DOI: 10.3886/DDILongitudinal04.

King, G. (1995). Replication, Replication. *PS: Political Science & Politics* 28(3): 444-452.

Emery, T. et Gauthier, A.H. (2017). *The Generations and Gender Programme: Innovating in an established Survey Research Infrastructure*. Poster présenté au XXVIII[e] Congrès international de la population, Le Cap (Afrique du Sud), 29 octobre – 4 novembre.

Emery, T. et Caporali, A. (en préparation). The added value of cross-national studies: Compliance and usage in the Generations and Gender Programme.

Fokkema, T., Kveder, A., Hiekel, N., Emery, T., et Liefbroer, A. C. (2016). Generations and Gender Programme Wave 1 data collection: An overview and assessment of sampling and fieldwork methods, weighting procedures, and cross-sectional representativeness. *Demographic Research* 34(18): 499-524. DOI: 10.4054/DemRes.2016.34.18.

Gauthier, A.H. et Emery, T. (2016), The Generations and Gender Programme: Past, present and future. *Demos: bulletin on population and society* 32 (7), special issue: 7. (http://www.nidi.nl/shared/content/demos/2016/demos-32-07-gauthier.pdf, dernier accès 28/03/2018).

Granda, P., Wolf, C. et Hadorn R. (2010). Harmonizing Survey Data. In: Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L.E., Mohler, P.P., Pennell, B.-E., et Smith, T.W. (dir.). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, N.J.: Wiley & Sons: 315-332. DOI: 10.1002/9780470609927.ch17.

Marker, H. J. (2013). Strengthening cooperation between European social science data archives: The evolving role of CESSDA. In: Kleiner, B., Renschler, I., Wernli, B., Farago, P., Joye, D. (dir.), *Understanding Research Infrastructures in the Social Sciences*. Zurich, Seismo Press: 39-46.

Mohler, P.P., Hansen, S.E., Pennell, B.E., Thomas, W., Wackerow, J., et Hubbard, F. (2010). A survey process quality perspective on documentation. In: Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L.E., Mohler, P.P., Pennell, B.-E., et Smith, T.W. (dir.). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, N.J.: Wiley & Sons: 299–314.

DOI: 10.1002/9780470609927.ch16.
UNECE – United Nations Economic Commission for Europe. (2005). *Generations and Gender Programme - Survey Instruments*. New York/ Geneva: United Nations. (http://www.unece.org/pau/pub/ggp_survey_instruments.html, dernier accès 28/03/2018).
UNECE - United Nations Economic Commission for Europe. (2007). *Generations and Gender Programme - Concepts and Guidelines*. New York/ Geneva: United Nations. (https://www.unece.org/pau/pub/ggp_concepts_guidelines.html, dernier accès 28/03/2018).
Vardigan, M., Heus, P., et Thomas, W. (2008). Data Documentation Initiative: Toward a standard for the social sciences. *The International Journal of Digital Curation* 3 (1): 107-113. DOI: 10.2218/ijdc.v3i1.45.
Vikat, A., Spéder, Z., Beets, G., Billari, F., Bühler, C., Desesquelles, A., Fokkema, T., Hoem, J.M., MacDonald, A., Neyer, G., Pailhé, A., Pinnelli, A., et Solaz, A. (2007). Generations and Gender Survey (GGS): Towards a better understanding of relationships and processes in the life course. *Demographic Research* 17(14): 389–440. DOI: 10.4054/DemRes.2007.17.14.