

SECTION 7

# Gestion des données





# Gestion des données

- **Planifier la gestion des données avant leur collecte.**
- **Ne pas sous-estimer l'ampleur de la tâche.**
- **Accorder autant d'attention à l'extraction des données de la base de données qu'à leur saisie.**
- **S'assurer que la base de données est correctement documentée.**

Les données rassemblées pendant une enquête doivent être compilées, contrôlées, saisies dans une base de données adéquate, préparées pour l'analyse immédiate et archivées pour une future utilisation. Cette phase du travail est généralement appelée «gestion des données». Chaque enquête individuelle aura des dispositions différentes, et ces directives ne prévoient pas d'instructions progressives sur la planification ou sur la mise en œuvre des activités de gestion des données. Cependant, il faudra consacrer de l'attention aux thématiques décrites dans les sous-sections ci-après. Une liste de contrôle des points à inclure au plan de gestion des données est présentée à l'étape 8 de la section 4.

## PLANIFIER LA GESTION DES DONNÉES AVANT LA COLLECTE

Il est absolument essentiel de planifier la gestion des données avant leur collecte. Un plan de gestion des données devrait représenter un des produits principaux de la phase de planification de l'enquête (voir section 4). Le temps et les ressources nécessaires pour concevoir des procédures appropriées de gestion des données, y compris la gestion des risques, ne devraient pas être sous-estimés.

Il est également essentiel que le plan pour la gestion des données ne soit pas élaboré séparément des plans des opérations de terrain et d'analyse des données (voir section 4). Ces trois éléments doivent être alignés entre eux et sur les objectifs de l'enquête. Aussi bien planifiées et mises en œuvre que soient les opérations de terrain, toute l'enquête sera endommagée si les procédures de gestion des données ne sont pas en mesure de gérer le matériel recueilli sur le terrain. De même, si les procédures de gestion des données n'assurent pas de mécanismes efficaces en ce qui concerne l'extraction des données pour l'analyse et l'établissement de rapports, l'enquête ne fournira pas les informations nécessaires à l'amélioration de la gestion des ressources zoogénétiques. La planification de la gestion des données se concentre souvent sur la première phase du travail – comment stocker les données de façon systématique – et beaucoup moins sur l'extraction efficace des données. C'est une erreur. Il est crucial de consacrer la même attention tant à l'extraction des données du système de gestion de l'enquête qu'à leur saisie. La réflexion sur les modes d'analyse et d'archivage des données indiquera une méthode pour rassembler et gérer les données.

La gestion des données prévoit une série d'activités – conception de la base de données; création de la structure de la base de données; gestion du matériel rassemblé sur le terrain,



ainsi que les questionnaires complétés; contrôle, codage, saisie, stockage et archivage des données – dont chacune présente ses propres défis. Un autre défi est représenté par la nécessité d'intégrer efficacement ces activités, surtout si l'enquête est à grande échelle et si les activités de gestion des données sont réalisées dans des sites dispersés. Les sous-sections ci-après fournissent une vue d'ensemble de ces questions et quelques conseils pratiques. Les descriptions détaillées des procédures et des normes pour la gestion des données dans les enquêtes à grande échelle auprès des ménages se trouvent dans deux documents de la FAO (1995 et 2005a), qui ont été publiés dans le cadre du programme Recensement mondial de l'agriculture 2000. Fink (2008) et Fowler (2008) donnent des aperçus succincts de la gestion et de l'analyse des données des enquêtes et des recensements à grande échelle.

### **DISPOSITIONS POUR LA PROPRIÉTÉ ET LA TUTELLE DES DONNÉES DE L'ENQUÊTE**

Il faut aborder les questions juridiques et éthiques concernant la tutelle et la propriété des données de l'enquête, ainsi que l'accès à ces données et leurs utilisations (voir sections 3 et 4). Si le pays dispose d'une stratégie nationale d'enquête et de suivi, le cadre général pour la tutelle et la propriété des données rassemblées devrait avoir déjà été mis en place. Les dispositions pour les enquêtes individuelles devraient s'aligner sur le cadre national. Il est également important de considérer les aspects pratiques des dispositions sur la tutelle des données. Profiter des services offerts par les bureaux nationaux de recherche et de statistique est une possibilité. Quelles que soient les dispositions, les utilisateurs pertinents devront être en mesure d'accéder facilement aux données tant à court terme qu'à long terme. La nature des données rassemblées justifiera peut-être des liens opérationnels avec des bases de données existantes, comme celles qui contiennent les données du recensement de l'agriculture, les données biogéographiques géoréférencées, les données sur les établissements humains et sur les services socio-économiques (par exemple, les routes, les télécommunications, les écoles et les services vétérinaires) ou les cartes détaillées des frontières administratives. L'utilisation de ce genre de données peut ajouter une valeur considérable à la précision et à l'étendue des rapports de l'enquête. Il est par conséquent impératif que les plans pour la gestion des données consacrent toute l'attention qui s'impose au besoin d'assurer la compatibilité, les liens et éventuellement l'interopérabilité<sup>16</sup> entre les différentes bases de données. Ce genre de lien devrait être pris en compte pour l'élaboration d'accords pour la tutelle et la propriété des données.

### **CHOISIR LE LOGICIEL À UTILISER POUR LA GESTION DES DONNÉES**

En raison des évolutions rapides des logiciels de saisie et de gestion des données et du matériel y associé, ces directives ne proposent pas de recommandations spécifiques sur leur choix. Il faudrait rechercher des conseils actualisés auprès des experts compétents dans le domaine des statistiques et des technologies de l'information, surtout en ce qui concerne les applications personnalisées. Le logiciel choisi pour la gestion des données doit prévoir

<sup>16</sup> L'interopérabilité est atteinte lorsque les programmes peuvent échanger les données par un ensemble commun de formats, lisent et écrivent les mêmes formats de fichier et utilisent les mêmes protocoles.



## ENCADRÉ 44

**Excel n'est pas un logiciel de gestion des bases de données – un exemple au Royaume-Uni**

Au Royaume-Uni, certaines procédures expérimentales des instituts de recherche sont juridiquement réglementées. Au Roslin Institute, les procédures de conformité ont impliqué la saisie quotidienne des ensembles de données de base, ou articles, dans un tableur Excel, où chaque donnée était entrée dans une ligne qui chevauchait plusieurs colonnes, et chaque colonne contenait des aspects différents de l'information requise. Un des employés a voulu trier le tableur, mais il a fait une petite erreur. Ainsi, à la place de classer toutes les colonnes dans le même ordre, une seule colonne a été triée. Par conséquent, le lien contextuel entre les données stockées dans la colonne à peine triée et toutes les autres données a été interrompu. Tout le tableur, et les données saisies au cours de plusieurs mois, sont devenus inutilisables. Heureusement, dans ce cas, grâce aux procédures de sauvegarde de l'institut, seules les données entrées dans cette journée spécifique ont dû être saisies de nouveau. Cela aurait pu être bien pire.

Cet exemple démontre que Microsoft Excel n'est pas un logiciel conçu spécialement pour la gestion des bases de données. Les logiciels conçus spécialement à cette fin, comme Microsoft Access ou Sequel Server, Oracle ou CPro (logiciel libre gratuit du Bureau du recensement des États-Unis d'Amérique, disponible à l'adresse Internet <http://www.census.gov/ipc/www/cspro/index.html>) reconnaissent intrinsèquement que les données sont organisées en articles et que chaque article contiendra peut-être plusieurs parties différentes d'information. Ils reconnaissent également qu'un ensemble d'articles devra peut-être s'associer à d'autres ensembles d'articles de façon flexible. Par exemple, pendant une enquête simple réalisée dans un certain nombre de villages, où plusieurs éleveurs sont interviewés chaque jour sur les différents animaux d'élevage, un ensemble d'articles contiendra peut-être des informations sur les villages (taille de la population, accès à l'eau, proximité du marché, etc.); un autre ensemble, des informations sur les interviewés (âge, taille de l'exploitation, taille du ménage, etc.); et un autre encore, des informations sur chacune des espèces d'animaux d'élevage de chaque exploitation (nombre, races, etc.). Évidemment, il est important de pouvoir associer les données sur les animaux d'élevages aux données sur l'exploitation et, de même, les données sur l'exploitation aux données sur le village.

Fourni par John Woolliams, Roslin Institute, Royaume-Uni.

une saisie efficace des données, la création d'une base de données, la recherche dans la base de données pour extraire des données récapitulatives et doit s'adapter aux besoins des utilisateurs pour l'analyse. Dernièrement, le logiciel Microsoft Excel a été largement utilisé pour saisir et archiver les données, bien que ce logiciel n'ait pas été conçu à cette fin (voir encadré 44).



Les données à gérer seront peut-être de plusieurs types différents.

- Données métriques: c'est-à-dire, les variables quantitatives classiques, comme le rendement de lait. Il faudrait remarquer que ces variables nécessitent peut-être d'apparaître plus d'une fois dans la base de données: le rendement de lait sera peut-être enregistré sur le terrain en seaux standardisés (voir encadré 39) et ensuite converti en kg ou litres. Il ne faudrait pas omettre les enregistrements de terrain dans la base de données.
- Données ordinales: c'est-à-dire les données catégoriques qui ont un ordre hiérarchique. Par exemple, la question suivante pourrait être posée aux éleveurs: «La race A survit-elle mieux que la race B à la sécheresse? 1) beaucoup mieux, 2) mieux, 3) pareil, 4) pire, 5) bien pire». Dans ce cas, les chiffres allant de 1 à 5 représentent un ordre hiérarchique de l'adaptation relative à la sécheresse.
- Données nominatives: c'est-à-dire les codes numériques des variables qualitatives. Par exemple, «Que devient le lait produit? 1) vendu au marché, 2) consommé dans le ménage, 3) les deux, 4) autres utilisations». Ici, les codes allant de 1 à 4 facilitent l'enregistrement, mais l'ordre ne sous-entend pas la valeur quantitative.

Il sera peut-être également nécessaire de gérer les codes en langage distinct, les textes libres, les fichiers vidéo et les fichiers audio. Les options de logiciel pour une saisie et un archivage efficaces de ces types de données doivent être analysées pendant la phase de planification pour que les données puissent être rassemblées dans le format et dans la mesure corrects. Il est probable qu'au cours de l'enquête se présentent des opportunités de rassembler des informations imprévues, très importantes et explicatives. Voici un exemple: pendant l'enquête, on pourrait à l'improviste trouver que les propriétaires pasteurs d'un troupeau de chèvres de l'échantillon, pour réduire le risque de mortalité des jeunes, évitent de sélectionner les hases qui donnent des naissances multiples (ceci se produit, par exemple, parmi les pasteurs Borana, dans le sud de l'Éthiopie). Les dispositions de gestion des données devraient prévoir la saisie et le stockage de conclusions inattendues de ce genre, et ensuite de pouvoir y accéder pour analyse.

La base de données créée pour l'enquête nécessitera peut-être des mises à jour, surtout si l'enquête de base est suivie par une série d'enquêtes de suivi. Par conséquent, des ajustements à la structure et à l'organisation de la base de données seront peut-être nécessaires, et il faudrait en tenir compte lors du choix du logiciel.

Les bases de données pour les enquêtes principales ont été parfois établies en utilisant des logiciels sur mesure. L'expérience démontre que cette approche a été souvent regrettée. Il est fort recommandé de créer les bases de données pour l'enquête en utilisant des logiciels établis (par exemple, Microsoft Access, Sequel Server, Oracle, MySQL). Les raisons de ce choix sont les suivantes:

- un logiciel sur mesure requiert du temps additionnel pour écrire les codes à utiliser pour les opérations standard d'une base de données;
- la documentation est beaucoup plus vaste, car on ne peut pas utiliser des manuels déjà établis; et
- il devient nécessaire d'engager du personnel à long terme qui comprenne le logiciel sur mesure, plutôt que d'employer du personnel déjà formé dans l'utilisation des logiciels établis.



## DOCUMENTER LES MÉTHODES DE TRAVAIL

La documentation est une partie essentielle de la gestion des données; les meilleurs plans sont inutiles si ceux qui sont censés les mettre en œuvre ne connaissent pas la question. Par conséquent, pour assurer l'application cohérente des activités de gestion des données, un manuel de procédures approuvées devrait être rédigé et distribué à tous ceux qui sont engagés dans la collecte, dans le codage, dans la vérification, dans la recherche et dans l'analyse des données.

Il est facile de surestimer le temps pendant lequel on va se souvenir des détails des sujets qu'on ne traite plus activement, même si ces détails sont d'une énorme importance. Par exemple, l'utilisation d'un jargon peut créer des problèmes. Un acronyme est peut-être d'usage courant dans l'équipe de prospection au moment des opérations de terrain et de la saisie des données, mais peut devenir complètement incompréhensible quelques années plus tard, lorsque les données sont extraites d'une archive. Pour éviter ce genre de problèmes, les champs de données et leurs conventions de codage doivent être documentés de façon claire et non ambiguë, et explicitement associés au questionnaire de terrain ou à d'autres sources de données. L'adéquation de la documentation devrait être évaluée en demandant à quelqu'un, qui n'est pas engagé dans l'enquête, d'analyser la base de données à l'aide de la documentation et de signaler si les données peuvent être identifiées correctement et facilement.

## STANDARDISER LES PROCÉDURES POUR LA GESTION DES DONNÉES BRUTES

Il ne faudrait pas négliger l'interface entre les activités de terrain (section 6) et les activités de gestion des données. Les protocoles pour une journée sur le terrain devraient assurer que tous les enquêteurs savent quand, où et à qui transmettre les questionnaires et tout autre matériel (photographies, enregistrements audio, diagrammes préparés pendant les exercices de groupe, etc.) pour les premières phases de la gestion des données. La responsabilité de vérifier l'absence de pertes ou les erreurs d'étiquetage devrait être clairement attribuée. Il faudrait accorder de l'attention aux aspects pratiques de la gestion, du déplacement et du stockage en sécurité du matériel.

Comme il a été énoncé à la section 6, c'est une excellente idée que les équipes de terrain – enquêteurs et superviseurs – se réunissent après une journée de travail sur le terrain pour discuter de toute problématique rencontrée. Cette réunion représente une opportunité pour vérifier la présence de problèmes évidents dans les questionnaires complétés. Par exemple, on pourrait demander aux enquêteurs de préciser les écritures illisibles. Si, à ce stade, on découvre des données qui semblent des observations aberrantes et exceptionnelles, il sera peut-être utile d'enregistrer les commentaires sur les circonstances dans lesquelles ces données ont été rassemblées. Il faudrait envisager les procédures pour enregistrer ces commentaires. Dans certaines situations, il sera peut-être possible de revenir à la source et d'analyser davantage ces observations apparemment aberrantes. Toutefois, les enquêteurs, et même les superviseurs, n'ont pas le droit de modifier les données brutes uniquement parce qu'ils remarquent un résultat qu'ils considèrent comme imprévu. Il est essentiel de reconnaître que les enquêtes concernent aussi bien la quantification des variations naturelles que l'estimation des valeurs moyennes.



## ÉTABLIR DES PROCÉDURES POUR LA SAISIE DES DONNÉES

La «saisie des données», soit le transfert des données des questionnaires ou d'autres formats papier à la base de données électronique, peut représenter une entreprise majeure et il ne faudrait pas sous-estimer le temps, la patience et l'attention nécessaires pour la mettre en œuvre. Si l'enquête est à petite échelle et les activités sur le terrain ne sont pas effectuées dans une zone reculée, les arrangements physiques pour la saisie des données pourraient être assez simples. Les questionnaires complétés peuvent être apportés régulièrement à un seul endroit, pas loin du site de l'enquête, et entrés dans les ordinateurs par un petit groupe d'agents formés pour la saisie des données. Pour une enquête de plus grande taille, des arrangements plus élaborés, comme une des deux méthodes ci-après, seront nécessaires.

Méthode 1: Prévoir plusieurs stations de saisie des données et employer du personnel qualifié pour entrer et pour vérifier les données en utilisant un système adapté de saisie, avec une base de données centrale qui est continuellement mise à jour après vérification et approbation des données. Le nombre de stations nécessaires peut être estimé une fois que, grâce à l'expérience pratique, le temps moyen nécessaire pour saisir les données d'un questionnaire a été évalué (par exemple, FAO, 1995). Un exemple de système de saisie des données personnalisé et facile à utiliser, élaboré à cette fin, est BREEDSURV (Rowlands *et al.*, 2003). Il a été développé dans Microsoft Access 2000 par l'Institut international de recherches sur l'élevage (ILRI) et est utilisé pour les enquêtes sur les races d'animaux d'élevage dans les pays de la Communauté de développement en Afrique australe et en Éthiopie (Ayalew et Rowlands, 2004). À présent, d'autres logiciels sont disponibles pour la gestion des bases de données, spécialisés et plus adaptés, qui peuvent satisfaire ces besoins. Comme il a été énoncé plus haut, les possibilités d'utilisation de ces bases de données sur mesure après la conclusion de l'enquête sont limitées en raison de l'effort supplémentaire nécessaire pour la conception, pour la documentation et pour la formation du personnel.

Méthode 2: Prévoir un système de saisie des données en ligne par lequel les enquêteurs peuvent saisir directement les données dans la base de données en utilisant des ordinateurs portatifs. La vérification et l'édition automatisées des données peuvent être effectuées peu de temps après la saisie des données. Autrement, des lecteurs optiques peuvent être utilisés pour la saisie automatique des données, directement à partir de formats conçus exprès pour la collecte des données (FAO, 1995). Aujourd'hui (2010), ces options avancées seront peut-être trop coûteuses pour les pays en développement mais, compte tenu des progrès rapides dans les technologies, elles deviendront probablement plus largement applicables à l'avenir. Toutefois, les vérifications et la saisie des données manuelles peuvent être avantageuses car elles permettent de déceler des erreurs d'énumération et des réponses incohérentes. Ainsi, les sources des données peuvent être immédiatement visitées/interviewées de nouveau pour vérifier ou pour corriger les informations. L'encadré 45 décrit l'utilisation des technologies de l'information pour l'enregistrement et pour la surveillance des bovins en Irlande.

La saisie manuelle des données provoque inévitablement des erreurs aléatoires et systématiques. L'utilisation des systèmes automatisés de lecture optique et d'acquisition des données peut contourner de nombreuses erreurs humaines pendant la saisie. Comme il a





été indiqué plus haut, toutefois, il est peu probable que ces systèmes seront largement utilisables en ce moment. Il faudrait par conséquent se concentrer sur la réduction des erreurs humaines lors de la saisie manuelle des données. Une procédure avérée pour l'identification et pour la gestion de ces erreurs est la duplication de la saisie manuelle et la contre-vérification des données. Les logiciels spécialisés de gestion des bases de données, comme ceux qui ont été énumérés plus haut, donnent la possibilité de mettre en place ces actions. Ces bases de données prévoient peut-être également des procédures internes intégrées pour la vérification des données en ce qui concerne la cohérence et l'uniformité. Cependant, le logiciel ne peut pas remplacer la supervision minutieuse de l'Homme dans la saisie des données pour identifier et corriger les erreurs.

### ÉTABLIR DES PROCÉDURES POUR LA VÉRIFICATION DES DONNÉES

Outre les possibilités, mentionnées plus haut, d'identification et de correction des erreurs évidentes peu de temps après la collecte, il sera peut-être nécessaire, à un stade ultérieur, de vérifier et de réviser à fond les données de l'enquête pour atteindre la cohérence de l'ensemble des données (par exemple, en s'assurant que les mensurations d'événements impossibles, comme le rendement de lait des taureaux, ne sont pas enregistrées) et pour déceler, vérifier, corriger et éliminer les observations aberrantes qui autrement provoqueraient peut-être des écarts extrêmes dans les récapitulatifs. Les statisticiens experts devraient apporter ces modifications des données, après avoir considéré avec attention la nature et l'ampleur de toute valeur aberrante, car les observations aberrantes peuvent être valables ou pas.

Dans les enquêtes à grande échelle, la révision des données par ordinateur sera peut-être justifiée pour améliorer leur crédibilité, par exemple, en identifiant les données manquantes, en vérifiant les données par rapport aux gammes de valeur prévues, ou en vérifiant la cohérence logique et/ou numérique entre les données par rapport aux variables afférentes. De même, le codage des données sur des variables communes peut être automatisé et géré par ordinateur une fois que la gamme de valeurs et leur fréquence sont connues (FAO, 1995). Il sera peut-être impossible de corriger certaines erreurs sans visiter de nouveau les sources des données sur le terrain.

S'il est impossible de répéter les questionnaires sur le terrain, deux procédures statistiques standard peuvent être suivies au stade de l'analyse des données pour remédier aux erreurs sans produire des effets majeurs sur le résultat final: 1) ces entrées sont déclarées comme données manquantes, ou 2) sur la base des connaissances disponibles des données de l'enquête, les valeurs moyennes prévues peuvent être estimées et saisies – une procédure appelée imputation (pour de plus amples détails, voir FAO, 1995). L'imputation est décrite plus en détail à la section 8. Il est recommandé, à la phase de la saisie des données, de saisir toute donnée manquante, qui ne peut pas être acquise en revisitant la source, comme «manquante» et de conserver l'imputation pour la phase d'analyse des données. Cette approche est claire, non ambiguë et protège l'intégrité de la base de données en tant qu'enregistrement fidèle de l'opération sur le terrain.

Les taux d'erreurs de saisie des données devraient être estimés au début du processus de saisie pour que des mesures puissent être prises pour réduire les erreurs à des niveaux



## ENCADRÉ 45

**L'évolution des technologies d'enregistrement et de suivi – expériences en Irlande**

La République d'Irlande se concentre sur la fourniture d'excellents systèmes pour tous les aspects de la santé animale et de la traçabilité. En Irlande, les infrastructures d'enregistrement sont tellement développées que la base de données sur l'identification et sur le mouvement des animaux, gérée par le gouvernement et par ses organismes, contient l'information sur tous les bovins d'une exploitation, et tous les détails sur les naissances, les décès et les mouvements de ces animaux. La base de données est reliée à la base de données sur la santé animale (plate-forme Java J2EE, architecture à trois étages avec une base de données Oracle à la base). Les bureaux et les laboratoires pour la santé animale du gouvernement (1 500 personnes) sont également connectés en temps réel à ces bases de données par un Intranet officiel, tandis qu'un certain nombre d'employés du gouvernement et 1 000 vétérinaires privés, les abattoirs, les marchés de bovins, les centres de réunion, les points d'exportation et même quelques agriculteurs disposent de niveaux différents d'accès par Internet, selon leurs besoins et fonctions. Des modules pour les différents programmes de contrôle des maladies et de vaccination ont été élaborés. Le système facilite des réponses rapides à certaines situations et a transformé les pratiques de travail de certaines de personnes, éliminant ainsi beaucoup de formalités administratives.

Par exemple, la surveillance de la tuberculose bovine prévoit une visite par an à chaque exploitation. Pendant cette visite, tous les bovins de l'exploitation sont testés. Trois jours

après, une autre visite est prévue pour lire et pour enregistrer les résultats du test. Par le passé, cette surveillance se faisait uniquement sur papier. À présent, lorsqu'une ferme est à surveiller pour la tuberculose bovine, les identifications par marques auriculaires de tous les bovins qui devraient être présents à la ferme sont téléchargées sur un enregistreur/logger électronique de données dans les locaux du vétérinaire local par une interface XML<sup>1</sup> avec le système informatique de santé animale. Le vétérinaire apporte le logger de données à la ferme. Il enregistre par voie électronique la présence de l'animal et le lancement du test pendant la première visite et les résultats du test pendant la deuxième visite. Les enregistrements effectués à la ferme sont ensuite chargés sur la base de données officielle de surveillance de la santé animale et les informations sont traitées pour déterminer toute discordance entre les animaux prévus et les animaux présents, et ensuite pour programmer tout autre action qui sera peut-être nécessaire après les résultats du test. Les données sur la tuberculose bovine, cumulées dans la base de données de surveillance chaque année, comprennent 14 millions de contrôles individuels de santé animale pour la tuberculose et la brucellose, les conclusions de la surveillance des abattages et les résultats des laboratoires. Ces données sont également utilisées pour l'élaboration de procédures améliorées de contrôle et de différents projets de recherche.

Fourni par Margaret Good, Département de l'agriculture et de l'alimentation, Irlande.

<sup>1</sup> XML (langage de balisage extensible) est un ensemble de règles pour le codage des documents sous forme lisible par une machine.



acceptables: par exemple, en identifiant le personnel et le matériel qui sont associés aux erreurs et en évaluant si les problèmes peuvent être corrigés par une autre formation du personnel ou par la réparation des claviers. Les taux d'erreurs donnent une indication générale de la qualité et de la précision des données. La création de tableaux récapitulatifs des données à la fin de chaque séance de saisie constitue un autre moyen pour vérifier la qualité des données et pour identifier et gérer les observations aberrantes et les données manquantes.

## ÉTABLIR DES PROCÉDURES POUR LE CODAGE DES DONNÉES

Un aspect de la gestion des données qui peut représenter un défi de taille est le processus de codage pour la saisie dans la base de données. Son importance dépendra de la nature des données traitées, qui dépendra de la nature des questionnaires utilisés sur le terrain, qui à leur tour seront influencés par les objectifs de l'enquête.

Dans certaines situations, le codage est une question simple, qui se limite à convertir les noms complets d'articles ou de concepts clairement définis – qui peuvent être interprétés de façon cohérente le long de la chaîne de communication qui commence par l'éleveur, à travers les enquêteurs et les agents de saisie des données jusqu'aux analystes – en «codes» qui peuvent être manipulés plus facilement et plus efficacement par le logiciel de gestion et d'analyse des données. Par exemple, les noms des espèces d'animaux d'élevage peuvent être codifiés: bovins = 1; buffles = 2; chèvres = 3; moutons = 4; et ainsi de suite. De plus, pour certains types de questions de l'enquête, la gamme des réponses possibles sera prévisible à l'avance et peut être codifiée avant les activités sur le terrain; par exemple, les noms des espèces d'animaux d'élevage de la zone à l'étude ou les noms des villages à visiter. Pour ces articles, le codage ne devrait poser que quelques problèmes de théorie ou de logistique.

Par contre, pour certaines questions, la gamme des réponses ne sera peut-être pas prévisible à l'avance. Les différences locales de la langue relèveront peut-être de cette problématique. Par exemple, les maladies des animaux d'élevage seront probablement connues sous plusieurs noms locaux différents et, peut-être, que personne dans l'équipe de terrain ne sera en mesure de faire correspondre les noms utilisés par les éleveurs aux noms scientifiques des maladies. Également, les difficultés peuvent surgir parce que l'équipe de prospection ne disposait pas de connaissances suffisantes pour prévoir toutes les différentes manières de vivre et de travailler des éleveurs; par exemple les traitements sanitaires traditionnels et spécifiques pour les animaux. Le pilotage minutieux du questionnaire et des «évaluations rapides» préliminaires de consultation avec les informateurs clés ou les groupes d'éleveurs devrait contribuer à minimiser le nombre de réponses difficiles à codifier (voir sections 2, 5 et 6). Toutefois, le plan de gestion des données devrait établir des procédures qui mettent l'accent sur les éventuelles actions des agents de saisie des données en cas de problèmes et comment ces problèmes devraient être affrontés par les responsables de l'enquête.

Plus les questions de l'enquête sont «ouvertes», plus le problème du codage peut devenir délicat. Si l'enquête cherche à obtenir des informations, par exemple, sur les problèmes, les préférences ou les motivations des éleveurs ou à sonder les raisons pour lesquelles les animaux d'élevage ou certaines races particulières sont utilisés à des fins spécifiques, les



réponses décriront peut-être les processus, les relations ou les conditions, et ne seront pas une liste d'articles distincts et facilement codifiables. De plus, l'importance de certaines réponses ne sera peut-être complètement évidente qu'une fois que les réponses à d'autres questions, éventuellement d'autres interviewés, ont été prises en considération. Dans ces cas, il sera peut-être nécessaire de réaliser un codage et un reclassement substantiels des données brutes après la collecte.

En règle générale, les décisions sur le codage et sur le reclassement des données brutes ne devraient pas être prises par les enquêteurs, ou même par les superviseurs, car cette partie de la gestion des données devrait être fusionnée et gérée au niveau central. L'intégration de cette prise de décisions centralisée aux dispositions de gestion des données est probablement un défi, surtout pour une enquête à grande échelle où la saisie des données est gérée par un certain nombre de personnes dans des emplacements différents, et où la saisie des données et les activités sur le terrain se chevauchent dans le temps. Dans certains cas, les décisions finales sur le codage et sur le classement des données sont mieux gérées dans le cadre de leur analyse. Les conseils détaillés sur la gestion et sur l'analyse des données qualitatives vont au-delà de la portée de ces directives. Cependant, il est important de reconnaître que si les méthodes de collecte de données qualitatives «ouvertes», comme les interviews semi-structurées, doivent être utilisées pour une enquête à grande échelle, la gestion des données (et surtout le codage) nécessiteront l'attention des experts et une planification minutieuse. Si les complexités potentielles de la gestion des données qualitatives ne sont pas expliquées, les réponses détaillées des éleveurs pourraient rester juste sur papier et n'auraient aucune influence sur l'analyse et sur les résultats de l'enquête – auquel cas, peu serait acquis de l'utilisation des outils qualitatifs.

## ÉTABLIR DES PROCÉDURES POUR LE STOCKAGE ET POUR LA SÉCURITÉ DES DONNÉES

Il faudrait appliquer les mesures de stockage et de sécurité des données aux questionnaires complétés ainsi qu'aux données électroniques. Même si les bases de données de l'enquête sont considérées des ressources du domaine public, l'accès non autorisé aux données devrait être évité.

La base de données devrait être dupliquée en sécurité pour créer des copies de sauvegarde. Ces copies devraient être conservées au centre des données (le site primaire du stockage des données) ainsi que dans un site secondaire («station périphérique»). Les copies sauvegardées dans un emplacement et dans un système différents permettront de récupérer entièrement la base de données si les copies de travail et les copies de sauvegarde locales sont inutilisables en raison d'erreurs humaines, de problèmes informatiques ou de catastrophes d'autres types.

Il faudrait toujours archiver deux copies de la base de données primaire (pour sécurité), avant l'analyse, car ces copies demeureront l'enregistrement fidèle des opérations de terrain. C'est une bonne pratique d'utiliser pour l'analyse les copies de la base de données primaire plutôt que de manier les données originales.

Les formats originaux complétés de la collecte de données (par exemple, les questionnaires) devraient être toujours archivés en sécurité pour pouvoir ensuite y accéder pour



référence. La nature et l'étendue de l'utilisation continue des données décideront du temps de conservation des données, mais il est généralement conseillé de conserver le matériel au moins jusqu'à la conclusion des deux enquêtes successives du même genre. Par exemple, dans les pays où les enquêtes sur l'élevage sont entreprises tous les dix ans, les données brutes devraient être conservées pendant environ vingt ans. Il est probable que les copies électroniques des données brutes auront une valeur utile beaucoup plus longue. Par exemple, elles pourraient être utilisées pour l'analyse chronologique des implications pour la gestion des ressources zoogénétiques lorsqu'elle est affectée par le changement climatique.



