Département de recherche
Sciences Sociales, Agriculture et
Alimentation, Espace et Environnement

Démarche qualité appliquée à la gestion et à la préparation des données

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

La démarche qualité appliquée à la gestion et à la préparation des données de SAE2

Version 1.0 – 21 décembre 2012[1]

Au sein du département SAE2[2], les données sont des composantes stratégiques de nombreuses recherches. Une partie de ces données relèvent de la « Mission bases de données » et sont ainsi mutualisées pour les équipes de recherche. En complément à cette mission, des plateformes telles que l'ODR[3] ou le PAP[4] assurent l'assemblage et la gestion d'une part importante des sources dans le domaine de l'alimentation et du développement rural. Enfin, d'autres sources sont organisées et gérées de manière autonome dans les différentes unités qui les accueillent ou les produisent.

Pour entrer dans le processus de recherche proprement dit, ces données font l'objet de différents traitements décrits au travers des étapes du cycle de vie des données[5]. La fiabilité et la traçabilité de ces étapes, autrefois souhaitables, sont désormais nécessaires non seulement par souci d'efficacité, mais aussi parce que la profession l'exige. Certaines revues scientifiques demandent en effet que les programmes et données ayant généré les résultats soient mis à disposition de la communauté[6]. Par ailleurs, les données traitées dans le département sont mises à disposition de différents partenaires (INAO, MSA, ASP, Ministères, Instituts,…) qui, au même titre que la recherche, sont en droit d'exiger une qualité et une reproductibilité des processus qui ont servi à l'élaboration des données. En outre, compte tenu de l'investissement que représentent ces données, il apparaît indispensable d'en assurer la pérennité et de mettre en place des procédures visant à conserver ce patrimoine.

Aussi le département SAE2 a-t-il souhaité qu'une démarche « qualité des données » soit instruite. Cette instruction a été confiée à un groupe de travail[7] soutenu par le CATI IATISS[8] et supervisé par le CGD. Le travail mené concerne uniquement les données traitées dans le département, à savoir des informations recueillies lors d'enquêtes publiques, achetées auprès de fournisseurs, ou au cours de dispositifs *ad'hoc* mis en place dans le cadre de travaux de recherche. La démarche préconisée concerne donc des bases de données de taille raisonnable et ne semble pas pouvoir s'appliquer au traitement d'informations « à haut débit ».

---

[1] La dernière version de ce document est disponible sur https://esrcarto.supagro.inra.fr puis Réseau Recherche→ INRA SAE2→ Démarche Qualité.

[2] Sciences Sociales, Agriculture & Alimentation, Espace & Environnement.

[3] Observatoire du Développement Rural, plateforme web administrée et développée par une unité de service de l'INRA : US-ODR.

[4] Pôle Alimentaire Parisien, géré par l'UR ALISS qui assure également le suivi de l'Observatoire de la Qualité de l'Alimentation (OQALI).

[5] Ces traitements liés à la gestion et à la préparation des données sont parfois appelés "prétraitements" ou également regroupés dans l'appellation anglophone « data-management ».

[6] Par exemple, dans ses indications aux auteurs, l'AJAE demande : "Authors are expected to document their data sources, models, and estimation procedures as thoroughly as possible, and to make the data used available to others for replication purposes. If an exception to this rule is desired, reasons should be given and this should be explicitly noted in the cover letter".

[7] Ce groupe de travail était composé d'ingénieurs relevant de plusieurs unités : R. Chartier, E. Cahuzac et C. Gendre de l'ODR ; C. Bontemps et V. Orozco du GREMAQ ; C. Boizot-Szantai et de N. Guinet du PAP ; de J.L. Dupuis, C. Lanu et A. Lacroix de GAEL. Le document ainsi rédigé a bénéficié des critiques de S. Lecocq, O. Allais et S. Gojard (chercheurs à ALISS), T. Laurent, C. Bisière et C. Bonnet (ingénieur et chercheurs au GREMAQ), V. Piguet (ingénieur au CESAER). Toute remarque peut être adressée à dept-sae2-qd@supagro.inra.fr.

[8] Le CATI CITISES, successeur du CATI IATISS, a également soutenu et suivi ce travail.

Les préconisations s'adressent à toute personne qui traite des jeux de données quotidiennement, soit dans le cadre de la gestion de systèmes d'information, soit dans le cadre de travaux scientifiques. Elles visent à promouvoir l'homogénéisation des pratiques et garantir une qualité des processus de traitement en proposant trois outils :

- une charte au travers de laquelle ses signataires (plateformes de mise à disposition des données, unités proprement dites ou personnes *intuitu personae*) s'engagent à respecter un certain nombre de principes
- un guide des bonnes pratiques qui, pour chacune des étapes du cycle de vie des données, détaille différents outils et leur mise en œuvre pratique
- un pense-bête qui permet de s'auto-évaluer en repérant les étapes qui mériteraient d'être améliorées.

Dans chacune de ces parties, les mots soulignés sont définis dans le syllabus placé à la fin de ce document.

Charte assurant la qualité de la gestion et de la préparation des données

La démarche qualité de gestion et préparation des données, introduite par la présente Charte, vise à assurer la **fiabilité, la traçabilité** et la **pérennité** des données mises à disposition et utilisées au sein du département SAE2. A noter que cette Charte ne garantit pas la qualité des données proprement dites, que cette qualité soit *interne* (les données sont conformes aux spécifications de l'auteur) ou *externe* (les données répondent bien aux attentes de l'utilisateur). Par contre, elle garantit que tous les moyens ont été mis en œuvre pour gérer et traiter les données de manière fiable, pour conserver ces données elles-mêmes ou la trace des traitements qu'elles ont subis.

Via la présente Charte, **cinq principes** sont adoptés :

Recours aux métadonnées

La mise à disposition d'informations relatives aux données s'avère indispensable pour permettre une bonne compréhension de leur contenu et de leur degré d'élaboration (à quelle étape du cycle de vie elles se situent). Ce principe conduit à la standardisation de ces informations par l'utilisation de standards de métadonnées, d'ontologie et de thésaurus.

La personne ou le collectif signataire de cette Charte s'engage à appliquer un cadre informationnel normalisé (comme les standards Dublin Core, Data Documentation Initiative ou encore Statistical Data and Metadata eXchange) facilitant la diffusion et la compréhension des données à d'autres utilisateurs et a fortiori à d'autres systèmes d'information.

Traçabilité

La fiabilité des données réside dans la possibilité de suivre les différents états de celles-ci tout au long de leur cycle de vie. Pour cela, la présente charte préconise l'utilisation de techniques dites « *reproductibles* ». L'objectif est de tracer et de conserver les opérations effectuées sur les données lors des étapes de vérification, de correction, d'amélioration et d'enrichissement, avant leur mise à disposition aux utilisateurs.

La personne ou le collectif signataire de cette Charte s'engage à conserver les fichiers contenant les scripts et les procédures appliqués aux jeux de données et à ne pas recourir à des opérations manuelles ne laissant pas de traces (type copier/coller).

Stockage, sauvegarde et conservation

La démarche qualité des données promeut, en accord avec les règles de diffusion et de conservation données par l'auteur, le stockage, voire la conservation des jeux de données et de tous les éléments associés (métadonnées, thésaurus, fichiers log ou scripts), afin d'en permettre une utilisation ultérieure facilitée.

La personne ou le collectif signataire de cette Charte s'engage à définir et respecter des règles de nommage et d'archivage, à mettre en place des procédures de sauvegarde, de stockage et de conservation. Pour le stockage et la conservation, il s'engage à dédier des lieux ou espaces spécifiques accessibles à d'autres utilisateurs.

Respect des règles de diffusion et des droits d'auteur

Un objectif essentiel de la démarche qualité est de faire respecter les droits d'auteur ainsi que les règles d'utilisation des données établies par les auteurs, propriétaires ou par une instance décisionnelle (CNIL).

La personne ou le collectif signataire de cette Charte s'engage d'une part, à appliquer les règles de confidentialité et de conservation des données conformément aux conventions passées avec les propriétaires des données ou instances décisionnelles ; d'autre part, à citer ou faire citer par les utilisateurs, les sources exactes des données utilisées.

Signalement et correction des erreurs

L'amélioration de la qualité des données est un processus continu et itératif. A l'issue de certains traitements, l'utilisateur des données peut avoir repéré certaines erreurs qu'il convient de corriger au bénéfice de l'ensemble des utilisateurs (concept de *feedback*).

La personne ou le collectif signataire de cette Charte s'engage à mettre en place des outils pour faciliter les éventuels signalements d'erreurs par les utilisateurs, à effectuer les corrections nécessaires des données si ces erreurs sont validées, et enfin à signaler ces modifications aux utilisateurs.

Guide des bonnes pratiques pour la gestion et la préparation des données

Pourquoi ce guide ?

Après qu'un collègue ou un lecteur de votre article vous ait posé une question sur un résultat que vous avez obtenu, vous passez un temps considérable pour trouver les fichiers des programmes que vous avez utilisés pour le produire. Sans parler du temps pour comprendre ce que vous aviez vraiment fait dans ces programmes pour produire ce fameux résultat.

Parce que ce type de situations est familier, nous assistons depuis plusieurs années à un changement important dans les pratiques de la communauté scientifique. Cette dernière tend à promouvoir la « reproductibilité de la recherche » sous l'impulsion du mouvement « *Reproducible Research*[9] ». Suivi par des éditeurs de grandes revues (*Review of Economics and Statistics*, *Journal of Applied Econometrics*, etc..), ce processus se propage en amont, accentuant le contrôle sur toutes les étapes du processus de recherche menant à publication[10]: vérification des sources, vérification ou contrôle des résultats tirés de ces données, stockage dans des archives publiques ou privées, développement de plateformes.

Ce phénomène a conduit différentes institutions et universités (e.g. *National Science Foundation, American Statistical Association*, *Australian National University,* …) à proposer des règles ou des principes (*data management policies, ethics policies, data management plans, etc...*) s'appliquant aux chercheurs et ingénieurs traitant des données.

Au-delà de ces incitations institutionnelles, des études tendent à prouver que les articles permettant un accès aux données sont plus cités que les autres[11]. Enfin, avec le développement de l'informatique scientifique et des possibilités de collaborations à distance entre chercheurs et ingénieurs au sein d'un même projet, des méthodes et des outils se sont développés. Cette nouvelle façon de travailler est en train de s'imposer à tous les praticiens, et permettra à terme d'améliorer l'ensemble du processus de recherche, d'en améliorer l'efficacité et d'en augmenter la portée.

A qui s'adresse ce guide ?

Ce guide s'adresse à toute personne qui traite des jeux de données quotidiennement, soit dans le cadre de la gestion de systèmes d'information, soit dans le cadre de travaux scientifiques. Les

[9] Voir par exemple la première rencontre autour de la recherche reproductible organisée à l'université d'Orléans en avril 2012 (http://www.fdpoisson.fr/cascimodot/doc/RRRR/R4-050412.php) ou le site Reproductible Research (http://reproducibleresearch.net/).

[10] Ainsi de nombreuses revues (comme *l'European Review of Agricultural Economics*, *Economics Letters* ou *Sociologie du travail* par exemple ont adhérées au *Comittee On Publication Ethics* (COPE) et suivent les recommandations de cet organisme concernant les problématiques liées aux données (cf http://publicationethics.org/category/keywords/data-manipulation-/-falsification).

[11] Voir la page de Gary King à ce sujet http://gking.harvard.edu/pages/data-sharing-and-replication.

préconisations se centrent sur le traitement initial des données, dans cette zone généralement peu visible (le nuage de la figure 1), en amont du processus d'analyse proprement dit.

Les ingénieurs débutants et les doctorants y trouveront une aide pour structurer et organiser leurs travaux à partir de données brutes et y prendront de bonnes habitudes, puisque cette étape se reproduit dans chaque projet. Les ingénieurs et chercheurs confirmés pourront s'en servir pour analyser leurs pratiques en regard d'un objectif de traçabilité et de reproductibilité de leurs travaux et y découvrir de nouveaux outils simplifiant leur approche de ces traitements.

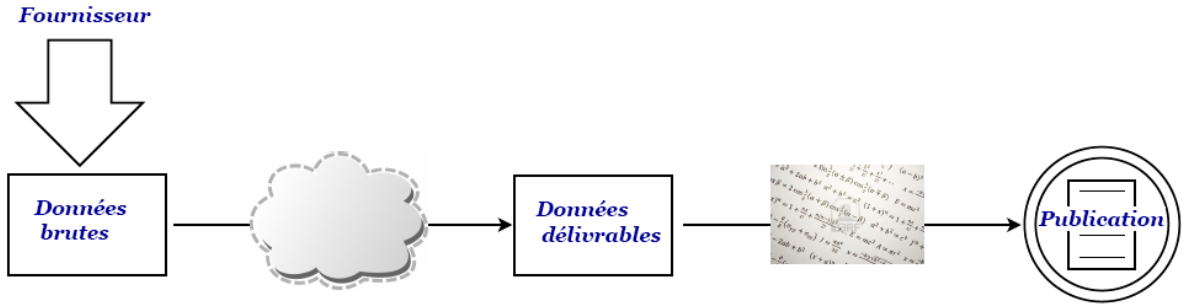


Figure 1 : Processus de recherche

En quoi consiste la démarche ?

La démarche qualité de gestion et de préparation des données s'inscrit tout au long du cycle de vie des données (voir figure 2), c'est à dire de la réception des données brutes[12] à leur utilisation dans les projets de recherche ou de leur mise à disposition aux utilisateurs (données délivrables), en passant éventuellement par des étapes de vérification, normalisation et enrichissement.

Les données évoluent au cours du temps, et passent de main en main dans un cycle dont une étape est représentée schématiquement par la figure 2. Il est à noter que ce processus se répète parfois, les fournisseurs pouvant se succéder les uns aux autres.

[12] Même si les données sont parfois collectées, nous nous plaçons ici à l'étape de réception des données, c'est à dire que nous considérons les données déjà disponibles.

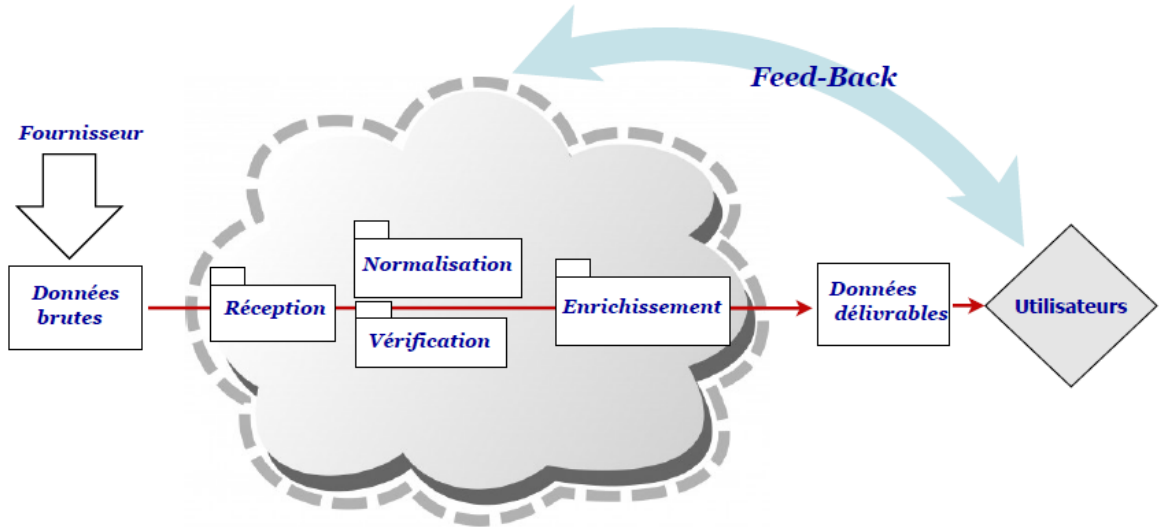


Figure 2 : Cycle de vie des données

Une chaîne de traitements, conforme à un cahier des charges, défini en amont, sera mise en place et créera les données délivrables à partir des données brutes. Tous ces traitements devront pouvoir être reproduits à l'identique ou en subissant une modification dans le but de générer de nouvelles données délivrables (par ex. : mise à jour des données en cas d'erreur détectée).

Aussi la chaîne de traitements devra être sauvegardée et archivée afin de maitriser le contenu des programmes et leur enchainement. Aucune intervention ne devra être effectuée hors de cette chaîne (ni manuellement). On devra "garder la trace" de toute intervention; aussi toute intervention, ainsi que tous les choix effectués, devront être documentés. En outre, il conviendra de travailler dans un souci d'amélioration continue, notamment en intégrant les remarques des utilisateurs. Enfin, la transmission du patrimoine devra être assurée : les données stockées seront documentées, et donc compréhensibles et utilisables par d'autres, si les règles d'utilisation et de diffusion des données le permettent.

En résumé, la démarche qualité se base sur une succession de règles de bon sens permettant d'assurer différents objectifs : traçabilité, reproductibilité, enrichissement et stockage.

Comment est organisé ce guide ?

Le guide détaille les actions qui devront être menées et les moyens pour les mener dans le but de satisfaire les cinq principes de la charte :

- le recours aux **métadonnées**
- la **traçabilité**
- le **stockage** et la conservation
- le respect des règles et **droit** d'auteur
- le **feedback** *i.e.* signalement et correction des erreurs

Le guide est organisé sous forme de fiches qui peuvent être consultées indépendamment les unes des autres et renvoient à l'un ou plusieurs des principes de la charte (surlignés en orange en haut de chaque fiche).

Sommaire des fiches disponibles

Fiches disponibles	Principes de la charte					A quel moment du cycle de vie ?
	Métadonnées	Traçabilité	Stockage	Droit	Feedback	
Fiche n°1. Définir et suivre un cahier des charges		**X**	X	X		*Réception/Vérification/ Normalisation/Enrichissement*
Fiche n°2. Se fixer des règles de stockage et d'arborescence		**X**	**X**			*Réception*
Fiche n°3. Utiliser des métadonnées	**X**	X		X		*Tout au long*
Fiche n°4. Suivre l'évolution des programmes et des fichiers		**X**	**X**			*Tout au long*
Fiche n°5. Se donner des règles de nommage	**X**	**X**				*Tout au long*
Fiche n°6. Maitriser l'enchaînement des étapes		**X**				*Tout au long*
Fiche n°7. Organiser le "feedback"		**X**			**X**	*Distribution*
Fiche n°8. Assurer la sécurité et la sauvegarde			**X**	X		*Réception*
Fiche n°9. Assurer la confidentialité				**X**		*Réception*

C'est quoi?

La définition d'un cahier des charges est indispensable, quelle que soit l'importance des données traitées. Même minime[13], ce cahier des charges doit clairement décrire le format des données attendues et les grandes lignes des traitements à effectuer (modifications de variables, création de nouvelles variables, vérifications à effectuer, type et format de fichiers, ...).

Afin de le définir, une réflexion avec les futurs utilisateurs des données doit avoir lieu pour connaître les besoins et bien fixer ensemble les étapes à réaliser.

Une réflexion sur la gestion des données (base de données...) doit aussi être menée individuellement ou collectivement.

Pour quoi faire?

Le cahier des charges permet d'assurer la **traçabilité** des traitements à effectuer en les listant clairement en amont. Il sera un document de référence formalisant les attentes des futurs utilisateurs ou clients concernant les données. Il est d'ailleurs à rédiger par la personne responsable des données, en accord ou avec les futurs utilisateurs ou clients.

A quel moment du cycle de vie ?

Le cahier des charges est à définir en amont de tout traitement, c'est à dire lors de la réception des données brutes (ou même avant). La conformité à ce cahier concerne principalement les étapes intermédiaires du cycle de vie : vérification, normalisation et enrichissement.

Exemples de bonnes pratiques

Organiser des réunions de travail avec les utilisateurs potentiels des données (chercheurs, clients) afin de connaitre les attentes concernant les données délivrables et bien formaliser avec eux ce cahier des charges.

Lorsqu'il n'y a pas de client, s'obliger à écrire quand même les principales étapes à suivre.

A minima, un « petit » cahier des charges doit lister par **écrit** les attentes (convention, membres rédacteurs, format des fichiers, modifications de variables, création de nouvelles variables, vérifications à effectuer...).

Exemples d'outils

Les outils collaboratifs (du type SilverPeas, Agora, Sharepoint …) facilitent l'élaboration écrite et collective du cahier des charges, ainsi que son suivi.

SilverPeas : Plateforme TCAO (Travail Collaboratif Assisté par Ordinateur).
Site officiel : http://www.silverpeas.com/
Site INRA : https://collaboratif.inra.fr/silverpeas

Agora : Plateforme TCAO non opensource.
Site officiel : http://www.online-agora.com/

Sharepoint : Plateforme TCAO de Microsoft non opensource.
Site officiel : http://sharepoint.microsoft.com/fr-be/Pages/default.aspx

[13] Un cahier des charges *minime* décrit ce qui est attendu à la fin d'un ou des traitements. Il précise, même sommairement, les variables à créer et le format du fichier (.xls, .dta, etc.).

C'est quoi?

Les règles de stockage et d'arborescence correspondent à la manière dont la personne gérant les données va les organiser sur son espace de travail (PC, serveurs...). Elles doivent être clairement définies en amont pour permettre **traçabilité**, **stockage** et **conservation**.

Pour quoi faire?

Des règles simples de **stockage** et de **partage** de fichiers permettront d'assurer l'unicité matérielle et temporelle (**intégrité**) des données, mais aussi des programmes et procédures.

A quel moment du cycle de vie ?

A définir à l'étape de réception.

Exemples de bonnes pratiques

Une pratique consiste à bien séparer les différentes étapes du cycle de vie des données en créant systématiquement au moins 4 répertoires : "DonneesBrutes", "DonneesTravaillees", "Programmes", "Outputs"[14]. Les données seront ainsi stockées selon une arborescence précise, séparant les données brutes, les programmes, et les données délivrables (finaux ou intermédiaires, fichiers de suivi, fichiers spécifiques). Des conventions pourront être établies afin de pouvoir disposer de la même arborescence sur plusieurs ordinateurs lors d'un travail en binôme[15]. Chaque programme utilise ensuite des chemins relatifs (par exemple « ../DonneesBrutes ») plutôt qu'absolus (« c:/MonProjet/DonneesBrutes ») dans l'appel des fichiers, programmes ou procédures, permettant à plusieurs co-auteurs de disposer d'un programme générique pouvant fonctionner sur n'importe quel poste.

Il est important de noter que les données brutes ne devront en aucun cas être modifiées (toute modification devra entrainer un changement de nom des fichiers).

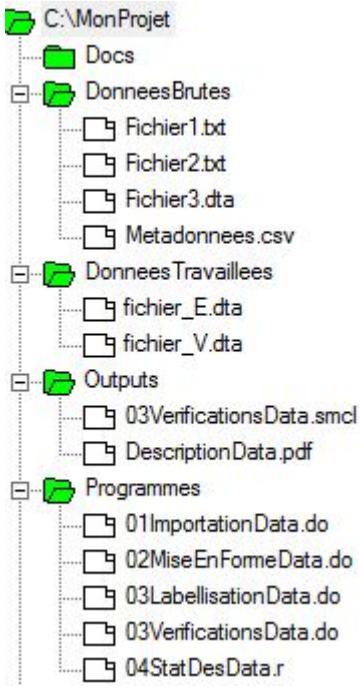


Figure 3 : Exemple d'arborescence

Un autre exemple est l'intégration des données dans un Système d'Information (cf. exemple du Système d'Information ODR dans le syllabus).

[14] A noter que si l'on se situe dans un cadre de travail international, le nom des répertoires devra être libellé en anglais.

[15] Il faudra toutefois veiller à ne pas enfreindre les règles d'intégrité. Il est en effet déconseillé de disposer de plusieurs copies d'un même fichier stockées à des endroits différents.

Exemples d'outils

Un petit programme (type ".bat" ou shell ou powershell) peut générer de manière automatique l'arborescence standard décrite plus haut).

Script permettant de créer une arborescence.

Le programme « CreateArborescence.bat » (code ci-dessous) permet de créer automatiquement l'arborescence de la Figure 3 :

```
REM "Programme de création des répertoires de base conformément à la démarche qualité"
REM " Ce programme devra être exécuté au niveau ou se situera la racine du projet (l'ensemble des répertoires)"
echo off
cls
REM "On demande le nom du projet qui sera le nom de la racine du projet"
set /p rep= Nom du nouveau projet :
REM "Creation du répertoire racine"
mkdir  "%rep%"
cd  "%rep%"

REM "Création des répertoires de nécessaires à tout traitement suivant (les noms peuvent être adaptés à chacun ci-dessous)"
mkdir  "DonneesBrutes"
mkdir  "DonneesTravaillees"
mkdir  "Programmes"
mkdir  "Docs"
mkdir  "Outputs"
REM " Un fichier portant la date de création est généré"
set Auto=Creation-%Date:~-4%-%Date:~-7,-5%-%Date:~-10,-8%.txt
echo Création de l'arborescence de ce projet %rep% le %Date:~-4%-%Date:~-7,-5%-%Date:~-10,-8%> %Auto%
cd ..
```

Ce programme peut être complexifié et adapté à chaque contexte de travail. Un autre exemple, plus complexe de génération automatique d'arborescence est disponible : http://www.grenoble.inra.fr/prog.zip

En cliquant sur "Cliquez_ici.bat", le script permet de créer une arborescence complète de projet ou une sous arborescence. Une première question permet de définir le type de cette arborescence :

- 1:projet : arborescence complète qui peut contenir au choix des " work package" de type "data" ou/et "expé".
- 2:expé : arborescence de la partie gestion des expériences, c'est-à-dire adaptée à la collecte des données (protocoles, gestion des sujets, …).
- 3:data : arborescence de la partie traitement des données, adaptée au traitement et à l'exploitation de données issues de sources externes ou d'expériences.

Il faut ensuite fournir le nom et l'année du projet, et pour un projet le nombre de "work package". Pour chacun d'eux, il faudra donner le type (1:expé, 2:data) et le nom. Le script construit alors l'arborescence en fonction de ces informations. Pour un projet, un répertoire "ADMIN" est présent quel que soit le nombre de "work package". De plus, deux répertoires supplémentaires ("WPxx_expe-Work_package_type" et "WPxx_data- Work_package_type") sont fournis comme modèles pour la création éventuelle d'autres "work package").

Remarques:

- les préfixes de chaque répertoire ont été choisis pour donner un classement homogène (classement par ordre alphabétique) de l'arborescence.
- des fichiers types (modèles) sont placés dans certain répertoire.
- le répertoire "init_prog" et le fichier "Cliquez_ici.bat" doivent être placés à la racine de l'emplacement choisi pour la création de l'arborescence.

C'est quoi?

Il s'agit de la description des données elles-mêmes (documentation des fichiers, dictionnaire des variables, leur format, modalités, propriété intellectuelle ...), qu'il s'agisse des données brutes ou des données délivrables. Même minime, cette description permettra de comprendre les données contenues dans un fichier, d'en connaître l'origine et d'en identifier la source.

Une métadonnée (mot à mot "donnée à propos de donnée") est une donnée servant à définir ou décrire le contenu d'une autre donnée[16] quel que soit son support. Elle peut être utilisée pour décrire une donnée sous la forme de catalogue, ou expliciter plus précisément le contenu d'un fichier ou d'une table. Elle permet en outre, de partager et de publier des informations dans un format commun.

Pour quoi faire?

Les métadonnées (et leur **standardisation**) sont nécessaires afin de disposer de l'ensemble des informations relatives aux données pour une parfaite compréhension et connaissance de leur contenu.

A quel moment du cycle de vie ?

Tout au long du cycle de vie.

[16]Une analogie simple avec la photographie permet de mieux comprendre. Lors de l'acquisition de données via un appareil photo numérique, *i.e.* lorsqu'on prend une photo, un fichier secondaire est enregistré. Ce petit fichier ne contient aucune information photographique mais des métadonnées relatives à la photo (date, objectif, distance focale, parfois localisation). Il s'agit d'un fichier de métadonnées, au format normalisé (format Exif).

Exemples de bonnes pratiques

- Une métadonnée simple consiste en l'écriture (papier, informatique) d'informations de base sur la provenance des données (nom de la source, format de fichier, identifiant, description du contenu, date de réception…) (cf figure 4).

- Une procédure courante consiste également à créer un fichier "lisez.moi" dans lequel ces informations seront écrites.

- Dans le cas où le fichier *lisez.moi* existe déjà, il n'est pas difficile d'ajouter des lignes à ce fichier pour structurer les éléments importants et les métadonnées liés à vos fichiers au cours de leur évolution dans la chaîne de traitements[17].

- Pour les bases les plus structurées, Dublincore est une norme internationale ouverte de métadonnées, développée par le Dublin Core Metadata Initiative. Elle comprend dans sa version de base 15 éléments de description des données (cf figure 5). Dans le domaine des SHS, on peut utiliser le standard DDI (Data Documentation Initiative) comme le réseau Quetelet et pour les données statistiques le standard SDMX (Statistical Data and Metadata eXchange) comme Eurostat.

- Dans le cas de bases de données, le schéma entité/association est à fournir puisqu'il indique les fichiers (tables) disponibles et leurs liens (cf figure 6).

Fichier	Contenu	Nb. Obs.	Clé
composition.dta	table nutriments (enerkc...)	N=1 342	id : code aliment (codal)
individu.dta	table individus (sexe_ps, v2_age...) pas poids, taille!!!	N=4 079	id : num individu (nomen)
individu_inca2_poids.dta (*)	idem avec poids, taille		
individu_nouvel_envoi.dta (*)	idem avec poids, taille + questionnaire fumeur, nb, nb maladies **+ infos sport**		
consoindiv.dta (*)	questionnaire	N=4 079	id : num individu (nomen)
compl_carnet.dta (*)		N=4 305	
nomenclature.dta	table aliments	N=1 343	id : code aliment (codal) et libellé (libal)
consorep.dta	table des consommations et détails (qté, marque...)	N=541 526	id : numéro ligne (numlig) par individu (nomen), jour (nojour), repas(tyrep)
consorep_nouvel_envoi.dta (*)	idem + var "avecqui" en +		
* 2ème envoi			

Figure 4 : Exemple de documentation des fichiers

[17]Evidemment, le nom du fichier devra être modifié.

Dublincore est une norme internationale ouverte, développée par le Dublin Core Metadata Initiative. Elle comprend dans sa version de base 15 éléments de description divisés en 3 groupes :

Contenu		Propriété intellectuelle		Instanciation	
Titre	SU.VI.MAX	**Créateur**	UREN	**Date**	2009-10-13
Sujet	Ménages, Produits …	**Collaborateur**	Ø	**Format**	SAS7, TXT
Source	https://intranet.inra …	**Editeur**	PAP	**Identifiant**	V1
Relation	Ø	**Droits**	Utilisable par les …	**Langue**	Français
Type	DataSet				
Description	Une table regroupant …				
Couverture	France métropolitaine				

Figure 5 : Description du Dublin Core (exemple à partir du jeu de données SU.VI.MAX)

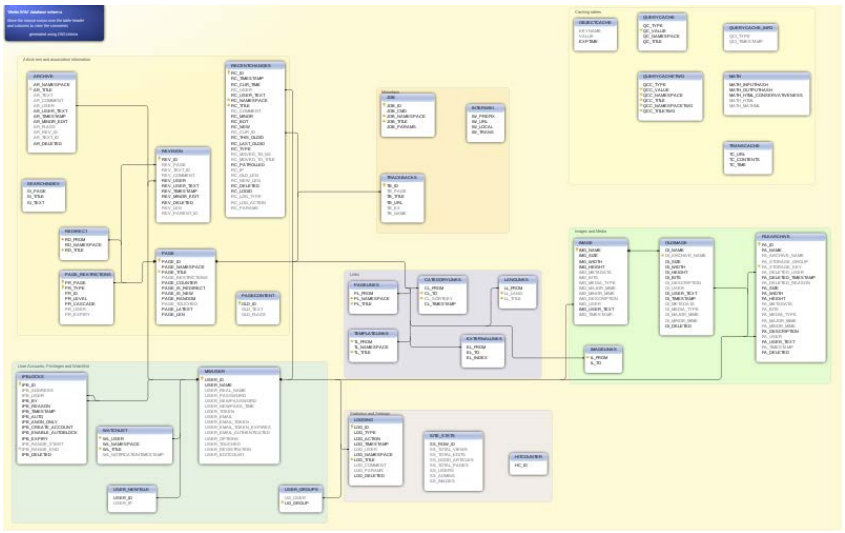


Figure 6 : Exemple de schéma entité/association

Exemples d'outils

- Norme internationale pour les métadonnées

Dublin Core.
Site : http://dublincore.org/

- Certains logiciels (Stata, Excel) permettent d'inscrire ces informations directement dans le fichier de données.

Stata : Logiciel de statistiques payant**.**
Site officiel : http://www.stata.com/

C'est quoi?

Il s'agit de mettre en place des règles simples de gestion de versions et de datation des programmes et des fichiers.

Pour quoi faire?

Les programmes et les fichiers évoluent, sont modifiés par une ou plusieurs personnes, et il est indispensable d'être capable de connaitre leur contenu et les différences entre les diverses versions existantes (**traçabilité**).

A quel moment du cycle de vie ?

Tout au long du cycle de vie.

Exemples de bonnes pratiques

- Numéroter chaque programme avec un numéro de version incrémentable à chaque modification. L'en-tête de chacun des programmes devra comporter un ensemble d'informations : date, auteur, traitements effectués, données brutes utilisées, données en sortie et historique des modifications importantes réalisées (ainsi que chaque numéro de version correspondant) (cf figure 7). Chaque fichier final doit aussi disposer de ce numéro de version (cf figure 8).

- Le système d'exploitation, la version du ou des logiciel(s) utilisé(s) devraient aussi être mentionnés afin d'assurer la reproductibilité à l'identique des traitements effectués (l'évolution technique implique parfois des changements).

- Transmettre les logs des programmes avec les fichiers de données (en plus des métadonnées).

- Créer et diffuser un dictionnaire des variables avec séparation entre variables créées et variables originales (voir Fiche n°5 sur les règles de nommage).

```
/*==================================================================================================*/
/* ============================== PROGRAMME DATAMAKERNFG ============================================*/
en sortie : fichiers d'achats du produit `prod' : p`prod'NF"

/*==================================================================================================*/
/* Ce programme NE TRAITE PAS LES FICHIERS MENAGES (voir MenagesMaker567.do)  */
/* DataMakerNF56 créé à partir de DataMakerNF345.do                           */
/* Version 1.1   29/01/08   sa3 = range.appellation et sa2 = range (pour cohérence
                            avec les fichiers ancienne formule) via un rename*/
/* Version 1.2   04/02/08   Initialisation des locales  (NbAnPasObs  AnSansObs) pour
                            test sur les années manquantes (r(N)>0 ) */
/* Version 1.3   13/02/2008 Modification du fichier ProduitNF56.dta si produit sans obs.  */
/* Version 1.4    7/03/2008 initialisation locale varlistannee_sans_prem   */
/* Version 1.41  26/03/2008 drop de an  */
/* Version 1.42  22/04/2008 FusionMarques56 intégrée et MN2NF corrigée (moda de Leader Price) */
/* Version 2.0   17/12/2008 Nouvel envoi 2005 et 2006                                          */
/* Version 2.1   15/06/2009 2005 2006 et 2007        */
/* Version 2.2   8/06/2010  Fichier des caractéristiques produits (Product_Desc_1aN.txt) modifié
                            pour 2005  */
/*                          (avant c'était le fichier 1er envoi 2005,
                            now celui 2ème envoi 2003-2006)  */
/* Version 2.3   25/08/2010 Pour nouvelles données 2006n, 2007n et 2008  */
/*                          correction labellisation fichier*/
/*Version 3.0  27/07/11 ajout condition pr liste des produits : besoin que l'info panel
                        soit renseignée*/
                        /*ex: ds Produits678.dta , pdt 538 n'a pas d'info panel donc
                        prog plante  (V) */
/* Version 3.1 30/08/11 : Changement de pu en Pu et qu en QU (conforme à notre règle
                            typographique)    */
/*Version 4.0  1/09/11 : version générique; chgt de boucle : 1 produit par année*/
/*                     : liste des produits à partir du fichier ProduitsNFXXX.dta
                        DANS la boucle année*/
/*==================================================================================================*/
local version "4.0"
```

Figure 7 : Exemple d'en-tête d'un programme Stata

```
. use J:\Secodip\Data2009\Produits\0005\p0005NF_E.dta
(PRODUITS SUCRANTS (0005) EPURE, annees (2007 2008 2009) ( 4 Jan 2012). PRIX EN
> E)

. note

_dta:
  1.  Qu= qorig*gawa*pweigh ; ptwa=ptwa*gawa*pweight (donc Pu=ptwa_orig/quorig)
  2.  Programme de labelisation des ménages : version M2.0
  3.  Créé avec la version 2.1 de MenageMakerG.do
  4.  Version 2.3 de MN2NF.ado
  5.  Créé avec la version 4.0 de DataMakerG.do
  6.  Créé avec la version 2.4 de VerifG.do
  7.  Version 1.0 de CreateUniteViasp10.ado
  8.  Attention, ce produit comporte plusieurs unités (cf sp10 ou tuwa)
  9.  Créé avec la version 1.1 de EpureG.do

achaber:
  1.  Rmq : si un achat a plusieurs erreurs d'achat, seule la dernière repérée
      sera indiquée dans la variable achaber
```

Figure 8 : Exemple de fichier Stata incorporant des informations de traitement

Exemples d'outils

- Des outils de gestion de versions comme Subversion ou GIT peuvent être utilisés afin de pouvoir comparer des versions de programmes (en mode texte) et ainsi revenir à des versions antérieures.

Subversion (cf figure 9)
http://subversion.apache.org/

GIT
http://git-scm.com/

- Total Commander permet également de comparer deux fichiers en mode texte et d'en illustrer les différences.

Total Commander (cf figure 10)
http://www.ghisler.com/accueil.htm

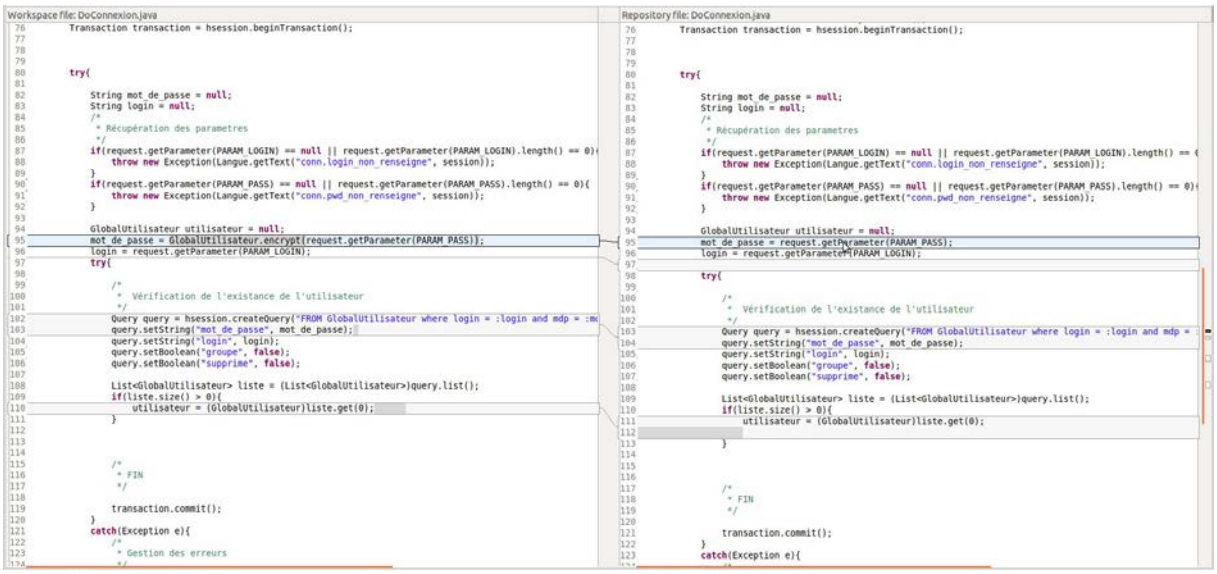


Figure 9 : Comparaison en mode texte de versions de programmes avec Subversion

Figure 10 : Comparaison en mode texte de versions de programmes avec Total Commander

C'est quoi?

L'archivage des données tout au long du cycle de vie demande la mise en place de certaines conventions s'appliquant aux données, aux scripts, procédures, logs. Ces règles pourront être clairement explicitées afin d'en partager la connaissance, par exemple entre deux co-auteurs qui sauront identifier les fichiers dont ils ont besoin parmi l'ensemble des fichiers à leur disposition.

Des règles simples et compréhensibles doivent être appliquées afin de retrouver les fichiers et d'identifier leur nature, en accord avec le cahier des charges. Au sein de ces fichiers, des règles d'identification des variables créées peuvent aussi être établies et décrites.

Pour quoi faire?

L'objectif est la bonne maitrise des traitements effectués (traçabilité) et la bonne compréhension des données disponibles et délivrables (normalisation du cadre informationnel).

A quel moment du cycle de vie ?

Tout au long du cycle de vie.

Exemples de bonnes pratiques

Dans le cas des fichiers

- Les noms des fichiers peuvent refléter les traitements qui ont été opérés. Par exemple, si le fichier d'origine s'appelle NomFichier.dta, la version "vérifiée" de ce fichier pourra simplement porter la lettre « V » en suffixe (NomFichier_V.dta) tandis que dans sa version "épurée", il portera le suffixe « E » (NomFichier_E.dta). Ceci permet une identification immédiate de la nature et du contenu des fichiers.

- Donner des noms de programmes explicites permet aussi de maitriser leur contenu. Par exemple, le programme de vérifications pourra se nommer "Verif.do". Cela facilitera la bonne maîtrise et la cohérence de l'information entre le programme et le fichier créé.

- Il peut aussi être intéressant d'ajouter le numéro de version dans son nom : Nom_Vx.y.z.dta avec x le numéro de réception, y le numéro de vérification et z le numéro de normalisation. Ces conventions permettent en un clin d'œil de savoir si un fichier est à jour ou non. (cf figure 11).

Dans le cas des programmes

- Dans le cas où on n'utilise pas un outil de gestion de versions, on peut donner un numéro de version incrémental aux programmes.

- Nous pouvons aussi penser à accompagner nos différentes versions d'explications. Par exemple, joindre un fichier (texte ou XML comme dans l'exemple de la figure 11) dans lequel nous donnons des informations sur sa création. Lorsqu'une nouvelle variante de la donnée est en cours de construction, il suffit de faire une copie du fichier de sa version précédente, et d'y ajouter les informations concernant les changements effectués.

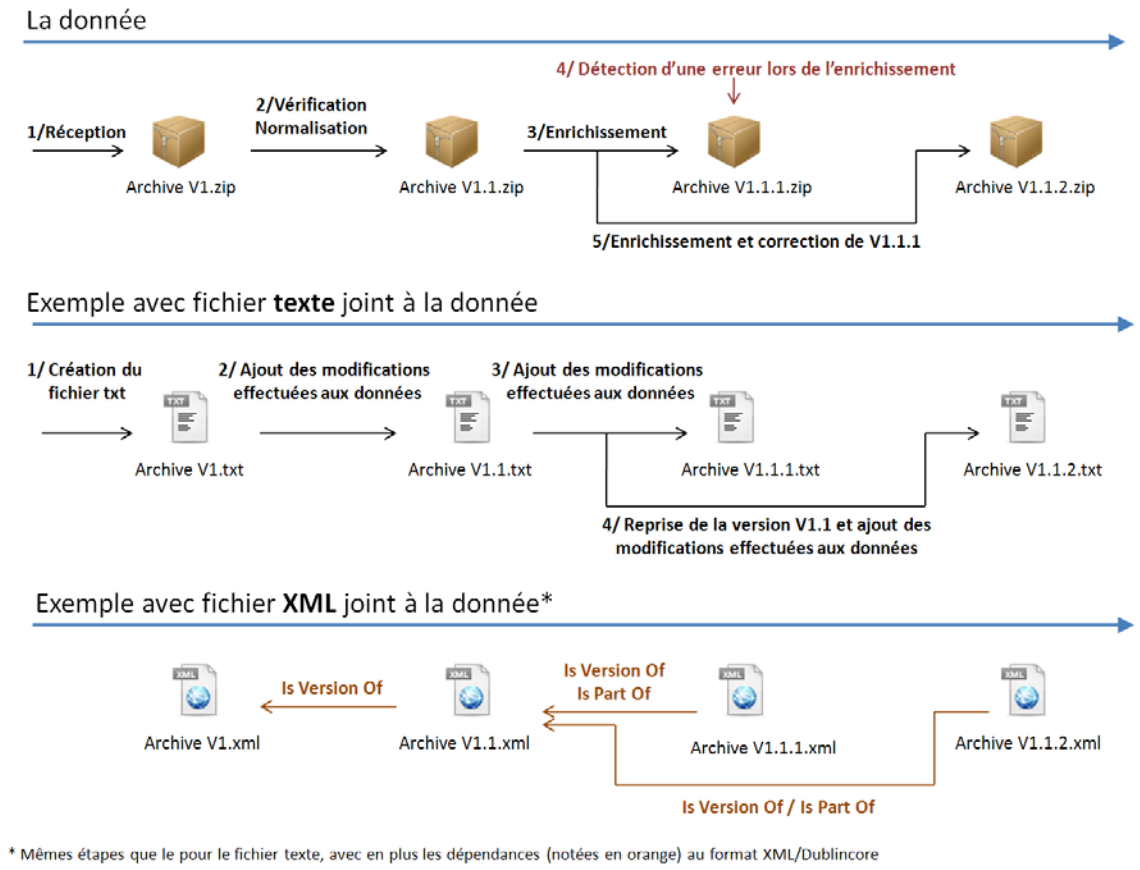


Figure 11 : XML (eXtensive Markup Language)

- On peut aussi préfixer les noms de programme par des numéros indiquant des "couches logiques" en termes de production de données. L'avantage est que la liste alphabétique des programmes reflète visuellement leur enchaînement.

Exemple :
- 01-xxx.sas
- 02-yyy.sas
- 02-zzz.sas
- 03-kkk.sas

xxx (couche 01) doit être exécuté avant yyy et zzz (couche 02),
mais l'ordre d'exécution de yyy et zzz importe peu. kkk doit être exécuté après yyy et zzz.

Dans le cas des variables

- Il faut se donner des conventions pour distinguer les variables d'origine des variables créées. Ces conventions sont dépendantes des outils utilisés. Pour des logiciels sensibles à la casse (comme Stata, R, Matlab par exemple), on peut se donner la convention suivante : les variables d'origine commencent par une minuscule, celles créées par une majuscule. D'autres conventions sont possibles comme par exemple ajouter un suffixe ou préfixe au nom de la variable créée, ou encore d'indiquer cette information dans le label de la variable.

Dans le cas de procédures ou de programmes

- Des programmes comme Mathematica ont pris l'option de mettre une majuscule au début de toutes leurs commandes et fonctions préprogrammées, afin que l'utilisateur sache distinguer immédiatement si la procédure ou la fonction utilisée est d'origine ou pas. Stata, comme beaucoup d'autres logiciels, permet de faire la distinction majuscule/minuscule ce qui est d'une grande aide et peut donc servir pour les conventions à adopter (noms de procédures/noms de variables).

Exemples d'outils

TheRename qui permet de renommer des fichiers ou des répertoires « par lot » en utilisant (ou pas) des expressions régulières. Ce programme permet également de récupérer et d'utiliser des caractéristiques de fichier pour s'en servir dans le nom des fichiers. Par exemple, les informations de date de création, les caractéristiques de prise de vue (contenus dans l'EXIF) pour les fichiers de photos ou de taille de fichier peuvent être utilisées.
http://www.herve-thouzard.com/the-rename

Total Commander permet d'affecter automatiquement des numéros incrémentés au nom d'une liste de fichiers, ou d'y adjoindre la date et autres caractères (préfixes, suffixes, etc..).
(http://www.ghisler.com/accueil.htm)

C'est quoi?

La maîtrise de l'enchaînement des étapes de traitements correspond à la maîtrise de l'ordre des traitements et de leurs contenus. Une personne gérant des données doit impérativement être capable de retrouver rapidement dans quel programme a eu lieu tel traitement (**traçabilité**). Il est en effet indispensable de savoir quel traitement a été opéré sur un fichier et quel traitement ne l'a pas été. L'exemple type est celui de la conversion de Francs en Euros, qui si elle est appliquée deux fois, peut avoir des conséquences dommageables[18].

Pour quoi faire?

Les traitements réalisés et l'ordre des étapes réalisées (si plusieurs programmes ou scripts sont écrits) doivent être maîtrisés. Une façon de maîtriser l'enchaînement des étapes de traitements des données peut-être de bien documenter les programmes et de documenter leur enchaînement.

A quel moment du cycle de vie ?

Tout au long du cycle de vie.

[18]Ceci n'arrive pas si les règles décrites ici sont suivies, notamment les règles de nommage et/ou de création de nouvelles variables.

Exemples de bonnes pratiques

- Chaque programme doit être correctement documenté. Les grandes étapes doivent être clairement identifiées. L'écriture d'un en-tête commentant ce que réalise le programme, et précisant la liste des fichiers utilisés en entrée et le nom de celui ou ceux produits en sortie est utile. De succincts commentaires doivent également être écrits tout au long du programme.
- L'écriture d'un schéma de l'enchaînement des programmes et des créations de fichiers permet de visualiser très rapidement les grandes étapes réalisées. Le minimum est d'écrire ce diagramme manuellement.
- Dans le cas où la chaîne des traitements nécessite plusieurs programmes, un programme général peut être créé, appelant chacun d'entre eux dans leur ordre d'exécution.
- Etablir un diagramme de Gantt. Le diagramme de Gantt est un outil utilisé en ordonnancement et gestion de projet et permettant de visualiser dans le temps les diverses tâches liées composant un projet. Il s'agit d'une représentation graphique représentant en abscisse les unités de temps (exprimées en mois, en semaine ou en jours) et en ordonnée les différentes tâches effectuées.
- Une façon de maîtriser le contenu des étapes est de donner des noms explicites aux programmes (cf. la fiche n°5 sur les règles de nommage).

Exemples d'outils

Le logiciel SAS établit un graphe du « Process Flow » (cf. figure 12)
SAS Logiciel de statistiques payant.
Site officiel : http://www.sas.com/software/sas9/

Le logiciel Dia (freeware) permet de créer des diagrammes très facilement (exemple sur la figure 13). Des conventions simples seront choisies notamment au niveau des couleurs et des formes (Par ex. les programmes sont représentés par des cercles bleus, les fichiers créés par des rectangles jaunes...)
Dia : http://projects.gnome.org/dia/

D'autres logiciels sont disponibles pour créer des diagrammes: Uml, Xmind, Freemind ou Lucid chart

Uml : Méthode de modélisation.
Page Wiki :
https://fr.wikipedia.org/wiki/Unified_Modeling_Language

Freemind : Logiciel de création de cartes heuristiques (cf figure 14)
Site officiel :
http://freemind.sourceforge.net/wiki/index.php/Main_Page

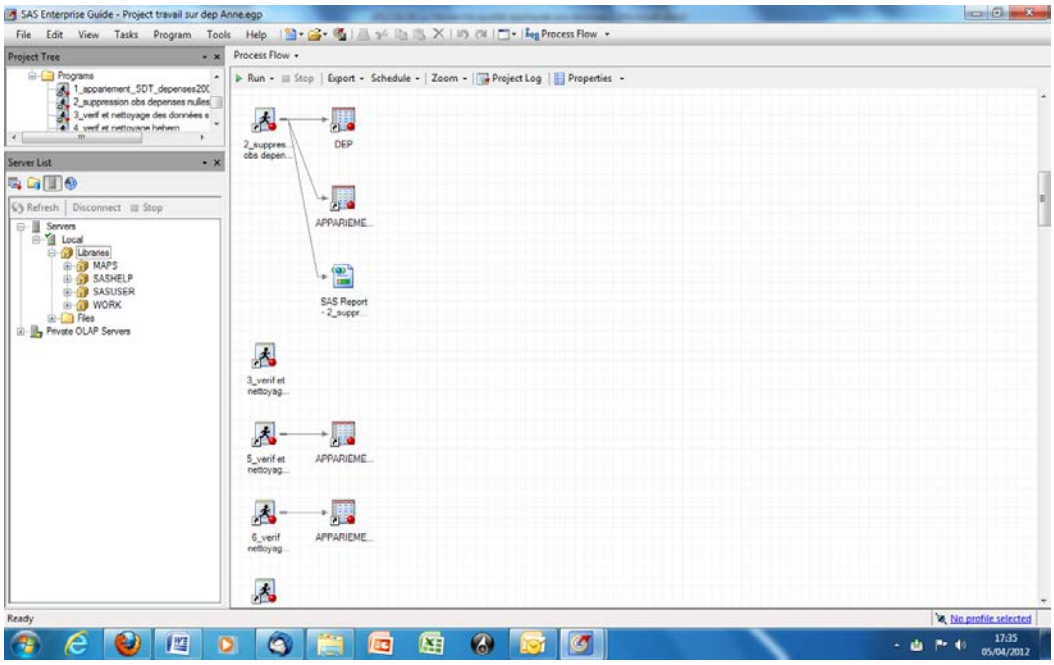


Figure 12 : Exemple de chaîne de traitements dans SAS

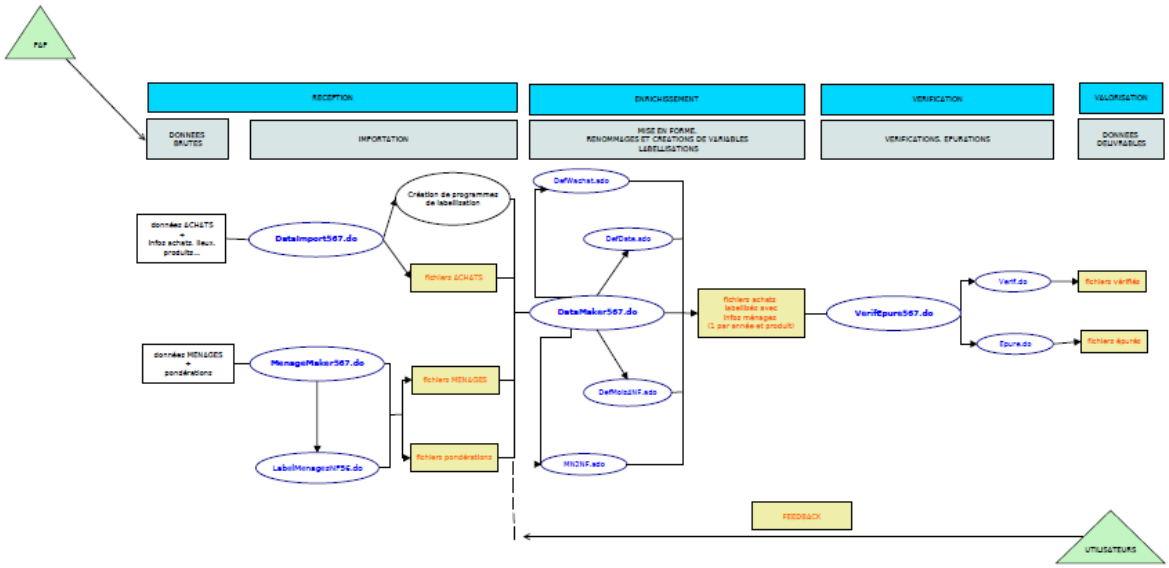


Figure 13 : Exemple de chaine de traitements illustrée avec DIA

.

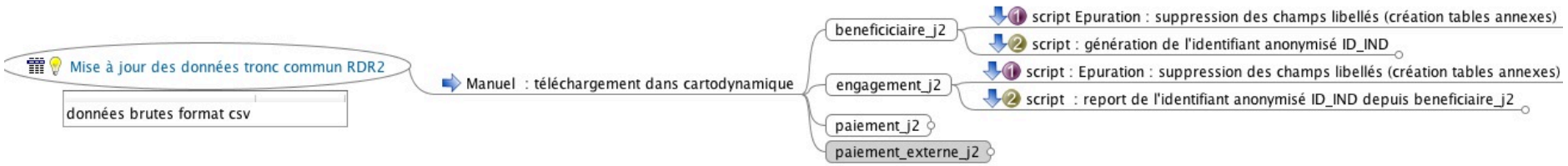


Figure 14 : Exemple de chaine de traitements illustrée avec FreeMind

C'est quoi?

Le *feedback* consiste à organiser l'écoute et la prise en compte des remarques ou des problèmes rencontrés par les utilisateurs des données (chercheurs, partenaires), à y répondre, afin d'intégrer leurs remarques et d'améliorer la chaîne des traitements. Un point important de la démarche qualité consiste à organiser l'amélioration constante des procédures de traitement.

Une fois les données délivrables créées et disponibles aux utilisateurs, il est donc nécessaire de mettre en place une organisation des remarques des utilisateurs. Ces retours peuvent être liés à des problèmes rencontrés sur les données (erreurs sur les données brutes ou erreurs commises dans la chaîne des traitements[19]), à des questionnements sur la nature des données, à des modifications de variables ...

Pour quoi faire?

Le *feedback* a pour objectif de signaler et corriger les erreurs repérées dans les données. Il assure la bonne communication de notre connaissance autour des données, et améliore les traitements effectués.

A quel moment du cycle de vie ?

Une fois les données délivrables créées et disponibles aux utilisateurs

Exemples de bonnes pratiques

- Une idée simple est d'organiser régulièrement des réunions autour des données avec les chercheurs concernés.

- Une autre idée est d'organiser un stockage centralisé des questions ainsi qu'un archivage des Questions/Réponses (FAQ).

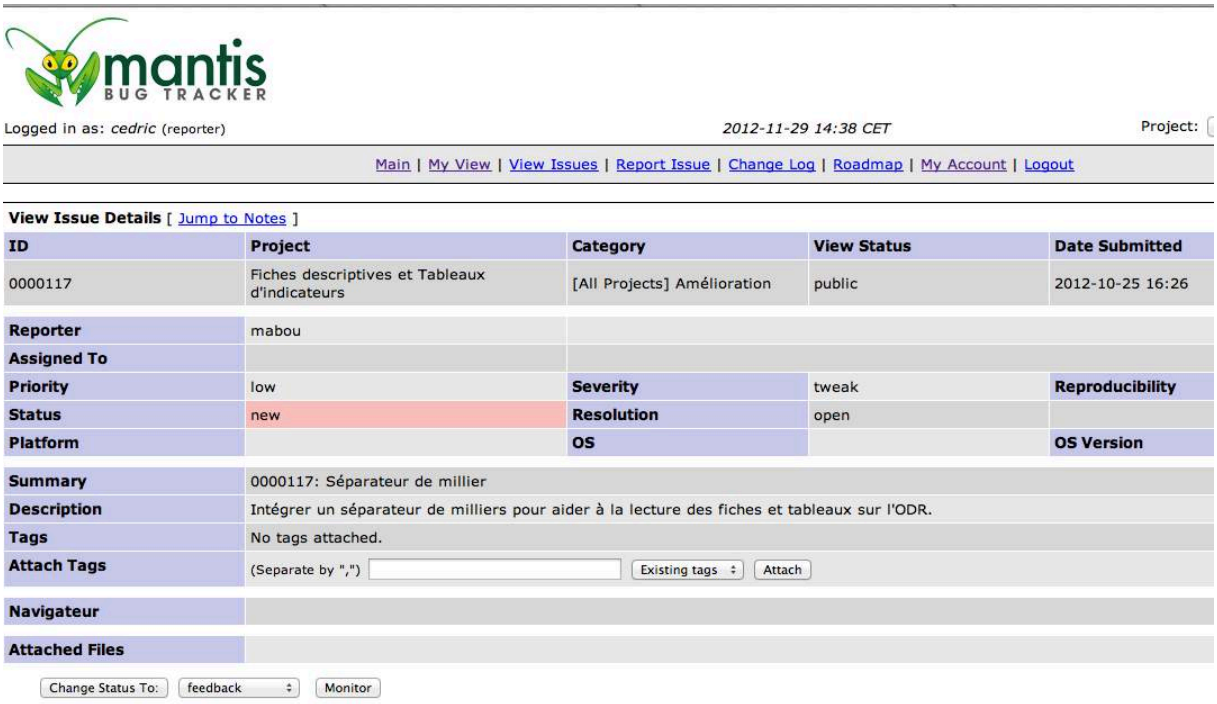
Exemples d'outils

- Sur des plateformes, les outils de discussion en ligne pourront être mis en place, signalés et organisés. Des outils Web (CMS, WIKIs, Blogs) permettent en effet aux utilisateurs de poster leurs questions, programmes, problèmes avec beaucoup de détail (upload des programmes). Des outils de signalement et de gestion des erreurs pourront également être utilisés (Bugzilla, Bug Tracker, RedMine, cf figure 15).

> **CMS** : Système de gestion de contenu
> Liste de CMS :
> https://fr.wikipedia.org/wiki/Liste_de_syst%C3%A8mes_de_gestion_de_contenu
>
> **WIKI** : Écriture collaborative de documents numériques.
> Liste de Wiki :
> https://en.wikipedia.org/wiki/List_of_wiki_software

- Des listes de diffusion pourront également être mises en place afin de prévenir les utilisateurs de problèmes, corrections ou autre. Pour les plateformes, des Newsletters ou des mails automatiques pourront avertir les utilisateurs d'une nouveauté sur la plateforme.

[19]Normalement, lorsqu'une erreur est commise dans la chaîne des traitements, si les bonnes pratiques ont été mises en place, il doit être possible de les identifier très vite (traçabilité) et surtout de recréer les données délivrables corrigées très vite (reproductibilité).



mantis BUG TRACKER

Logged in as: *cedric* (reporter) 2012-11-29 14:38 CET Project:

Main | My View | View Issues | Report Issue | Change Log | Roadmap | My Account | Logout

View Issue Details [Jump to Notes]

ID	Project	Category	View Status	Date Submitted
0000117	Fiches descriptives et Tableaux d'indicateurs	[All Projects] Amélioration	public	2012-10-25 16:26

Reporter	mabou			
Assigned To				
Priority	low	**Severity**	tweak	**Reproducibility**
Status	new	**Resolution**	open	
Platform		**OS**		**OS Version**

Summary	0000117: Séparateur de millier
Description	Intégrer un séparateur de milliers pour aider à la lecture des fiches et tableaux sur l'ODR.
Tags	No tags attached.
Attach Tags	(Separate by ",") Existing tags Attach
Navigateur	
Attached Files	

Change Status To: feedback Monitor

Figure 15 : Exemple de Bug Tracker, outil en ligne de signalement et de suivi d'erreur

C'est quoi?

Afin d'assurer la sécurité et la pérennité des données sous toutes leurs formes, il est important de bien choisir l'endroit où les entreposer et de procéder à des opérations de *sauvegarde*. Une opération de *sauvegarde* consiste à effectuer une copie d'un ensemble cohérent de données, tel qu'il existe à un moment bien identifié dans le temps.

Pour quoi faire?

Respecter les règles de sécurité vise à se prémunir des risques de détérioration des données. La sauvegarde vise à anticiper ces risques de perte ou de corruption de données (effacement par erreur, corruption par bogues, pannes matérielles, catastrophes) et donc à pouvoir restaurer une image cohérente des données après de tels incidents.

A quel moment du cycle de vie ?

A définir dès la réception des données.

Exemples de bonnes pratiques

- S'assurer au minimum que l'ordinateur sur lequel les données seront entreposées n'est pas facilement violable, qu'il est sécurisé au niveau de ses accès et par rapport aux attaques malveillantes et que les données sont régulièrement sauvegardées en plusieurs exemplaires et/ou sur différents supports et que ces *sauvegardes* ne se trouvent pas toutes au même endroit en cas de vol ou d'incendie.
- L'ordinateur où les données sont entreposées est connecté à un réseau sécurisé au niveau des accès extérieurs et des transferts de données en local et entre sites distants (VPN, SSH, SFTP, HTTPS…). En outre, il est préférable qu'il soit installé dans une pièce ou une baie climatisée disposant d'une alerte incendie et qu'il soit branché sur un réseau électrique relié à un onduleur. Si les données doivent être accessibles en permanence, vérifier également que le site dispose d'un groupe électrogène de secours en cas de panne.
- Le système (software) qui héberge les données est sécurisé en accès local et distant par un système de login/mot de passe (par ex. LDAP). Les communications distantes sont chiffrées (SSL). Les logiciels de sécurité (pare-feu, antivirus, antispyware, détection des connexions malveillantes …) ainsi que le système d'exploitation sont mis régulièrement à jour. Si le système héberge d'autres applications, son arborescence est sécurisée par des droits restrictifs pour les utilisateurs.
- Sur les ordinateurs (ou disques durs) des utilisateurs, il est souhaitable de stocker les données sur des partitions cryptées (TrueCrypt, cryptage système …). Cela permet en cas de perte, de vol ou d'intrusion, de garder les données invisibles et inexploitables par un tiers.
- L'ordinateur et les bases qui y sont installées doivent faire l'objet de sauvegardes régulières avec possibilité de récupérer aisément les données perdues et les versions antérieures le cas échéant.
- Les sauvegardes doivent être dupliquées à au moins deux endroits différents[20] et, le cas échéant, entreposées dans un coffre fermé et ignifugé.

Exemples d'outils

TrueCrypt : Logiciel de cryptage de fichiers / partitions.
Site officiel : http://www.truecrypt.org/

[20] Attention cette sauvegarde ne peut pas s'effectuer dans certains cas. En effet, certaines conventions de recherche imposent de ne disposer que d'un seul exemplaire des données.

C'est quoi ?

Dans le cas de données soumises à des règles d'utilisation (contrat, convention) définies par le propriétaire des données ou une instance décisionnelle (partenaire, CNIS, CNIL), toute personne responsable de ces données devra se donner les moyens de faire respecter ces règles. Les règles peuvent concerner l'utilisation des données (qui a le droit et pour quoi faire), leur accès, leur conservation (ou destruction) ainsi que leur citation. En outre, la personne en charge des données devra tracer les utilisations qui en sont faites, pour pouvoir éventuellement en rendre compte auprès du propriétaire.

Pour quoi faire ?

Pour faire respecter ces règles de confidentialité, les utilisateurs qui doivent être répertoriés précisément, devront en être informés. Des procédures adaptées de stockage, diffusion et utilisation de ces données devront être prises.

A quel moment du cycle de vie ?

L'ensemble des contraintes d'usage ou de restriction d'utilisation potentielle ou supposée des données doit être discuté avec leurs propriétaires, en amont de toute réception.

La réflexion sur l'utilisation des données, leur accès, leur conservation et leur citation devra aussi se faire avant ou au moment de la réception des données (dans tous les cas, avant toute mise à disposition).

Exemples de bonnes pratiques

- Avant la mise à disposition des données, organiser une réunion avec les utilisateurs précisant clairement les règles. Ces règles, si elles ne sont pas déjà précisées dans une convention ou contrat, devront être écrites dans un document diffusable (éventuellement un fichier Licence.txt).

- Une idée (parfois imposée par certains propriétaires de données) est de faire signer une charte aux utilisateurs sur leur engagement à respecter et faire respecter ces règles (par ex. par les stagiaires…).

- Enregistrer l'ensemble des utilisateurs de ces données : nom de l'utilisateur, date de mise à disposition des données, nom du (ou des) fichier(s) délivré(s). De façon minimale, un fichier type tableur sera alimenté. De façon plus élaborée et dans le cas de plateforme de téléchargement, des logs seront stockés à chaque téléchargement.

- Si des restrictions s'appliquent à l'utilisation des données, une vérification *ex-post* (récapitulatif des travaux les ayant utilisées) ou une demande *ex-ante* (examen d'un projet) devra être faite.

- Dans chaque document valorisant les données, citer systématiquement les sources exactes des données utilisées. La façon de citer les sources doit être indiquée dans le document expliquant les règles d'utilisation.

Pense-bête pour des bonnes pratiques de gestion et de préparation des données

Ce pense-bête, présenté sous forme de questions chronologiques, se donne pour objectifs de :

1) Vérifier que l'on n'a rien oublié dans la démarche de gestion des données
2) Donner des repères pour la mise en œuvre de bonnes pratiques permettant d'auto-évaluer nos pratiques de gestion des données.

Les réponses peuvent être les suivantes O : oui, N : non, SO : sans objet, NSP : ne sait pas.

Une progression dans la démarche qualité se mesure notamment par la réduction, voire la disparition, du nombre de réponses NSP à ce pense-bête.

Réception		*O*	*N*	*SO*	*NSP*
1	**Ces données ont-elles été fournies avec un cahier des charges ?**	☐	☐	☐	☐
2	**Avez-vous reçu des informations complémentaires à ces données ?**	☐	☐	☐	☐
	Type de données				
	Sur le Contexte du recueil ?	☐	☐	☐	☐
	Sur l'échantillon ?	☐	☐	☐	☐
	Un dictionnaire des variables vous a t'il été fourni ?	☐	☐	☐	☐
	Un questionnaire d'enquête ?	☐	☐	☐	☐
	Statut				
	S'agit-il de données publiques (diffusables) ?	☐	☐	☐	☐
	Existe-t-il une convention pour leur utilisation ?	☐	☐	☐	☐
	Une liste d'utilisateurs autorisés ?	☐	☐	☐	☐
	Sont-elles soumises au comité du secret (CNIS) ?	☐	☐	☐	☐
	Ont-elles nécessité une déclaration CNIL ?	☐	☐	☐	☐
	Y a-t-il une exigence de dépôt dans un endroit sécurisé ?	☐	☐	☐	☐
	Format				
	Connaissez-vous leur encodage (ascii, ebcdic, ansi, utf8,...) ?	☐	☐	☐	☐
	Connaissez-vous leur stockage (type de fichier : plat, bdd, propriétaire) ?	☐	☐	☐	☐
	Connaissez-vous le support original (CD, DVD, clé, mail, ftp, web,…) ?	☐	☐	☐	☐
	Connaissez-vous la taille de ces données ?	☐	☐	☐	☐
	Le nombre de fichiers ?	☐	☐	☐	☐
	Réception				
	Connaissez-vous la date de réception des données ?	☐	☐	☐	☐

	Connaissez-vous la date de création des données ?	☐	☐	☐	☐
	Est-ce une première réception ?	☐	☐	☐	☐
	Si non : est-ce une mise à jour ?	☐	☐	☐	☐
	Si non : est-ce une correction complète ?	☐	☐	☐	☐
	Si non : est-ce une correction partielle ?	☐	☐	☐	☐
3	**Existe-t-il des métadonnées liées à ces données ?**	☐	☐	☐	☐
	Si oui : Elles étaient livrées avec	☐	☐	☐	☐
	Si oui : Elles ont été créées	☐	☐	☐	☐
4	**Avez-vous fait une sauvegarde du support original ?**	☐	☐	☐	☐
	Avez-vous doublé cette sauvegarde ?	☐	☐	☐	☐
	Cette 2ème sauvegarde est-elle dans un endroit différent ?	☐	☐	☐	☐
5	**Avez-vous une convention de nommage des fichiers sauvegardés ?**	☐	☐	☐	☐
	Est-elle portée à la connaissance des utilisateurs ?	☐	☐	☐	☐
6	**Les données ainsi réceptionnées sont-elles destinées à être archivées ?**	☐	☐	☐	☐
	Savez-vous qui en est le responsable ?	☐	☐	☐	☐
Vérification		*O*	*N*	*SO*	*NSP*
7	**Existe-t-il un cahier des charges pour cette étape ?**	☐	☐	☐	☐
8	**Interdisez-vous les modifications manuelles ?**	☐	☐	☐	☐
9	**Les données sont-elles intégrées dans un SI existant ?**	☐	☐	☐	☐
	L'intégration est-elle de la responsabilité d'une seule personne ?	☐	☐	☐	☐
	Disposez-vous d'un système de suivi de l'intégration des données (traçabilité) ?	☐	☐	☐	☐
	Est-ce que le système de suivi est consultable ?	☐	☐	☐	☐
	Le SI doit-il conserver en ligne plusieurs versions des données ?	☐	☐	☐	☐
10	**Avez-vous fait les premières vérifications (interne) ?**	☐	☐	☐	☐
	Avez-vous vérifié que vous pouvez les lire ?	☐	☐	☐	☐
	Avez-vous vérifié les nb lignes, nb colonnes ?	☐	☐	☐	☐
	Avez-vous identifié les variables ?	☐	☐	☐	☐
	Avez-vous fait des statistiques et tests élémentaires ?	☐	☐	☐	☐
11	**Les données peuvent-elles être vérifiées par rapport à un système de référence externe (normalisation) ?**	☐	☐	☐	☐
	Avez-vous mis en place une procédure de normalisation (tests d'intégrité) ?	☐	☐	☐	☐
12	**Avez-vous une convention de nommage de ces nouveaux fichiers ?**	☐	☐	☐	☐
13	**Existe-t-il des métadonnées liées à chaque nouvelle table de données ?**	☐	☐	☐	☐

14	**Chaque nouvelle table est-elle destinée à être archivée ?**	☐	☐	☐	☐
15	**Avez-vous fait une sauvegarde de vos procédures (scripts, programmes) de vérifications ?**	☐	☐	☐	☐
	Vos procédures sont-elles bien reproductibles ?	☐	☐	☐	☐
	Vos procédures et tables de données suivent-ils des règles de nommage cohérentes ?	☐	☐	☐	☐
Enrichisse-ment		*O*	*N*	*SO*	*NSP*
16	**Existe-t-il un cahier des charges définissant les variables complémentaires à créer ?**	☐	☐	☐	☐
17	**Interdisez-vous les modifications manuelles ?**	☐	☐	☐	☐
18	**Avez-vous une convention de nommage de ces variables complémentaires ?**	☐	☐	☐	☐
19	**Disposez-vous d'un système de suivi de ces traitements complémentaires (logs) ?**	☐	☐	☐	☐
	Est-ce que le système de suivi est consultable ?	☐	☐	☐	☐
20	**Existe-t-il des métadonnées liées à cette nouvelle table de données ?**	☐	☐	☐	☐
21	**Cette nouvelle table est-elle destinée à être archivée ?**	☐	☐	☐	☐
22	**Avez-vous fait une sauvegarde des données enrichies?**	☐	☐	☐	☐
23	**Avez-vous fait une sauvegarde de vos procédures (scripts, programmes) d'enrichissement ?**				
	Vos procédures sont-elles bien reproductibles ?	☐	☐	☐	☐
	Vos procédures et tables de données suivent-ils des règles de nommage cohérentes ?	☐	☐	☐	☐
Diffusion		*O*	*N*	*SO*	*NSP*
24	**Dans le cas où le fournisseur le désire, avez-vous la possibilité de connaître les destinataires de ses données ?**	☐	☐	☐	☐
25	**Dans le cas où le fournisseur le désire, avez-vous la possibilité de connaître les utilisations de ses données ?**	☐	☐	☐	☐
26	**Existe-t-il des règles d'obtention de ces données livrables ?**	☐	☐	☐	☐
	Lettre d'engagement ?	☐	☐	☐	☐
	Demande écrite à un comité responsable des données ?	☐	☐	☐	☐
	Suivi des téléchargements ou des envois ?	☐	☐	☐	☐
Feed-back		*O*	*N*	*SO*	*NSP*
27	**Existe-t-il une procédure pour faire remonter les remarques et les erreurs détectées par les utilisateurs ?**	☐	☐	☐	☐
28	**Avez-vous intégré les remarques et erreurs signalées par les utilisateurs ?**	☐	☐	☐	☐
29	**Existe-t-il un moyen de diffusion de ces modifications ?**	☐	☐	☐	☐

Données brutes : Jeu de données initial provenant d'une source extérieure constituant la matière première sur laquelle vont s'appliquer les différents traitements. Ce jeu de données est souvent accompagné de documents descriptifs, clauses d'utilisation et de dictionnaires qui serviront à enrichir le jeu de métadonnées associé. Ces données doivent être archivées[21] et ne doivent subir aucune modification.

Données délivrables : Jeu de données issu d'un ou de plusieurs traitements effectué(s) en suivant un cahier des charges plus ou moins formel à partir de données brutes. Ce fichier est accompagné de métadonnées. Ce jeu de données est plus connu des chercheurs sous le nom de "fichier(s) de travail" puisque le processus de recherche empirique est réalisée à partir de ces fichiers exploitables. Ces fichiers peuvent être reproduits à l'identique ou modifiés en suivant la chaine des traitements décrits dans la démarche qualité.

Intégrité : Unicité matérielle et temporelle des fichiers, des procédures et des programmes.

Normalisation : La normalisation consiste à :

- identifier, dans un fichier de données, les attributs qui correspondent à une ou des clés référentielles pour ces données ; ce peut être par exemple un identifiant d'individu, un repérage géographique
- vérifier l'intégrité référentielle, c'est-à-dire que ces clés référentielles correspondent aux clés des données de référence existantes par ailleurs
- détecter d'éventuelles redondances ou incohérences avec les données existantes.

Ces opérations permettent de rapprocher celles-ci d'autres informations existantes ou à venir, pour en accroitre les possibilités de valorisation, par exemples pour effectuer des jointures avec d'autres données disponibles, agréger les données selon des règles pré-établies (agréger les communes d'un canton)…

Ontologie : Ensemble structuré des termes et concepts représentant le sens d'un champ d'information. Il existe des logiciels permettant de créer des ontologies autour de champs disciplinaires ou de thématiques (arbres heuristiques). L'ontologie est particulièrement utile dans le traitement de données textuelles, mais peut permettre aussi de définir et représenter les liens sémantiques entre des données, facilitant ainsi la compréhension globale du contenu des données par rapport au champ ou à la thématique auquel elles se rattachent.

Thésaurus : Outil linguistique recensant les termes contenus (en langage naturel) dans un ensemble de ressources (jeux de données, documents) et permettant leurs indexations. Un thésaurus a pour vocation de faciliter la recherche de documents ou données par un utilisateur : par exemple permettre, via un moteur de recherche informatique, une recherche de données par mots clefs sur un ensemble de jeux de données.

Stockage, Sauvegarde, Archivage
Ce sont des éléments de la gestion du cycle de vie de l'information (ILM *i.e* .Information Life-cycle Management)

[21] En fonction des clauses de leur délivrance (règles de diffusion, confidentialité, usage à durée limitée. Voir aussi confidentialité).

Les données, pour être traitées, doivent avoir subi une opération de **stockage** sur un support informatique.
Il s'ensuit un certain nombre de questions :

- Capacité, coût du stockage
- Disponibilité, intégrité des informations
- Performance
- Besoin de chiffrement de données sensibles

La **sauvegarde** doit répondre au risque de perte ou de corruption de données (effacement par erreur, corruption par bogues, pannes matérielles, catastrophes).
Une opération de sauvegarde réalise une copie d'un ensemble cohérent de données telles qu'elles existaient à un moment bien identifié dans le temps.
Les objectifs des opérations de sauvegarde sont la restauration des données, c'est-à-dire la persistance des données même après un incident majeur.
La sauvegarde ne suffit donc pas, il faut pouvoir, après une perte, restaurer une image cohérente des données.

L'archivage correspond à un ensemble de pratiques et d'outils utilisés pour conserver des informations stables ('mortes') et pouvoir les consulter quels que soient leurs dates, format et support matériel.

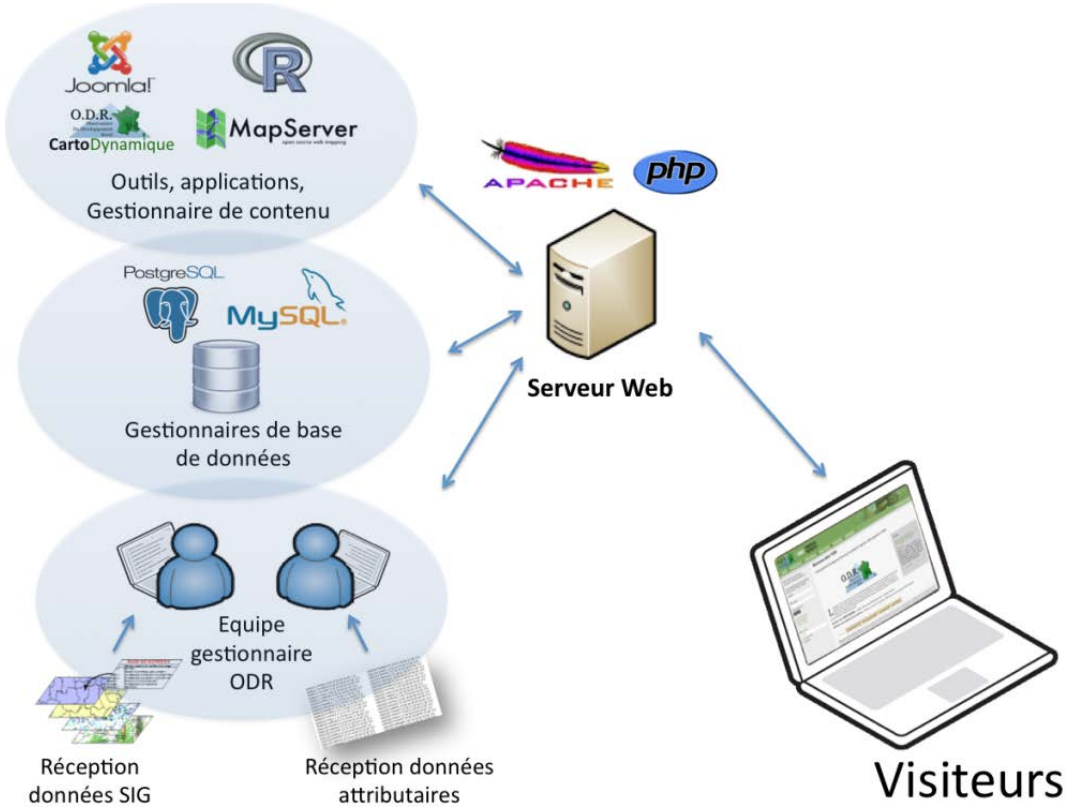
Sauvegarde	Archivage
Copie de données pour recours en cas d'incident	Conservation des données d'origine pour accès ultérieur
Les données sont remplacées périodiquement par celles des nouvelles sauvegardes	Aucune modification des données après archivage : données stables
Utilisation des données en cas de reprise après incident	
Les données relèvent de l'exploitation	Les données relèvent du domaine d'un projet

Système d'information : Un système d'information (SI) est défini comme un ensemble organisé de ressources humaines, logicielles, matérielles et données permettant de traiter, stocker et communiquer des informations. Outre l'aspect informatique[22], les spécificités d'un système d'information sont ses capacités d'interconnections et d'interopérabilités entre les ressources qui le composent.

Par exemple, l'Observatoire du Développement Rural (ODR) est un système d'information au sens qu'il regroupe un ensemble de ressources interopérables entre elles pour favoriser la conservation, le traitement et la diffusion de jeux de données en utilisant les technologies du web. Il se compose :

- d'une équipe de gestionnaires ODR
- d'un ensemble d'outils de stockage, traitement et diffusion de données connectés entre eux
- d'un serveur web facilitant la communication avec les utilisateurs.

[22] L'utilisation de moyens informatiques et de télécommunications sont à l'origine de la notion de système d'information.



(SIG = Système d'Information Géographique)

Dans le cadre de l'ODR, l'interopérabilité des données est assurée par le géocodage (référencement géographique) de l'ensemble des jeux de données. Celui-ci permet de rapprocher des données de sources et d'origines diverses à l'intérieur du système d'information.

Pour aller plus loin

American Statistical Association (1999). "*Ethical Guidelines for Statistical Practice.*" ASA. http://www.amstat.org/about/ethicalguidelines.cfm

ANU Data Management Manual "*Managing Digital Research Data at the Australian National University*"(2012) http://information.anu.edu.au/training_and_skills_development/information_literacy/resources/ANU_DM_Manual-v11.09.20_v2.pdf

Dublin Core Metadata Initiative Web site. http://dublincore.org/

Eurostat (2005), « *Code des bonnes pratiques de la statistique européenne pour les services statistiques nationaux et communautaires* » http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-77-07-026/FR/KS-77-07-026-FR.PDF

Groupes de travail communs UNECE/Eurostat/OCDE sur les métadonnées statistiques (METIS), 2009. *Modèle générique du processus de production statistique* (version 4.0)

Inter-university Consortium for Political and Social Research (ICPSR), 2012. *Guide to social science data preparation and archiving. Best practice throughout the data life cycle* (5th ed.). Ann Arbor, MI. http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Lyman, Peter and Hal R. Varian, (2003) "*How Much Information?*" University of California., Berkley http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Pienta, Amy M., Alter George , Lyle Jared (2010) *The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data* Inter-university Consortium for Political and Social Research (ICPSR) , Population Studies Center. Working paper. http://deepblue.lib.umich.edu/bitstream/2027.42/78307/1/pienta_alter_lyle_100331.pdf

Strasser, Carly, Cook, Robert; Michener, William; & Budden, Amber. (2012). "*Primer on Data Management: What you always wanted to know*". DataOne Working paper. http://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

TGE ADONIS, 2010. *Le guide des bonnes pratiques numériques. Entrepôt OAI-PMH* (version1) http://www.tge-adonis.fr/sites/default/files/ressourcesdoc/pdf_guide_oai10_vf.pdf

TGE ADONIS, 2011. *Le guide des bonnes pratiques numériques.* (version2) http://www.tge-adonis.fr/sites/default/files/ressourcesdoc/guide_des_bonnes_pratiques_v2.pdf

Van den Eynden V., Corti L., Woollard M., Bishop L., Horton L. 2011. *Managing and sharing data. Best practice for researchers* (3rd ed.) Universityt of Essex. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Vardigan, M., Heus, P. and Thomas, W. (2008) *Data documentation initiative: Toward a standard for the social sciences.* Int. J. Digital Cur., 3, 107-113.

Suivi des versions

N° Version	Date	Commentaires	Diffusion
V0.1	20/01/2012	Compilation des différentes contributions	Envoi à des relecteurs extérieurs (09/03/2012).
V0.2	01/06/2012	Incorporation des commentaires des relecteurs	Envoi à la délégation qualité, et au CGD.
V0.3	10/10/2012	Incorporation remarques F. Jacquet et discussion groupe; Changement de titre ("préparation" au lieu de traitement)	Interne au Groupe.
V1	21/12/2012	Modifications finales. Modification du pense-bête. Dernières modifications de forme.	SAE2 et Délégation Qualité INRA