LES PRINCIPES DE QUALITÉ DES DONNÉES

Arthur D. Chapman[1]

Although most data gathering disciples treat error as an embarrassing issue to be expunged, the error inherent in [spatial] data deserves closer attention and public understanding …because error provides a critical component in judging fitness for use. (Chrisman 1991).

GBIF
www.gbif.org

[1] Australian Biodiversity Information Services
PO Box 7491, Toowoomba South, Qld, Australia

Citation suggérée:

Chapman, A. D. (2005). *Les principes de qualité des données*, version 1.0. Trad. Chenin, N. Copenhague: Global Biodiversity Information Facility, 76 pp. Disponible en ligne sur http://links.gbif.org/gbif_qualite_donnees_manual_fr_v1.pdf

ISBN/DoI: non disponible.
Persistent URI: http://links.gbif.org/gbif_qualite_donnees_manual_fr_v1.pdf
Langage du document: Français (Titre original: *Principles of Data Quality*)
Date de publication de la traduction en Français: Avril 2011.
Date de publication de la version originale en Anglais: Juillet 2005

La traduction en Français de ce texte a été cofinancée par le Système Mondial d'Information sur la Biodiversité (GBIF) et la Direction Générale de la Coopération au Développement Belge (DGD) au travers du projet *Central African Biodiversity Information Network* (CABIN) coordonné par le Musée Royal d'Afrique Centrale (RMCA). Le GBIF France a aussi contribué à cette traduction.

GBIF DGD CABIN RMCA GBIF France

Dégagement de responsabilité: Même si tous les efforts ont été mis en œuvre par l'auteur et l'éditeur pour veiller à ce que le contenu de ce texte soit précis et à jour, tous les renseignements contenus ici sont donc présentés «tels quels», sans autre déclaration ou garantie de quelque nature que ce soit. Ni l'auteur ni l'éditeur sont en aucun cas responsable de tout dommage pouvant découler de l'utilisation de l'information contenue dans le présent texte.

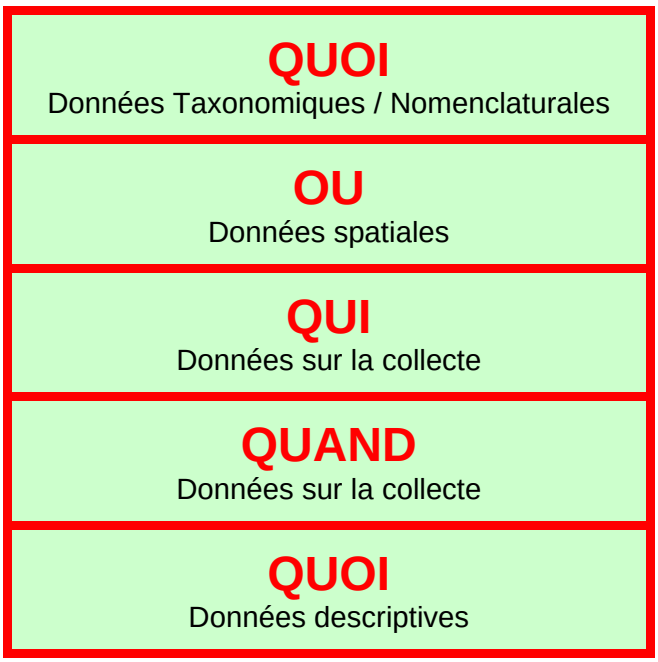
Couverture: GBIF Secretariat 2011. Photo par © Per de Place Bjørn, 2005. Amata phegea (Linnaeus 1758)

Contenue

Ce Chapitre est équivalent à :

Chapman, A. 2005. *Principles of Data Quality,* version 1.0. Copenhagen: Global Biodiversity Information Facility. 58 pp. ISBN: 87-92020-03-8 (disponible en fichier PDF autonome sur http://www.gbif.org)

Introduction



Les principes de qualité des données ont pris ces dernières années une place centrale dans des domaines tels que les affaires (SEC 2002), la médecine (Gad et Taulbee 1996), les SIG (Zhang et Good child 2002), la télédétection (Lunetta et Lyon 2004) et beaucoup d'autres ; mais ce n'est que récemment qu'ils sont progressivement acceptés partout par la communauté des muséums et de la taxonomie.

L'intensification rapide des échanges et la disponibilité croissante des données taxonomiques et d'occurrence d'espèces, accroît l'importance de ces principes, les utilisateurs de ces données étant de plus en plus exigeants quant à leur qualité. En effet, certains en dehors de la communauté des muséums considèrent la qualité des données que ceux-ci fournissent comme généralement insuffisante pour qu'elles puissent être utilisées dans l'aide à la décision en matière de conservation ; mais cette critique est-elle due à la qualité des données en elle-même ou bien à une documentation insuffisante de ces données ? Or, ces données sont d'une importance capitale. Accumulées sur une longue période, elles fournissent en effet une base de référence irremplaçable sur la biodiversité à une époque où l'humanité a sur celle-ci un impact inouï (Chapman et Busby 1994). Elles constituent une ressource essentielle pour tout effort de conservation de l'environnement, puisqu'elles fournissent le seul enregistrement pleinement documenté de la présence des espèces dans des zones qui peuvent avoir vu leur habitat modifié par le défrichement pour la mise en culture, l'urbanisation, le changement climatique, ou toute autre cause (Chapman 1999).

Voici quelques unes des idées que j'ai essayé de développer ci-dessous, tout en mettant en évidence un certain nombre de principes relatifs à la qualité des données qui devraient être au centre des préoccupations des muséums et des herbiers lorsqu'il fournissent leurs données à un large public.

La qualité des données et le taux d'erreur dans les données sont des aspects souvent négligés dans les bases de données environnementales, les systèmes de modélisation, les SIG, les

systèmes d'aide à la décision, etc. Trop souvent, les données sont utilisées sans un regard critique, sans prendre en considération les erreurs qu'elles recèlent, et ceci peut conduire à des résultats erronés, à des informations trompeuses, à des décisions environnementales déraisonnables et à des coûts accrus.

> *Les données sur les spécimens de plantes et d'animaux détenues dans les muséums et les herbiers constituent une vaste source d'information, indiquant où se trouvent les espèces non seulement aujourd'hui, mais aussi en remontant sur plusieurs siècles (Chapman et Busby 1994).*

De nombreux principes de qualité des données doivent être respectés lorsqu'on traite de données sur les espèces, particulièrement en ce qui concerne l'aspect spatial de ces données. Ces principes sont en jeu à toutes les étapes du processus de gestion des données. Une perte de qualité à une de ces étape quelle qu'elle soit réduit leur applicabilité et les usages pertinents que l'on peut en faire. Ces étapes comprennent :

- La saisie des données et leur enregistrement au moment de la collecte,
- La manipulation des données préalable à leur numérisation (la préparation des étiquettes, l'inscription des données dans un registre, etc.),
- L'identification de la collecte (spécimen ou observation) et son enregistrement,
- La numérisation des données,
- Leur documentation (saisie et enregistrement des métadonnées),
- Leur stockage et leur archivage,
- La présentation et la diffusion des données (publications imprimées ou électroniques, interface Web de bases de données, etc.),
- L'utilisation des données (analyse et manipulation).

Toutes ces étapes influent sur la qualité finale ou « aptitude à l'usage » des données et toutes concernent l'ensemble des aspects des données – la partie taxonomique ou nomenclaturale des données – le « quoi », l'aspect spatial - le « où » et d'autres aspects comme le « qui » et le « quand » (Berendsohn 1997).

Avant d'entrer dans une discussion sur la qualité des données et son application aux données d'occurrence d'espèces, il faut définir et décrire un certain nombre de concepts. A commencer par le terme de qualité lui-même, ainsi que les termes d'exactitude et de précision qui sont souvent employés à tort, et ce que nous entendons par données primaires sur les espèces et par données d'occurrence d'espèces.

Il ne faut pas sous-estimer l'élégance toute simple de l'amélioration de la qualité. Outre le travail en équipe, la formation et la rigueur, elle n'exige pas de talents particuliers. Quiconque le veut peut être un contributeur efficace.
(Redman 2001).

Définitions

Données d'occurrence d'espèces

La locution « données d'occurrence d'espèces » recouvre ici les données inscrites sur les étiquettes attachées aux spécimens ou aux lots de spécimens hébergés dans les muséums ou les herbiers, les données d'observation ou les données d'inventaires environnementaux. En général, la référence spatiale des données est un point, quoique le référencement spatial soit aussi parfois linéaire : données issues d'une enquête environnementale le long d'un transsect, collecte le long d'une rivière, polygonal : observations à l'intérieur d'une aire définie, comme un parc national, maillé : observations ou données d'enquête recueillies selon un maillage régulier.

En général on parle de données géoréférencées - c'est-à-dire des enregistrements dotés de références géographiques qui les relient à un endroit particulier – que ce soit sous forme de coordonnées géographiques (longitude et latitude, UTM) ou pas (description textuelle d'une localité, altitude, profondeur) - et à un moment particulier (date, heure dans la journée). Les données sont généralement aussi reliées à un nom taxonomique, mais on trouve aussi des spécimens non identifiés. La locution a occasionnellement été utilisée de manière interchangeable avec la locution « données primaires sur les espèces ».

Données primaires sur les espèces

La locution « données primaires sur les espèces » est utilisée pour décrire des données brutes de collecte et des données sans aucun attribut spatial. Y compris des données taxonomiques et nomenclaturales sans attribut spatial, comme des noms, des taxons ou des concepts taxonomiques dépourvus de références géographiques.

Exactitude et précision

Exactitude et *précision* sont régulièrement confondues et les différences ne sont généralement pas bien comprises. Les différences s'expliquent mieux au travers d'exemples (figure 1).

L'*exactitude* fait référence à la proximité des valeurs mesurées, des observations ou des estimées avec la valeur réelle ou vraie (ou avec une valeur qui est acceptée comme étant vraie – par exemple, les coordonnées d'un point de contrôle d'une enquête), comme illustré dans la figure 1.

La *précision* (ou *résolution*) peut être divisée en deux notions principales. La *précision statistique* est la proximité entre des observations répétées. Ceci n'a rien à voir avec leur relation avec la vraie valeur, et on peut avoir un haut degré de précision, mais une faible exactitude, comme illustré dans la figure 1a. La *précision numérique* est le nombre de chiffres significatifs avec lequel une observation est enregistrée : cette notion s'est renforcée avec l'avènement des ordinateurs. Par exemple une base de données peut fournir un enregistrement de latitude/longitude décimales avec une précision à 10 chiffres –c'est-à-dire une précision de 0,1 mm- alors qu'en réalité l'enregistrement a une résolution maximale de 10 à 100 m –soit 3 à 4 chiffres-. Ceci conduit souvent à une fausse impression à la fois de précision et d'exactitude.

Ces termes –exactitude et précision- peuvent aussi s'appliquer à des données non spatiales. Par exemple, une collecte peut avoir une identification au niveau de la sous-espèce (c'est à dire avec une grande précision), mais avec un taxon erroné (c'est à dire avec une faible exactitude), ou bien être identifiée correctement mais seulement au niveau de la famille (grande exactitude, mais faible précision).

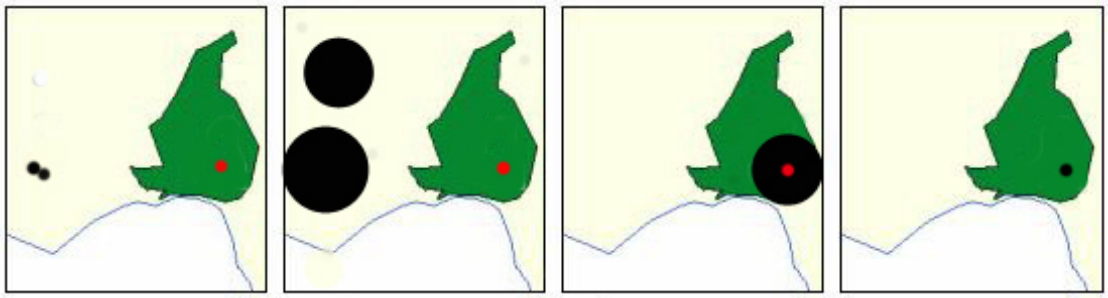


Fig. 1. *Montre les différences entre exactitude et précision dans un contexte spatial. Les points rouges indiquent les vraies localisations, les points noirs représentent les localisations rapportées par un collecteur.*

a. Haute précision, faible exactitude,
b. Faible précision, faible exactitude indiquant une erreur aléatoire,
c. Faible précision, haute exactitude,
d. Haute précision et haute exactitude.

Qualité

La notion de qualité, appliquée aux données, a des définitions diverses, mais dans le domaine géographique une définition est maintenant largement acceptée – celle de « aptitude à l'usage » (Chrisman 1983) ou « usage potentiel ». Cette définition a été adoptée par la plupart des standards modernes de transfert de données spatiales (ANZLIC 1996a, USGS 2004). Elle est aussi de plus en plus utilisée dans des domaines non spatialisés comme en économie ou dans le commerce et l'industrie. Certains (English 1999, par exemple) croient que cette définition « aptitude à l'usage » est un peu restrictive et ils militent pour une définition qui inclue aussi l'aptitude à des usages futurs ou potentiels.

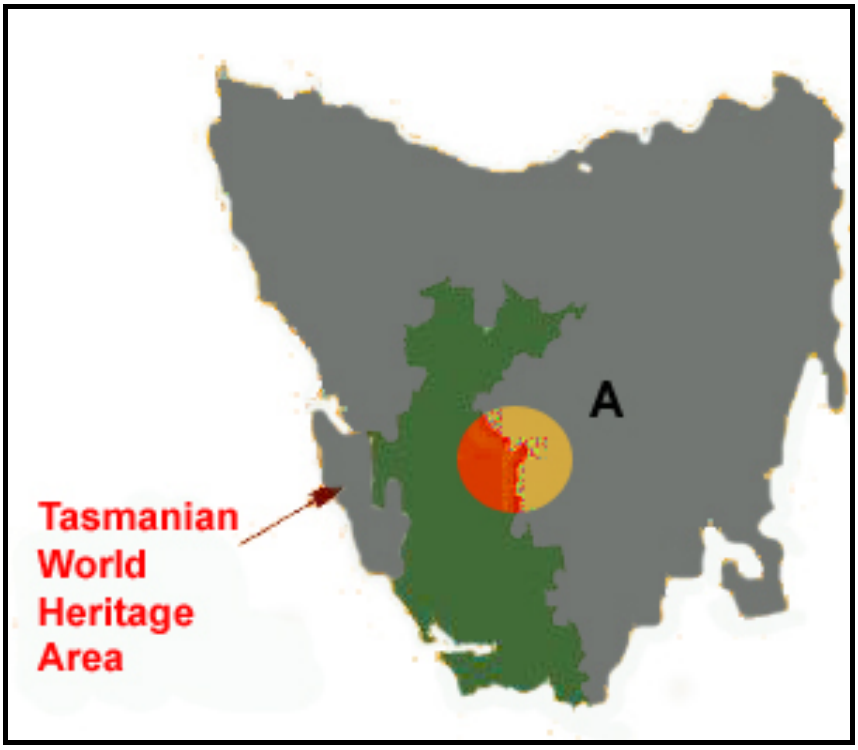


Fig. 2. *Carte de la Tasmanie, Australie, montrant un enregistrement (A) collecté avec une précision de 0,5} (environ 50 km) comme l'indique le cercle. La zone possible de la collecte (déterminée en utilisant le degré de précision) recouvre partiellement la zone de la Tasmanie classée au Patrimoine Mondial.*

La figure 2 montre un exemple d'utilisation du concept d'*aptitude à l'usage*. La collecte de l'espèce en question (marquée « A ») a une précision de 0,5° de latitude (soit environ 50 km). Si l'on prépare une liste des espèces de Tasmanie, et que l'on veut savoir si cette espèce se trouve en Tasmanie, alors l'enregistrement peut être utilisé pour répondre à cette question – la collecte est « apte à l'usage » et peut donc être considérée comme de bonne qualité dans ce contexte. D'un autre côté, si l'on veut savoir si cette espèce se trouve ou non dans la zone classée, alors l'enregistrement ne permet pas de répondre – la collecte peut s'y trouver ou pas. La donnée n'est pas « apte à l'usage » dans ce cas et est donc de faible qualité dans ce contexte. Les valeurs de latitude et longitude dans la base de données peuvent être très précises en apparence –codées avec de nombreuses décimales-, et ceci peut induire en erreur l'utilisateur de l'enregistrement si le degré réel de précision de la mesure ou de l'estimation n'est pas indiqué.

On rencontre des cas similaires avec les composantes non spatiales des données, lorsqu'une identification erronée par exemple, peut rendre la donnée de peu de valeur et ainsi inapte à l'usage. Si l'on étudie la distribution d'une espèce (ou encore sa physiologie ou son écologie, etc.), avoir un nom erroné attaché à un spécimen ou à une observation peut conduire à des résultats erronés et susceptibles d'induire en erreur ceux qui les utiliseraient.

La qualité des données est multidimensionnelle : elle implique la gestion des données, la modélisation et l'analyse, le contrôle et l'assurance qualité, le stockage et la présentation. Comme l'ont dit indépendamment Chrisman (1991) et Strnng *et al*. (1997), la qualité des données est liée à l'usage et ne peut pas être jugée indépendamment de l'utilisateur. Dans une base de données, les données n'ont pas de qualité ou de valeur intrinsèques (Dalcin 2004) ; elles n'ont qu'une valeur *potentielle* qui se *réalise* seulement quand quelqu'un utilise des

données pour faire quelque chose d'utile. La qualité de l'information est liée à sa capacité à satisfaire ses consommateurs et leurs besoins (English 1999).

Redman (2001) a suggéré que pour que des données soient aptes à l'usage il faut qu'elles soient accessibles, exactes, disponibles quand on en a besoin, complètes, cohérentes avec les autres sources, pertinentes, aussi exhaustives que possible, qu'elles fournissent un niveau de détail correct, et qu'elles soient aisées à lire et à interpréter.

Un problème qu'un gestionnaire de données doit prendre en compte est ce qui doit être fait pour la rendre utilisable par un public plus large (c'est à dire accroître son potentiel d'usage et sa pertinence) et ainsi étendre le spectre de ses domaines d'utilisation. Il faudra trouver là un compromis entre l'extension du potentiel d'usage et l'effort que cela demande : il peut par exemple être nécessaire d'éclater des champs, d'ajouter de l'information sur le géo-référencement, etc.

> *Les données sont de haute qualité si elles sont aptes à être utilisées dans le but qui a conduit à les recueillir, que ce soit pour l'aménagement, l'aide à la décision, ou la planification*
>
> (Juran 1964).

Assurance Qualité / Contrôle Qualité

La différence entre le contrôle qualité et l'assurance qualité n'est pas toujours claire. Taulbee (1996) fait la distinction entre Contrôle Qualité et Assurance Qualité et souligne que l'un ne peut exister sans l'autre si l'on veut atteindre les objectifs de qualité. Elle définit

- Le *Contrôle Qualité* comme un jugement de la qualité fondé sur des normes internes, des processus et procédures établis pour contrôler et vérifier la qualité, et
- L'*Assurance Qualité* comme un jugement de la qualité fondé sur des nomes externes au processus et l'audit des activités et des processus de contrôle qualité assurant que les produits finaux respectent les normes de qualité prédéfinies.

Dans une approche plus industrielle, Redman (2001) définit l'*Assurance Qualité* comme

> *Les activités qui sont conçues pour fournir des produits informationnels sans défaut qui satisfont les principaux besoins des principaux clients, au meilleur coût.*

Comment ces notions doivent être appliquées en pratique n'est pas clair, et dans la plupart des cas ces deux notions semblent utilisées de manière largement synonyme pour décrire l'ensemble des pratiques de la gestion de la qualité des données.

Incertitude

L'incertitude peut être comprise comme une « *mesure de l'incomplétude de nos connaissances ou de notre information sur une quantité inconnue dont la vraie valeur pourrait être établie si un dispositif de mesure parfait était disponible »* (Cullen et Frey 1999). L'incertitude est une caractéristique de la compréhension des données par l'observateur, et elle dépend plus de l'observateur que des données en elles-mêmes. Il y a toujours de l'incertitude dans les données ; la difficulté consiste à enregistrer, comprendre et

visualiser cette incertitude de sorte que les autres puissent aussi la comprendre. L'*incertitude* est le mot clé dans la compréhension et l'évaluation du risque.

Erreur

L'erreur recouvre à la fois l'imprécision des données et leurs inexactitudes. De nombreux facteurs concourent à produire des erreurs.

> *« Erreurs et incertitudes sont habituellement perçues comme mauvaises. Ce n'est pas nécessairement le cas, cependant, car il peut être utile de savoir comment apparaissent les erreurs et incertitudes, comment elles peuvent être gérées et si possible réduites… Un bonne compréhension des erreurs et de leur propagation conduit à un contrôle qualité actif* » (Burrough et McDonnel 1998).

L'erreur est généralement considérée comme soir aléatoire, soit systématique. La notion d'*erreur aléatoire* se réfère plutôt à un écart par rapport à l'état vrai dont la valeur est aléatoire.. L'*erreur systématique*, ou biais, provient d'un déplacement uniforme des valeurs et est parfois décrit comme ayant une 'exactitude relative' dans le monde de la cartographie (Chrisman 1991). S'il s'agit de déterminer l''aptitude à l'usage', l'erreur systématique ,peut être acceptable pour certaines applications, et pas pour d'autres. On peut prendre l'exemple de l'utilisation d'un système géodésique différent[1], qui peut ne pas poser de difficulté majeure s'il est utilisé sciemment, d'un bout à l'autre de l'analyse. Des difficultés apparaissent toutefois lorsqu'une analyse utilise des données de plusieurs sources différentes ayant des biais différents – par exemple des données fondées sur des systèmes géodésiques différents, ou bien où l'identification a été effectuée selon des systèmes nomenclaturaux différents.

« Comme on ne peut pas échapper à l'erreur, celle-ci devrait être reconnue comme une dimension fondamentale des données » (Chrisman 1991). Ce n'est que lorsque l'erreur est incluse dans la représentation des données qu'il est possible de répondre à des questions sur leurs limitations et même celles de la connaissance actuelle. Les erreurs connues dans les trois dimensions de l'espace, des attributs et du temps doivent être mesurées, calculées, enregistrées et documentées.

Validation et nettoyage

La validation est un processus utilisé pour déterminer si les données sont exactes, incomplètes, ou aberrantes. Le processus peut comprendre des vérifications de format, de complétude, de normalité, ou de limites, le revue des données pour identifier les aberrations (sur les plans géographique, statistique, temporel ou environnemental) ou d'autres erreurs, et l'évaluation des données par des experts du domaine (comme des taxonomiste spécialisés). Ces processus présentent habituellement leurs résultats sous forme de marqueurs et de documentation qui permettent ensuite de vérifier les enregistrements suspects. La validation peut aussi consister à vérifier le respect des normes, règles et conventions applicables. Une étape clé de la validation et du nettoyage des données est l'identification des causes profondes

[1] L'utilisation dans une application d'un système géodésique différent de celui qui a été utilisé pour créer les données peut donner lieu à un glissement systématique par rapport à la position réelle (d'un point de longitude et latitude données) qui peut atteindre environ 400 mètres, dans certaines parties du la Terre.

des erreurs détectées et les mesures prises pour empêcher ces erreurs de se reproduire (Redman 2001).

Le nettoyage des données fait référence au processus de correction des erreurs qui ont été identifiées dans les données durant la phase de validation. Le terme « nettoyage » (en anglais : « cleaning ») est synonyme de « lavage » des données (en anglais : « cleansing »), quoique certain utilise « lavage » pour recouvrir à la fois la validation et le nettoyage des données. Il est important dans le processus de nettoyage que des données ne soient pas perdues par inadvertance, et que les changements opérés sur l'information existante soit effectués avec soin. Il vaut souvent mieux conserver côte à côte dans la base de données à la fois les données originales et les nouvelles données issues de la correction, se sorte que si des erreurs sont faites au cours du nettoyage, l'information originale puisse être récupérée.

Nombre d'outils et de manuels ont été produits ces dernières années pour aider à valider et nettoyer les données sur les espèces. Ceux-ci seront traités dans les documents associés sur les *Principes et méthodes de nettoyage des données*. Le processus de nettoyage manuel des données est laborieux et demande du temps, et il est lui-même sujet à erreurs (Maletic et Marcus 2000).

Le déroulement général du nettoyage des données (d'après Maletic et Marcus 2000) est le suivant :

- Définir et déterminer les types d'erreurs,
- Rechercher et identifier les occurrences d'erreurs,
- Corriger les erreurs,
- Documenter les occurrences d'erreurs et les types d'erreurs,
- Modifier les procédures de saisie afin de réduire le nombre d'erreurs à l'avenir.

Exactitude de l'étiquetage

L'exactitude de l'étiquetage est habituellement entendue comme consistant à documenter la qualité des biens et produits vendus ou mis à la disposition de tiers. En ce qui concerne les données d'occurrence d'espèces, ceci comprend habituellement les méta données, si celles-ci documentent de manière complète les aspects de la qualité, des procédures et méthodes du contrôle qualité, et/ou des statistiques adaptées aux données mesurant leur qualité. L'exactitude de l'étiquetage est primordiale pour obtenir la certification et l'accréditation lorsque celles-ci sont utiles. La plupart des muséums et des herbiers assurent déjà cette exactitude quant à l'identité de l'expert et à la date à laquelle l'identification a été réalisée (information sur le determinavit), mais elle est rarement étendue aux autres informations présentes dans l'enregistrement ou assurée dans le cas des données d'observation et des données d'inventaire non liées à des spécimens.

Utilisateurs

Qui sont les utilisateurs ? Le terme d'utilisateur des données recouvre toute personne intervenant dans toute phase de la chaîne de l'information (figure 3). Dans le cas des données primaires sur les espèces, le terme recouvre les utilisateurs du muséum ou de l'herbier, tels que les taxonomistes, les gestionnaires, les chercheurs, les techniciens, les collecteurs, ainsi que les utilisateurs externes ou travaillant en aval tels que les décideurs, les scientifiques, les agriculteurs, les forestiers et les horticulteurs, les gestionnaires de l'environnement, les ONGs (environnementales ou impliquées dans la production), les professionnels de santé, les

pharmaciens, les professionnels de l'industrie, les conservateurs des jardins et des zoos, et le grand public (y compris les jardiniers). Les données d'occurrence d'espèces ont une liste infinie d'utilisateurs et concernent toute la communauté d'une manière ou d'une autre.

Les données primaires sur la biodiversité ont été souvent collectées sans prendre en considération l'ensemble des utilisateurs potentiels. Traditionnellement, les données, particulièrement celles des muséums et des herbiers, ont été collectées avec pour objectif principal de fournir de l'information à la recherche en taxonomie ou en biogéographie. C'était là un objectif essentiel, mais aujourd'hui les bailleurs de fonds de ces institutions, qui sont souvent des Ministères ou autres agences gouvernementales, souhaitent que leur investissement bénéficie à une communauté plus large, et donc que les données soient mieux valorisées en étant aussi utilisables à d'autres fins. En particulier, les gouvernements cherchent à utiliser les données pour améliorer la prise décision en environnement, les gestion environnementale et la planification de la conservation (Chapman et Busby 1994), et les conservateurs de ces données ne peuvent ignorer ces utilisateurs ni leurs besoins. Avec de bons mécanismes de retour d'information, les utilisateurs peuvent fournir un retour sur la qualité des donnée, et ainsi peuvent constituer un maillon important dans la chaîne de la qualité des données comme discuté plus loin.

Déterminer les besoins des utilisateurs est une tâche ardue. Mais cet effort est indispensable et il est source de grandes satisfactions.

Principes de la qualité des données

> *L'expérience a montré que le fait de traiter les données comme un atout à long terme et de les gérer dans un cadre coordonné apporte des économies considérables et une valeur durable* (NLWRA 2003)

Les principes de la qualité des données doivent être appliqués à toutes les étapes du processus de gestion des données (saisie, numérisation, stockage, analyse, présentation et utilisation). Il y a deux éléments clés pour améliorer la qualité des données : la prévention et la correction. La prévention des erreurs est étroitement liée à la fois à la collecte des données et à la saisie des données dans la base. Bien que des efforts considérables puissent et doivent être consacrés à la prévention des erreurs, il n'en reste pas moins que dans de grands jeux de données des erreurs continueront à se produire (Maletic et Marcus 2000) et la validation et la correction des données ne peuvent être ignorées.

La prévention des erreurs est considérée comme bien plus efficace que la détection des erreurs après coup, du fait que cette détection est souvent coûteuse et qu'on ne peut jamais garantir que toutes les erreurs sont détectées (Dalcin 2004). La détection d'erreurs joue cependant un rôle particulièrement important dans le cas des collections patrimoniales (Chapman et Busby 1994, English 1999, Dalcin 2004) qui fournissent un grand nombre des données primaires sur les espèces et des données d'occurrence d'espèces dont il est question ici.

Il faut commencer par établir une vision des données, développer une politique de gestion et d'utilisation des données, et implémenter une stratégie – et non se lancer de but en blanc dans des activités de nettoyage des données sans objectifs précis, sans coordination et d'une manière non systématique..

La vision

Il est important que les organismes développent une vision de la qualité de leurs données. Ceci concerne particulièrement les organismes qui comptent mettre leurs données à la disposition de tiers. Une bonne vision de la qualité de leurs données améliorera généralement la vision d'ensemble de l'organisme (Redman 2001) ainsi que les procédures opérationnelles de l'organisation. En développant cette vision, les gestionnaires devraient se focaliser sur l'obtention d'un cadre intégré de gestion dans lequel les responsables, le personnel, le matériel informatique, les applications logicielles, le contrôle qualité et les données elles-mêmes sont intégrés à l'aide d'outils, règles de conduite et normes appropriés, afin d'assurer la maintenance des données et d'en faire des produits d'information de qualité (NLWRA 2003).

Une vision de la qualité des données

- oblige un organisme à réfléchir à ses besoins à long terme en matière de données et d'information et à leur relation avec sa performance à long terme,
- oriente l'activité dans la bonne direction : c'est-à-dire vers la qualité,
- fournit une base solide pour la prise de décision à la fois à l'intérieur et à l'extérieur de l'organisme,
- formalise la reconnaissance des données et de l'information comme atouts essentiels de l'organisme,
- maximise l'utilisation des données et de l'information de l'organisme, évite la duplication, facilite les partenariats et améliore l'équité de l'accès, et
- maximise l'intégration et l'interopérabilité.

La politique

Tout autant qu'une vision, un organisme a besoin d'une politique pour implémenter cette vision. Le développement d'une politique saine de qualité des données permettra vraisemblablement de

- forcer l'organisme à réfléchir de manière plus ouverte à la qualité et à réexaminer ses pratiques quotidiennes,
- formaliser les processus de gestion des données,
- aider l'organisme à clarifier ses objectifs concernant
    - la réduction des coûts,
    - l'amélioration de la qualité des données,
    - l'amélioration du service et des relations avec ses clients,
    - l'amélioration du processus de décision,
- apporter aux utilisateurs confiance et stabilité lorsqu'ils consultent et utilisent les données de l'organisme,
- améliorer les relations et la communication avec les clients de l'organisme (aussi bien les fournisseurs que les utilisateurs des données),
- améliorer la réputation de l'organisme au sein de sa communauté, et

- améliorer les chances d'un meilleur financement lorsque les objectifs de bonne pratique sont atteints ou approchés.

La stratégie

Comme les grandes institutions détiennent des quantités importantes de données, il est crucial que ces institutions mettent en place une stratégie pour saisir et vérifier leurs données (voir aussi le paragraphe *Priorétiser*, plus loin). Une bonne stratégie de qualité des données (à la fois pour la saisie et le contrôle qualité) doit établir des objectifs à court, moyen et long termes. Par exemple (d'après Chapman et Busby 1994) :

- **Court terme**. Les données qui peuvent être assemblées et vérifiées sur une période de 6 à 12 mois (ceci concerne généralement à la fois les données déjà présentes dans la base et les nouvelles données qui nécessitent une vérification moins poussée).
- **Moyen terme**. Les données qui peuvent être saisies dans la base sur une période d'environ 18 mois avec des moyens réduits et les données dont la qualité peut être vérifiée à l'aide de méthodes internes simples.
- **Long terme**. Les données qui peuvent être entrées et/ou vérifiées sur une longue durée en mettant en œuvre des dispositifs collaboratifs, des méthodes de vérification plus sophistiquées, etc. Ceci peut impliquer un travail systématique sur toute la collection, en sélectionnant :
    - Les groupes taxonomiques révisés récemment ou en cours d'étude taxonomique au sein de l'institution.
    - Les collections importantes (les types, les spécimens de référence particuliers, etc.)
    - Les groupes clés (les familles importantes, les taxons significatifs sur la plan national, les taxons menacés, les taxons importants sur le plan écologique ou environnemental).
    - Les taxons issu de régions géographiques clés (par exemple provenant de pays en développement avec lesquels l'institution souhaite partager ses données, ou de régions géographiques qui intéressent particulièrement l'institution).
    - Les taxons impliqués dans des accords de coopération avec d'autres institutions (par exemple lorsqu'un ensemble d'institutions ont convenu de coopérer pour numériser les données sur les mêmes taxons).
    - Un parcours systématique d'un bout à l'autre d'une collection.
    - Les acquisitions récentes, de préférence aux collections anciennes.

Certains des principes d'une bonne gestion des données qui devraient faire partie de la stratégie sont (d'après NLWRA 2003) :

- Ne pas réinventer la roue dans la gestion de l'information,
- Rechercher l'efficacité dans la collecte des données et les procédures de contrôle qualité,
- Partager les données, l'information et les outils chaque fois que cela est possible,
- Utiliser les normes existantes ou développer de nouvelles normes avec un souci de robustesse et en synergie avec les autres acteurs,
- Renforcer le développement des réseaux et les partenariats,
- Présenter un modèle de travail sain pour la collecte et la gestion des données,

- Réduire la duplication dans la collecte des données et le contrôle qualité des données,
- Regarder au delà des utilisations immédiates et examiner les demandes des utilisateurs,
- S'assurer de produire une bonne documentation et de bonnes procédures de métadonnées.

Prévenir vaut mieux que corriger

Le coût de saisie d'une collection dans une base de données peut être important (Armstrong 1992) mais ce n'est qu'une fraction du coût de la vérification et de la correction ultérieures des données. Il vaut mieux prévenir les erreurs que de devoir ensuite les corriger (Redman 2001) et c'est de loin l'option la moins coûteuse. Effectuer des corrections a posteriori peut aussi impliquer que des données incorrectes auront déjà été utilisées dans un certain nombre d'analyses avant d'être corrigées, induisant en aval le coût de décisions prises en fonction de données de mauvaise qualité, ou celui de recommencer les analyses.

La prévention des erreurs, toutefois, n'impacte pas les erreurs déjà introduite dans la base de données, c'est pourquoi la validation et le nettoyage des données reste une partie importante du processus de qualité des données. Le nettoyage est important pour identifier la cause des erreurs qui ont déjà été introduites, et il devrait conduire à établir des procédures qui assurent que ces erreurs ne soient pas répétées. Cependant le nettoyage ne doit pas intervenir de manière isolée ; sinon les problèmes ne disparaîtront pas. Les deux opérations, nettoyage des données et prévention des erreurs, doivent être conduites conjointement. Décider de nettoyer les données et de s'occuper plus tard de prévenir les erreurs, signifie généralement que la prévention des erreurs n'est jamais réalisée de manière satisfaisante, et en attendant les erreurs se multiplient dans la base..

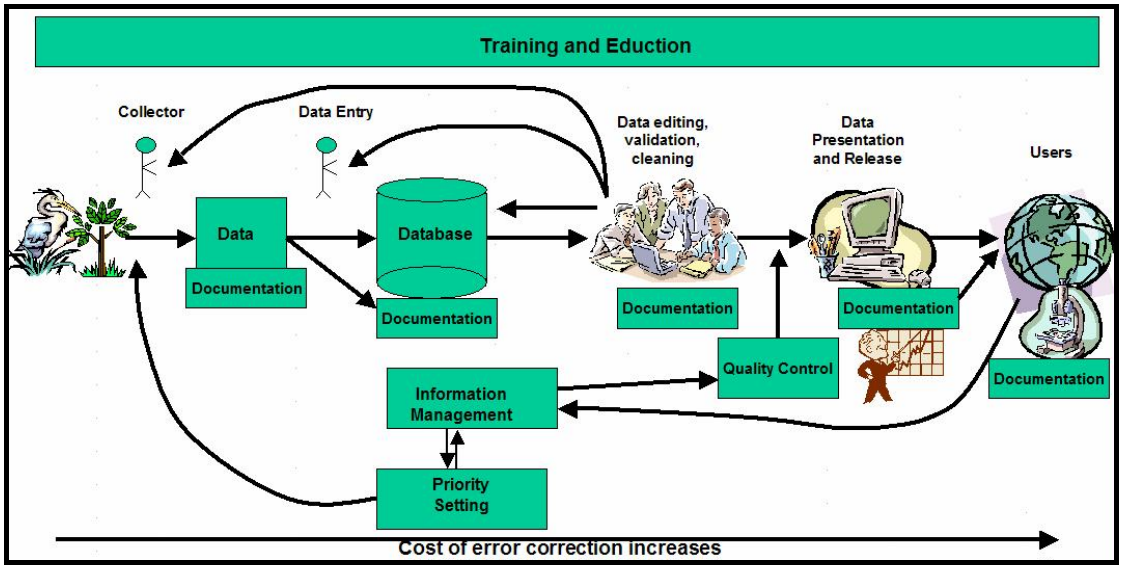


Fig. 3. *Chaîne de gestion de l'information montrant que le coût de la correction des erreurs augmente à mesure que l'on progresse dans la chaîne. Une bonne documentation et une bonne formation sont essentielles à toutes les étapes.*

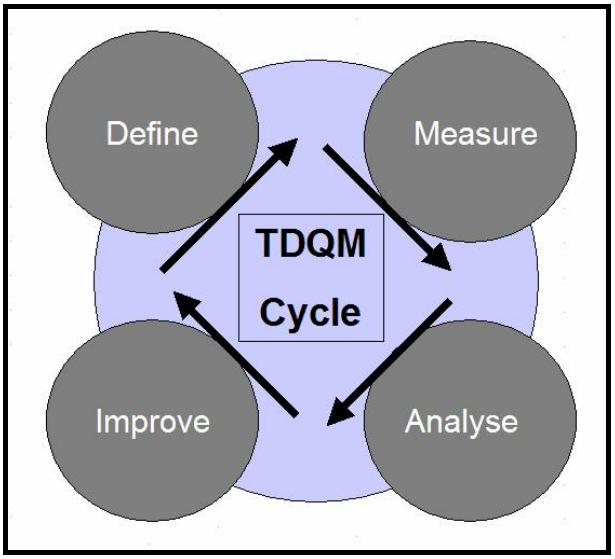


Fig. 4. *Cycle de gestion globale de la qualité des données*[2] *montrant la nature cyclique du processus de gestion des données (d'après Wang 1998).*

Les conservateurs et les propriétaires des données (organismes gérant des collections, comme les muséums et les herbiers) sont largement responsables de la qualité de leurs données. Il n'en reste pas moins que ceux qui fournissent les données et ceux qui les utilisent ont aussi leur part de responsabilité.

> *Assigner la responsabilité de la qualité des données à ceux qui les produisent. Si cela n'est pas possible, en assigner la responsabilité aussi près que possible de leur production*
>
> (Redman 2001)

Le collecteur a la responsabilité première

La première responsabilité de la gestion de la qualité des données est celle du collecteur. Il est de sa responsabilité de s'assurer que :

- l'information portée sur l'étiquette est correcte,
- l'information de l'étiquette est enregistrée et documentée avec exactitude,
- l'information relative à la localisation est aussi exacte que possible, et que son degré d'exactitude comme sa précision sont documentées,
- les méthodologies de collecte sont pleinement documentées,
- l'étiquette ou les notes de terrain sont claires et sans ambiguïté, et
- l'information inscrite sur l'étiquette est lisible par les opérateurs de saisie.

Si l'information portée sur l'étiquette ou inscrite dans le carnet de collecte n'est pas claire et exacte, il est ensuite extrêmement difficile de la corriger après coup. Ceci est moins important pour la partie taxonomique des données lorsque des spécimens sont déposés, puisqu'elle peut –et elle l'est habituellement- être ultérieurement vérifiée par des experts.

[2] « TDQM » = « Total Data Quality Management » : Gestion globale de la qualité des données, avec les étapes Définir, Mesurer, Analyser et Améliorer.

Il est également important que la localisation et les informations subsidiaires soient notées lors de la collecte ou de l'observation, et non en fin de journée ou au retour au laboratoire comme cela a souvent été le cas par le passé.

La plupart des données d'un organisme proviennent de « fournisseurs », et il est beaucoup plus facile de développer de bonnes pratiques de collecte de données que de corriger les erreurs en aval.

Le gestionnaire ou conservateur a la responsabilité essentielle et à long terme

Le conservateur des données (muséum, herbier, université, organisme de conservation, ONG, ou particulier) a la responsabilité à long terme de maintenir et améliorer la qualité des données aussi longtemps qu'il en conserve la responsabilité (voir, par exemple, une liste des responsabilités des conservateurs dans Olivieri *et al.* 1995, p. 623). Il est important que l'organisme conservateur assigne en interne la responsabilité prédominante de gérer la qualité des données, mais il est aussi essentiel que l'organisme ait une culture de la qualité des données de sorte que chaque individu au sein de l'organisme sache qu'il a une part de responsabilité dans la qualité des données dont l'organisme à la charge. Il est de la responsabilité du conservateur d'assurer que :

- les données sont transcrites dans la base correctement et avec exactitude à partir des notes du collecteur,
- les procédures de contrôle qualité sont en place et mises en œuvre lors de la saisie,
- les données et la qualité des données sont documentées de manière appropriée et exacte,
- des vérifications de la validité des données sont effectuées régulièrement,
- les vérifications effectuées sont bien documentées,
- les données sont stockées et archivées de manière appropriée (voit les notes sur le stockage, plus loin),
- les versions antérieures sont systématiquement conservées pour permettre des comparaisons et le retour aux données originales,
- l'intégrité des données est maintenue,
- les données sont disponibles en temps utile et de manière correcte avec la documentation qui permet aux utilisateurs de déterminer leur « aptitude à l'usage »,
- les responsabilités des conservateurs concernant la confidentialité, les droits de propriété intellectuelle, le copyright, et les sensibilités des propriétaires traditionnels ou autochtones sont assumées,
- les conditions d'utilisation des données sont tenues à jour et disponibles en même temps que toute restriction sur l'usage ou les domaines connus où les données ne sont pas utilisables,
- toutes les requêtes légales au regard des données sont satisfaites,
- le retour d'information de la part des utilisateurs sur la qualité des données est traité en temps utile,
- la qualité des données est maintenue au plus haut niveau possible à tout moment,
- toutes les erreurs connues sont soigneusement documentées et indiquées aux utilisateurs.

La propriété et la charge de conservation des données ne confère pas seulement des droits à gérer les données et contrôler leur accès ; il confère aussi des responsabilité quant à leur gestion, au contrôle de leur qualité et à leur maintenance. Les conservateurs ont aussi une responsabilité morale quant à la transmission des données aux générations futures.

Responsabilité des utilisateurs

Les utilisateurs des données ont aussi leur part de responsabilité dans la qualité des données. Ils doivent informer en retour les conservateurs de toute erreur ou omission qu'ils peuvent trouver, y compris des erreurs dans la documentation des données, ou de toute information supplémentaire qu'ils souhaiteraient voir enregistrer à l'avenir. Ce sont souvent les utilisateurs, en regardant des données dans le contexte d'autres données, qui peut identifier des erreurs ou des valeurs aberrantes dans les données qui autrement passeraient inaperçues. Un muséum peut ne posséder qu'un sous ensemble de toutes des données disponibles (provenant d'un seul pays ou d'une seule région par exemple), et ce n'est que quand les données sont combinées avec celles d'autres sources que les erreurs deviennent évidentes.

Selon l'objectif de la collecte des données dans une institution, l'utilisateur peut aussi apporter une contribution utile à l'établissement des priorités futures concernant la collecte et la validation des données (Olivieri *et al.* 1995).

L'utilisateur a aussi une responsabilité dans la détermination de l'aptitude à l'usage des données et dans le fait de ne pas utiliser les données de manière inappropriée.

Utilisateurs et collecteurs ont un rôle important à jouer dans l'aide à apporter aux conservateurs pour maximiser la qualité des données de leurs collections, et les uns comme les autres ont particulièrement intérêt à ce que les données soient de la meilleure qualité possible.

L'établissement de partenariats

L'établissement de partenariats pour la maintenance de la qualité des données peut être gratifiant et limiter les coûts. C'est particulièrement vrai dans le cas des muséums et des herbiers, où des enregistrements sont souvent dupliqués dans un certain nombre de muséums. Dans le domaine des bibliothèques, de nombreuses communautés utilisent une forme de collaboration et de partenariat pour améliorer le catalogage de leurs ouvrages (Bibliothèque du Congrès 2004), et les muséums et herbiers pourraient facilement en faire de même. De tels partenariats et accords de collaboration pourraient être développés avec :

- des collecteurs de données importants (afin d'améliorer le flux d'information – par exemple en développant des formulaires normalisés de collecte de données et de rapports, fournir des GPS, etc.),

- d'autres institutions hébergeant des données similaires (par exemple des doubles des spécimens que l'on possède),
- d'autres institutions du même type ayant des besoins similaires en qualité des données et qui peuvent être en train de développer des méthodes, outils, normes et procédures de contrôle qualité,
- des intermédiaires clés dans la diffusion des données (comme le GBIF) qui jouent un rôle important dans le regroupement et la diffusion d'informations issues d'un grand nombre de fournisseurs de données,
- les utilisateurs des données (en particulier ceux susceptibles d'effectuer des tests de validation des données durant ou avant l'analyse), et
- des statisticiens et des auditeurs des données qui peuvent être capables d'améliorer les méthodologies de gestion des données et des flux de données et les techniques relatives à la qualité des données.

Votre organisation n'est pas la seule à se préoccuper de la qualité des données

Priorétisation

Si l'on veut obtenir assez vite des données hautement utiles à un grand nombre d'utilisateurs, il peut être nécessaire d'établir des priorités quant à la saisie et/ou la validation des données (voir aussi les commentaires au paragraphe *Complétude*, plus loin). Dans ce but il peut être nécessaire de :

- mettre l'accent d'bord sur les données les plus critiques,
- se concentrer sur des groupes circonscrits (dans les domaines taxonomique ou géographique, etc.),
- traiter en priorité les spécimens types et de référence,
- ignorer les données qui ne sont pas utilisées, ou dont la qualité ne peut pas être garantie (c'est-à-dire des enregistrement avec une mauvaise information de géo-référencement - mais avoir à l'esprit l'importance de certaines données historitques malgré leur mauvais géo-référencement),
- s'attacher aux données qui sont utiles au plus grand nombre d'utilisateurs et d'utilisations,
- porter l'effort sur les domaines où de nombreuses données peuvent être nettoyées pour un coût réduit (par exemple en utilisant des traitements de masse).

Les données ne sont pas toutes d'égale importance, il faut donc se concentrer sur les plus importantes, et si un nettoyage des données s'impose, faire en sorte qu'il n'y ait ***jamais*** *à y revenir.*

Complétude

Les organismes devraient s'efforcer d'avoir des données complètes (au moins pour un groupe prioritaire – pour un groupe taxonomique, une région, etc.) afin que tous les enregistrements éligibles soient exploités dans la compilation des données. Il vaut mieux avoir des données

complètes pour un groupe, qu'avoir une quantité de données incomplètes, car les analyses effectuées sur des données incomplètes ne pourront pas être poussées. Il est aussi important d'avoir une politique relative aux données manquantes, qui définisse des seuils de données manquantes et les réponses appropriées, autant qu'une politique pour documenter l complétude des données (voir le paragraphe *Documentation*, plus loin).

Durée de validité et délai de disponibilité

Il y a trois facteurs clés liés au délai de disponibilité et à la durée de validité des données :

- la période durant laquelle les données ont été collectées,
- la date de dernière mise à jour des données,
- la durée vraisemblable de validité des données.

La durée de validité des données préoccupe beaucoup les utilisateurs. Beaucoup de conservateurs de données la relient à la période de la collecte ou de l'inventaire. Du fait du délai qui s'écoule entre la collecte et la publication (ce qui pour les données biologiques peut être excessivement long), l'information publiée représente « ce qui a été » et non « ce qui est ». La plupart des utilisateurs de données sur la biodiversité en sont conscients, et cet aspect est l'un des paramètres qui distinguent cette catégorie de données des autres.

En termes de gestion de la qualité des données, on se réfère souvent à la période durant laquelle les données peuvent être utilisées, qui peut être liée à la date de dernière vérification ou mise à jour. Ceci concerne particulièrement les noms attachés aux données. Quand ont-ils été mis à jour pour la dernière fois, et sont-ils conformes à la taxonomie récente ? Est-ce que les règles modernes de la nomenclature taxonomique ont été respectées ? Par exemple, lorsqu'une espèce est divisée en plusieurs taxons de périmètre réduit, l'un de ces taxons retient le nom de l'ancienne espèce : il peut être important pour l'utilisateur de savoir si le nom utilisé se réfère à l'ancien taxon, plus large, ou au nouveau, plus restreint. La date de validité peut être utilisée comme la date « consommer avant … » des produits alimentaires : au delà de cette date, le conservateur ne garantit plus l'information nomenclaturale attachée à l'enregistrement.

Il se peut aussi que pour certains jeux de données, la notion de délai de disponibilité et de date de validité ne soient pas applicables ou que cette indication ne puisse être incorporée ou tenue à jour. Ce peut être notamment le cas des grandes collections de muséums ou d'herbiers. Mais cette notion peut être importante dans le cas des observations ou données d'inventaire qui ne se réfèrent pas à un spécimen, ou sur lesquelles les dernières révisions taxonomiques n'ont pas été répercutées. Elle est aussi importante pour les collections recombinées à partir de contributions diverses. Ce peut être le cas par exemple si plusieurs institutions dans des pays en développement confient une copie de leurs données à une institution tierce qui en assure l'hébergement aux fins de connexion au portail GBIF.

Fréquence de mise à jour

La fréquence de mise à jour des données au sein d'un jeu de données est liée à la durée de validité et au délai de disponibilité : elle doit être formalisée et documentée. Ceci concerne autant l'ajout de nouvelles données que la correction de données existantes : les deux ont un impact sur la qualité globale du jeu de données et sont donc aussi importants pour les utilisateurs. Un utilisateur ne voudra pas prendre la peine de télécharger ou se procurer un jeu de données s'il est sur le point d'être mis à jour et amélioré.

Cohérence

Redman (1996) a distingué deux aspects de la cohérence : la cohérence sémantique, qui requiert que la présentation des données soit claire, non ambiguë et cohérente ; et la cohérence structurelle, selon laquelle les types et attributs des entités devraient avoir les mêmes structure et format de base.

Un exemple simple de cohérence sémantique concerne le fait que les données soient toujours dans les mêmes champs, et soient de ce fait aisées à retrouver – par exemple il doit y avoir des champs distincts pour le rang infra spécifique et le nom infra spécifique, de sorte qu'il soit toujours clair que le champ du nom infra spécifique ne contient qu'un nom, et que l'indication du rang (« variété. » ou « sous-espèce ») est placée dans un autre champ (voir Table 2).

Tableau 1. *Incohérence sémantique dans le champ « Niveau infra spécifique »*

Genre	Espèce	Niveau infra spécifique
Eucalyptus	globulus	sous espèce bicostata
Eucalyptus	globulus	bicostata

Tableau 2. *Cohérence sémantique rétablie par l'ajout d'un champ supplémentaire permettant de distinguer rang et nom*

Genre	Espèce	Rang infra spécifique	Nom infra spécifique
Eucalyptus	globulus	sous espèce.	bicostata
Eucalyptus	globulus		bicostata

Un bon modèle de base de données relationnelle ne permettrait pas de nombreuses incohérences de type ; mais de nombreuses bases de données existant actuellement dans les institutions en charge des collections ne sont pas aussi conçues de manière aussi rigoureuse.

La cohérence structurelle concerne le fait qu'au sein d'un même champ –par exemple le champ » Rang infra spécifique » de la table 2-, le rang « sous espèce » soit toujours noté de la même manière –et non selon le cas « sous esp. », « ssp. », « ss esp. », etc. Ceci peut être évité grâce à un bon modèle de données et une bonne gestion des valeurs inscrites dans les champs.

La cohérence est aussi importante dans la méthode que dans la documentation car elle permet aux utilisateurs de savoir quels tests ont été effectués et comment ils l'ont été ; où trouver l'information ; et comment interpréter les éléments d'information importants. Il faut toutefois trouver un équilibre entre souci de cohérence et flexibilité (Redman 2001).

Flexibilité

Les conservateurs de données doivent ménager une certaine flexibilité dans leurs méthodes de contrôle qualité. En effet, bien que de nombreuses données biologiques soient de nature similaire, des approches différentes en matière de qualité peuvent être mieux adaptées. Selon les régions (par exemple, si des jeux de données différents sont disponibles selon les régions pour effectuer des tests comparatifs), ou les groupes taxonomiques (organismes aquatiques ou organismes marins, etc.), ou les méthodes de saisie des données (enregistrements issus d'observations ou d'inventaires, par opposition à ceux-ci issus de spécimens de collections).

Les opinions taxonomiques sont en fait des hypothèses, et des opinions (ou hypothèses) taxonomiques différentes (mais valides) peuvent conduire le même organisme à être classé différemment par des taxonomistes différents et ainsi avoir un ou plusieurs noms alternatifs – chacun pouvant être aussi valide (Pullan *et al*. 2000, Knapp *et al*. 2004). C'est le cas lorsque deux taxonomistes sont en désaccord sur la position de taxons au sein de genres – par exemple, certains taxonomistes placent des espèces dans le genre *Eucalyptus*, alors que d'autres croient qu'elles appartiennent au genre *Corynthia*. En pratique, et particulièrement en zoologie, la vision du réviseur le plus récent est acceptée à moins qu'il n'y ait une bonne raison de la rejeter.

La flexibilité ménage la possibilité d'adapter la présentation à la demande. Des travaux récents du Groupe de travail sur les bases de données taxonomiques (en anglais, « Taxonomic Databases Working Group » : TDWG[3]) et d'autres ont été consacrés aux structures de base de données qui permettent de présenter des points de vue taxonomiques différents (Berendsohn 1997). Bien qu'en surface une telle flexibilité puisse sembler réduire la qualité, elle permet en réalité à l'utilisateur une plus grande souplesse pour déterminer l'aptitude à l'usage, et de ce fait elle améliore en pratique la qualité vu de l'utilisateur.

Transparence

La transparence est un élément important qui améliore la confiance des utilisateurs. La transparence signifie que les erreurs ne sont pas cachées, mais au contraire identifiées et explicitement documentées, que les procédures de validation et contrôle qualité sont documentées et accessibles, et que les mécanismes de retour d'information sont ouverts et encouragés.

Un exemple où la transparence est importante est celui de la documentation des méthodologies de collecte (particulièrement dans le cas des observations et inventaires). Elle est en effet précieuse pour que l'utilisateur puisse déterminer si les données sont appropriées pour une usage donné.

Mesures et objectifs de performance

Les mesures de performance sont un ajout intéressant aux procédures de contrôle de qualité : elles permettent aux utilisateurs d'avoir confiance dans le niveau d'exactitude ou de qualité des données. Les mesures de performance peuvent inclure des tests statistiques :

- sur les données : par exemple, 95 % des enregistrements sont dans un rayon de 1.000 mètres de leur position indiquée ;
- sur le niveau de contrôle qualité : par exemple, 65 % des enregistrements ont été vérifiés par un taxonomiste qualifié dans les 5 dernières années, et 90 % ont été vérifiés par un taxonomiste qualifié dans les 10 dernières années ;
- sur la complétude : toutes les mailles de 10 minutes d'arc ont été échantillonnées ;
- etc.

Les mesures de performance aident à quantifier la qualité des données. Les avantages en sont que :

- l'organisme peut s'assurer lui-même que certaines données sont de haute qualité,

[3] Voir la site www.tdwg.org

- elles contribuent à la gestion d'ensemble des données et à la réduction de la redondance, et
- elles aident à coordonner les divers aspects de la chaîne de la qualité des données, de sorte l'on puisse s'organiser pour partager le travail entre plusieurs opérateurs.

Avant de mesurer la qualité des données, il faut envisager la manière dont les utilisateurs vont exploiter les résultats, afin de les structurer pour un usage optimal.

Nettoyage des données

Les principes du nettoyage des données sont traités dans le document associé *Principes et méthodes de nettoyage des données*. Il suffit ici d'indiquer le cadre général du nettoyage des données, adapté de Maletic et Marcus (2000) :

- définir et déterminer les types d'erreur,
- rechercher et identifier les erreurs,
- corriger les erreurs,
- documenter les erreurs et les types d'erreur,
- modifier les procédures de saisie pour limiter de telles erreurs à l'avenir.

Ne vous laissez pas leurrer par l'apparente simplicité des outils de nettoyage des données. Ils sont précieux et utiles à court terme, mais sur le long terme, rien ne remplace la prévention des erreurs.

Valeurs aberrantes

La détection des valeurs aberrantes (sur les plans géographique, statistique et environnemental) peut nous fournir l'un des tests les plus utiles pour identifier les erreurs potentielles dans les données spatiales. Il est cependant important que ces tests ne détruisent pas sans examen critique des données pour la seule raison qu'elles contiendraient des valeurs statistiquement aberrantes. Les données environnementales sont connues pour contenir des enregistrements apparemment aberrants statistiquement, bien qu'ils soient parfaitement valables. Ils peuvent être dus à des évolution historiques, à des changements de régime climatique, aux conséquences d'activités humaines, etc. L'exclusion sans examen critique de valeurs aberrantes peut ôter des enregistrements utiles du jeu de données et fausser les analyses futures.

D'un autre côté, les utilisateurs peuvent décider de détruire des données aberrantes de leur jeu d'analyse s'ils ne sont pas certains de leurs validité. L'identification des valeurs aberrantes permet donc aux conservateurs de données d'identifier des erreurs potentielles, mais elle aide aussi les utilisateurs à déterminer si certains enregistrements sont ou non aptes à être intégrés dans leurs analyses.

La détection des valeurs aberrantes peut être une méthode de validation précieuse, mais il faut bien noter que toutes les valeurs aberrantes ne correspondent pas à des erreurs.

Etablir des objectifs d'amélioration

L'établissement d'objectifs simples, faciles à quantifier, peut conduire à une amélioration rapide de la qualité des données. Un objectif comme celui de diviser par deux le pourcentage d'enregistrements nouveaux mal géocodés tous les six mois pendant deux ans peut conduire à une réduction totale du taux d'erreurs de 94 % (Redman 2001). De tels objectifs devraient se concentrer sur :

- ces calendriers clairs et exigeants,
- des taux d'amélioration, plutôt que des degrés absolus de qualité,
- des définitions claires (plus précises par exemple que « mal géocodés »),
- des objectifs simples et accessibles.

On peut aussi se donner des objectifs à plus long terme comme de diviser par deux la durée du nettoyage des données, en améliorant les techniques de saisie et de validation.

Les objectifs de performance sont un bon moyen pour un organisme de maintenir un niveau cohérent de vérification et validation de qualité – par exemple, que 95 % des enregistrements soient documentés et validés dans les six mois suivant leur réception.

Traçabilité

Les conservateurs doivent garder la trace des opérations de vérification : quelles données ont été vérifiées et quand. Ceci permet d'éviter les redondances et d'empêcher que des données ne passent à la trappe. Le meilleur moyen de réaliser cela est de tenir un journal des opérations de validation.

Contrôles d'édition

Les contrôles d'édition sont des règles de travail qui déterminent les valeurs permises dans un champ donné. Par exemple, la valeur dans la champ mois doit être comprise entre 1 et 12, la valeur du champ jour doit être entre 1 et 31 – la limite maximale dépendant du mois-, etc. On parle de règles « univariées » lorsqu'elles s'appliquent à un champ,et de règles « bivariées » lorsqu'elles s'appliquent à deux champs simultanément (par exemple, la combinaison du jour et du mois).

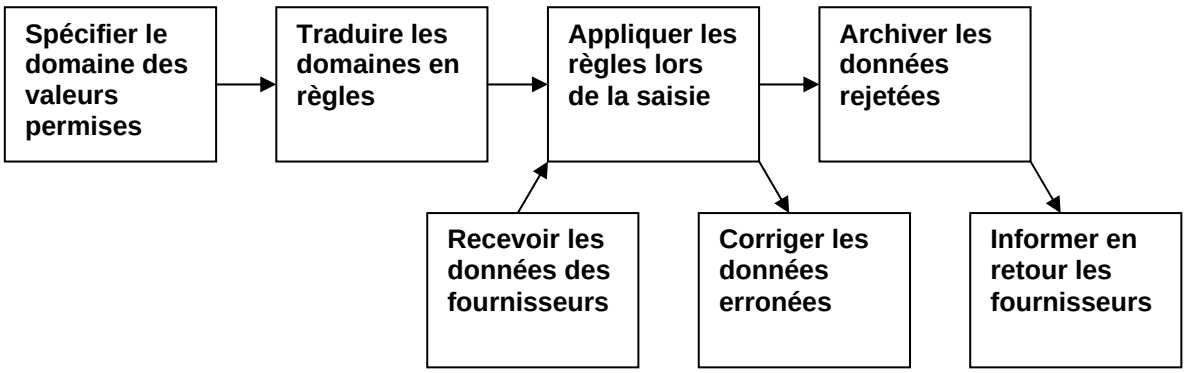


Fig. 5. *Utilisation des contrôles d'édition (adapté de Redman 2001)*

Un deuxième exemple concerne les coordonnées géographiques. Si elles sont exprimées en latitude et longitude, des tests simples de domaine de valeur permettent de vérifier que la latitude est comprise entre 0 et 90 degrés, que les minutes et secondes sont entre 0 et 60, etc. Lorsque l'on a affaire à des données en UTM, en revanche, c'est plus compliqué. Très souvent une base de données concernant une petite région à l'intérieur d'une zone UTM n'inclura pas l'identifiant de la zone. Ceci peut semble acceptable tant que les données ne sont pas combinées avec des données d'autres régions. Mais dès qu'on tente une telle combinaison, les données deviennent complètement inutilisable. Donc les contrôles d'édition doivent assurer que l'identifiant de la zone est toujours inclus.

Minimiser la duplication et le besoin de retoucher les données

L'expérience dans le monde de l'industrie et des services a montré que la mise en œuvre d'une chaîne de gestion de l'information peut limiter la duplication et le besoin de retoucher les données, et conduire à une réduction des taux d'erreur jusqu'à 50 %, et ainsi réduire le coût résultant de l'utilisation de mauvaises données parfois des deux tiers (Redman 2001). Ceci est largement dû à des gains d'efficacité obtenus en attribuant à chacun des responsabilités claires dans la gestion des données et le contrôle qualité, en minimisant les goulots d'étranglement et les temps d'attente, en minimisant la duplication liée aux contrôles qualité répétés par plusieurs personnes, et en améliorant l'identification de meilleures méthodes de travail.

Conservation des données originales

Il est important que les données originales telles que le collecteur les a enregistrées, ou même ultérieurement telles qu'elles ont été intégrées dans la collection ou saisies dans la base, ne soient pas perdues au cours du processus d'édition et de nettoyage des données. Les changements effectués dans la base de données au cours du processus de nettoyage devraient être inscrits comme une information supplémentaire, tout en maintenant aussi l'information originale. Une fois que l'information a été détruite, il est difficile, sinon impossible de la recouvrer. Ceci peut être particulièrement important en ce qui concerne l'information sur les collecteurs et les localisations. Ce qui apparaît à un moment donné à un gestionnaire de données comme une erreur peut ne pas en être une en réalité. Les changements que subissent les noms des localités (par exemple le passage de la Tchécoslovaquie à la République Tchèque) ne sont parfois pas seulement des changements de nom mais aussi de périmètre. Il

peut être important ensuite de savoir ce qui avait été inscrit à l'origine au lieu de n'avoir que la version « corrigée ». Voir aussi les commentaires au paragraphe Archivage.

La discrétisation peut entraîner la perte de données et de leur qualité

La discrétisation des données peut souvent entraîner la perte de données et ainsi réduire leur qualité d'ensemble. Par exemple, si l'on dispose de données dotées d'une information de localisation détaillée (peut-être même un géo référencement), mais que l'on stocke les données sur la base d'un maillage. Il vaut toujours mieux stocker les données avec leur résolution la plus fine, et ensuite les discrétiser à la sortie si c'est nécessaire pour un usage particulier. Si un utilisateur a besoin de produire une carte de présence/absence sur un maillage de 10 x 10 minutes d'arc, il peu facilement le faire à partir de données stockées comme des points, mais si les données sont stockées sur la base d'un maillage, il est impossible de les exploiter à une échelle plus fine. Il peut être aussi extrêmement difficile - voire impossible- de combiner des données qui ont été discrétisées selon un maillage différent (taille des mailles ou origine du repère). Il en va de même pour les données descriptives – si les données sont discrétisées selon des états adaptés à une clé (par exemple : arbre au dessus de 6 mètres et buisson en dessous), et que de nouvelles données sont obtenues d'une autre source qui utilise un seuil de 4 mètres au lieu de 6 mètres, que fait-on des valeurs entre 4 et 6 mètres ? Il vaut bien mieux stocker les données avec la taille en mètres, et se préoccuper de décider s'il s'agit d'un arbre ou d'un buisson plus tard.

Un cas typique est celui du stockage du degré d'exactitude du géocodage. J'ai toujours recommandé de stocker ce degré en mètres, mais de nombreuses bases de données stockent cette information en catégories (<10m, 10-100m, 100-1000m, 1000-10.000m). Si vous avez un enregistrement dont vous avez pu déterminer qu'il est exact à 2 km près, vous perdez immédiatement de l'information en devant placer cet enregistrement dans la catégorie 10-10.000m.

Documentation

Une bonne documentation est un élément clé de la gestion des données. Sans une bonne documentation, l'utilisateur ne peut pas déterminer l'aptitude des données à l'usage qu'il a en tête, et ne peut donc pas déterminre la qualit des données pour cet usage. Une discussion plus détaillée de la documentation se trouve au chapitre *Documentation,* plus loin.

Retour d'information

Il est essentiel pour les conservateurs d'encourager le retour d'information de la part des utilisateurs de leurs données, et de traiter avec sérieux le retour qu'ils reçoivent. Comme mentionné au paragraphe Responsabilité de l'utilisateur, ci-dessus, l'utilisateur a souvent de bien meilleures chances de détecter certains types d'erreur en combinant les données avec celles d'un ensemble d'autres sources, que chaque conservateur ne peut le faire en travaillant isolément.

Le développement de bons mécanismes de retour d'information n'est pas toujours tâche aisée. Un bouton de retour d'information peut être placé sur la page de l'interface de requête, ou bien un document attaché peut être envoyé aux utilisateurs au moment du téléchargement, qui établisse des méthodes pour effectuer le retour d'information sur les erreurs et adresser des

commentaires au conservateur. Certains de ces mécanismes sont traités en détail dans le document associé *Principes et méthodes de nettoyage des données*.

Etablir des canaux efficaces pour le retour d'information de la part des utilisateurs et des fournisseurs est un mécanisme aisé et productif pour améliorer la qualité des données.

Education et formation

L'éducation et la formation à tous les niveaux de la chaîne de l'information peuvent conduire à de grandes améliorations de la qualité des données (Huang et al. 1999). Cela commence par la formation et l'éducation des collecteurs à l'utilisation de bonnes procédures de collecte qui respectent les besoins des utilisateurs des données, cela passe par la formation des opérateurs de saisie et des personnels techniques en charge de la gestion au jour le jour des bases de données, jusqu'à l'éducation des utilisateurs finaux quant à la nature des données, leurs limitations et leurs usages potentiels. Les aspects éducation et formation de la qualité des données reposent largement sur une bonne documentation.

On trouve un exemple d'intégration de vérification de la qualité des données, d'éducation et de formation dans le projet de géo référencement MaPSTeDI (Université du Colorado 2003). Le processus comprend la vérification d'un certain nombre d'enregistrements de chaque opérateur de saisie du géocode. Dans le cas d'un nouvel opérateur, l'exactitude des premiers 200 enregistrements est vérifiée par un superviseur. Non seulement cela maintient la qualité des données, mais cela permet aussi à l'opérateur d'apprendre et de s'améliorer à partir de ses erreurs. Selon l'opérateur, un lot supplémentaire de 100 enregistrements peut être vérifié et à mesure que l'opérateur devient plus expérimenté, la vérification est restreinte à une sélection au hasard de 10 % des enregistrements, pour se réduire à la fin à environ 5 % ; si un fort pourcentage d'erreurs est encore trouvé,alors des enregistrements supplémentaires sont vérifiés.

Des procédures bien conçues telles que celles-ci peuvent aider à éduquer le nouvel utilisateur. Inversement, s'il n'y a pas de procédures, il y a peu de moyen d'assurer la cohérence entre les opérateurs et entre les tâches.

Responsabilité

L'attribution de la responsabilité de l'ensemble de la qualité des données peut aider les organismes à atteindre un niveau raisonnable de contrôle qualité, à fournir un référent pour le retour d'information sur les erreurs, et un contact pour la documentation et les requêtes.

Une mauvaise formation est la cause de nombreux problèmes de qualité des données.

Données taxonomiques et nomenclaturales

> *De mauvaises données taxonomiques peuvent contaminer des domaines d'étude connexes (Dalcin 2004).*

La taxonomie est la théorie et la pratique de la classification des organismes (Mayr et Ashlock 1991). La plupart des données sur les espèces qui nous intéressent ici incluent une part taxonomique (ou nomenclaturale) (par exemple le nom de l'organisme et sa classification) – dénommée le « domaine de classification des données » par Dalcin (2004). La qualité de cette part des données et la manière dont cette qualité peut être déterminée diffère considérablement de la part spatiale des données, car elle est habituellement plus abstraite et plus difficile à quantifier.

Les données taxonomiques consistent en (tous ces aspects ne sont pas toujours présents) :

- le nom (scientifique, commun, la hiérarchie, le rang)
- le statut nomenclatural (synonyme, accepté, typification)
- la référence (auteur, date et lieu de publication)
- la détermination (par qui et quand l'enregistrement a été identifié)
- les champs qualité §exactitude de la détermination, qualifiants)

Une des principales sources d'erreur dans les noms taxonomiques est la mauvaise orthographe. La détection des erreur d'orthographe dans les bases de données taxonomiques peut être une tâche simple quand il s'agit des noms scientifiques qui représentent la hiérarchie comme les noms de Famille et de Genre (Dalcin 2004). Dans ces cas-là, des fichiers de référence sont généralement disponibles pour la plupart des groupes taxonomiques. Il y aussi des listes de noms d'espèces de plus en plus étoffées disponibles grâce à des projets comme Species 2000 (http://www.species2000.org) et le programme de travail ECat du GBIF (http://www.gbif.org/prog/ecat).. L'utilisation des noms d'espèces ou épithètes seuls sans le genre associé dans un fichier de référence est rarement satisfaisant du fait que nombre d'épithètes spécifiques peuvent varier légèrement d'un genre à un autre. Une méthode d'identification des erreurs d'orthographe consiste à détecter et isoler des erreurs dans les noms scientifiques qui ont un haut degré de similarité sans être exactement identiques (Dalcin 2004, CRIA 2005)[4].

La méthode de loin la plus satisfaisante de réduire la possibilité d'orthographes erronées dans les noms scientifiques consiste à intégrer des fichiers de référence dans le processus de saisie en utilisant des menus déroulant de noms de familles, genres et espèces, etc. Dans une situation idéale où les fichiers de référence sont disponibles, la mise en œuvre de ces techniques devrait réduire l'occurrence de ce type d'erreur à quasiment zéro. Malheureusement, il y a de larges régions du monde et nombre de groupes taxonomiques importants pour lesquels de telles listes ne sont toujours pas disponibles.

Lorsque des fichiers de référence sont importés d'une source externe comme Catalogue of Life ou ECat, l'identifiant de la source devrait être enregistré dans le base afin que les changements apportés lors des mises à jour de ces fichiers puissent être aisément répercutés dans la base, et que celle-ci soit mise à jour en conséquence. On peut espérer que ceci sera

[4] http://www.webopedia.com/TERM/G/GUID.html

bientôt facilité grâce à la mise en œuvre d'identifiants globaux uniques (en anglais, « Globally Unique Identifiers » : GUIDs).

L qualité taxonomique des données repose largement sur l'expertise taxonomique disponible. Le handicap taxonomique (Environment Australia 1998) et le déclin dans le monde entier du nombre de chercheurs taxonomistes bien formés vont entraîner une diminution de la qualité à long terme de la taxonomie opérationnelle et donc de la qualité des données primaires sur les espèces (Stribling *et al.* 2003). L'initiative taxonomique mondiale (en anglais, « Global Taxonomic Initiative » : GTI) (CBD 2004) tente de réduire ce handicap, mais le problème risque fort de perdurer. La qualité peut aussi se détériorer avec le temps, surtout dans les cas où il n'y a pas de spécimens disponibles et bien conservés (c'est ce qui se passe avec la plupart des données d'observation et d'inventaires) ou dans les domaines où il n'y a pas de bonne compétence taxonomique.

La capacité d'une institution à produire des résultats d'une bonne qualité taxonomique (y compris des données primaires sur les espèces qui soient bien documentées) est influencée par (d'après Stribling *et al.* 2003) :

- le niveau de formation et d'expérience du personnel,
- le niveau d'accès à la littérature technique, aux spécimens de référence et aux spécialistes taxonomistes,
- la possession d'équipements de laboratoire et locaux appropriés, et
- l'accès à l'Internet et les ressources qui y sont disponibles.

Enregistrement et exactitude de l'identification

Traditionnellement, les muséums et les herbiers utilisent un système de determinavit où les experts examinent de temps en temps les spécimens de leur compétence et en font la détermination. Ceci se fait souvent à l'occasion de révisions taxonomiques ou de visites dans les collections. C'est une méthode éprouvée mais qui prend du temps et très dépendante des circonstances. Mais il n'y a pas de solution alternative, du fait que l'identification automatisée ne semble pas près d'être opérationnelle à court ni à long terme.

L'enregistrement de l'identification pourrait inclure un champ contenant l'indication de son degré d'exactitude. La date de la détermination est habituellement notée dans la plupart des bases de données. Le champ dédié à l'exactitude serai codé et ses valeurs possibles exprimeraient les cas suivants (Chapman 2004) :

- identifié par un expert mondial dans le taxon avec une grande certitude,
- identifié par un expert mondial dans le taxon avec une bonne certitude,
- identifié par un expert mondial dans le taxon avec des doutes,
- identifié par un expert régional dans le taxon avec une grande certitude,
- identifié par un expert régional dans le taxon avec une bonne certitude,
- identifié par un expert régional dans le taxon avec des doutes,
- identifié par un non expert du taxon avec une grande certitude,
- identifié par un non expert du taxon avec une bonne certitude,
- identifié par un non expert du taxon avec des doutes,
- identifié par le collecteur avec une grande certitude,
- identifié par le collecteur avec une bonne certitude,
- identifié par le collecteur avec des doutes,

La manière dont on pourrait ordonner ces valeurs est ouverte à discussion, de même que le fait que le choix des catégories. Je crois comprendre que certaines institutions ont déjà un champ de cette nature, mais à ce stade, je n'ai pas pu trouvé d'exemple. La version 4 de la norme HISPID[5] (Conn 2000) en inclut en fait une version simplifiée – l'« indicateur du niveau de vérification », avec cinq valeurs, présentées dans le tableau ci-dessous.

Tableau 3. *Indicateur de niveau de vérification dans HISPID (Conn 2000).*

0	Le nom n'a pas été vérifié par un expert
1	Le nom a été déterminé par comparaison avec d'autres spécimens identifiés
2	Le nom a été déterminé par un taxonomiste ou par une autre personne compétente à l'aide de l'herbier et/ou la bibliothèque et/ou du matériel vivant documenté
3	Le nom de la plante a été déterminé par un taxonomiste qui a entrepris de réviser le groupe
4	L'enregistrement provient d'une collecte de type de la propagation asexuée à partir d'un matériel type

De nombreuses institutions enregistrent déjà le degré de certitude, en utilisant des expressions comme « aff », « cf », « *s. lat* », « *s. str* », « ? ». Bien que certaines de ces expressions aient des définitions rigoureuses,leur utilisation par les individus varient considérablement. L'utilisation de *sensu stricto* et de *sensu lato* implique des variations dans le concept taxonomique.

En outre, lorsque les noms sont dérivés d'ailleurs que l'expertise taxonomique, on pourrait lister leurs sources (d'après Wiley 1981) :

- descriptions de nouveaux taxons,
- révisions taxonomiques,
- classifications,
- clés taxonomiques,
- études faunistiques ou floristiques,
- atlas,
- catalogues,
- checklists,
- manuels,
- enseignement en taxonomie / règles de nomenclature,
- analyse phylogénétique.

L'incertitude peut généralement être réduite, et la qualité améliorée, en faisant appel à au moins deux publications ou spécialistes. Les divergences d'identification entre taxonomistes n'impliquent pas toutefois que l'une des identifications soit erronée : il peut s'agir d'une divergence d'opinion (ou d'hypothèse) taxonomique.

[5] « Herbarium Information Standards and Protocols for Interchange of Data », norme d'échange d'information sur les spécimens d'herbiers, développée par un comité issu des herbiers d'Australie.

Précision de l'identification

Selon Stribling et al. (2003), la précision de l'identification (qu'ils ont à tort désignée par « précision taxonomique ») peut être évaluée en comparant les résultats du traitement par deux taxonomistes ou spécialistes d'un échantillonnage tiré au hasard. On peut aussi l'évaluer en comparant les noms donnés à des doubles, détenus (et identifiés) par différentes institutions. Il s'agit de notions relativement abstraites, et je ne suis pas sûr de l'intérêt d'enregistrer ce type d'information.

Un deuxième aspect de la précision de l'identification, toute fois, concerne le niveau auquel le spécimen est identifié. Une identification à l'espèce, ou à la sous-espèce, est plus précise que si elle est prononcée seulement au genre ou à la famille. Dans la documentation d'un jeu de données, il peut être intéressant pour les utilisateurs de savoir que 50 % de l'identification est au genre seulement – ce qui est le cas dans beaucoup de groupes d'animaux.

Biais

Le biais est une erreur systématique qui provient d'une dérive uniforme des valeurs (Chrisman 1991). Il se produit souvent du fait d'une méthodologie appliquée régulièrement et qui produit une erreur systématique par nature. Des biais dans la nomenclature taxonomique peuvent provenir d'une identification précise mais inexacte. De tels biais peuvent résulter de la mauvaise interprétation d'une clé dichotomique ou d'une structure morphologique, de l'usage d'une nomenclature invalide ou d'une publication obsolète (Stribling et al. 2003) ou de l'usage d'une publication inadéquate (par exemple la Flore d'une autre région, qui peut ne pas mentionner les bons taxons pour la zone étudiée).

Cohérence

L'incohérence peut se produire sur le plan de la classification dans les bases de données lorsque deux noms ou plus sont acceptés et représentent le même taxon (par exemple *Eucalyptus eremaea* et *Corymbia eremaea*). Ceci peut correspondre à des opinions divergentes, ou à des erreurs dues à des orthographes multiples (par exemple, Tabernaemontana hystrix, Tabernaemontana histryx, Tabernaemontana histrix – CRIA 2005).

Complétude

Motro et Rakov (1998, d'après Dalcin 2004) désignaient la complétude comme « le fait que toutes les données soient disponibles » et distinguaient la complétude des fichiers (aucun enregistrement ne manque), et la complétude des enregistrements (tous les champs sont connus dans chaque enregistrement).

La complétude en termes taxonomiques (c'est à dire dans une base de noms ou de taxons) se réfère à la couverture des noms. Est-ce que la base inclut des noms à tous les niveaux de la hiérarchie (par exemple jusqu'à la sous-espèce ou seulement à l'espèce) ? Quelle portion du règne animal ou végétal la base couvre-t-elle ? Inclut-elle les synonymes ? Tous ces aspects sont importants pour l'utilisateur s'il veut déterminer l'aptitude des données à l'usage qui l'intéresse. Dalcin (2004), par exemple, distingue la *complétude nomenclaturale*, qui correspond à l'inclusion de tous les noms dans un contexte donné (par exemple, dans un contexte taxonomique – la présence de tous les noms pour un groupe taxonomique particulier - ; ou dans un contexte spatial – la présence de tous les noms d'une région donnée) et la

complétude de classification, qui correspond à la présence de tous les noms reliés à nom « accepté » d'un taxon donné (c'est à dire le synonymie complète).

Dans le cas d'une base de spécimens ou d'observations, la complétude peut correspondre à quelque chose comme « tous les champs du schéma Darwin Core sont-ils inclus ? » et « tous les champs du schéma Darwin Core sont-ils renseignés ? ». Dans une base de caractères, « tous les caractères pour toutes les phases de la vis sont-ils présents ? » (par exemple, les fruits pour les plantes, les stade de développement chez les insectes).

Collections de spécimens

On ne saurait trop insister sur l'importance des collections de spécimens, mais les bases de données ne peuvent pas toujours traiter de spécimens. De nombreuses bases de données d'observations sont constituées sans pour autant que des spécimens soient déposés. Il n'est pas non plus possible, pour des raisons politiques, juridiques, de conservation ou autres, de prélever un échantillon pour constituer un spécimen de référence dans tous les cas ou dans toutes les régions.

Lorsque le prélèvement de spécimens est possible, c'est souvent un exercice intéressant au début d'un programme qui traite d'espèces d'établir des accords de coopération entre les collecteurs et des institutions comme les muséums et les herbiers afin de déposer des spécimens de référence (Brigham 1998). De tels programmes devraient aussi comprendre un archivage approprié et des stratégies de dépôt, y compris des délais maximaux entre la collecte et le dépôt ou l'archivage.

Données spatiales

Les données spatiales ont souvent été en pointe dans le développement des normes pour la documentation des données : la norme « Spatial Data Transfer Standards » (Normes d'échange de données spatiales) (USGS 2004), le programme INSPIRE[6] (« Information for Spatial Information in Europe » : Information sur l'information spatiale en Europe), et beaucoup d'autres. Elles ont donc montré le chemin dans le développement des normes de qualité des données (comme la norme ISO 19115[7] qui régit les méta données relatives à l'information géographique). La nature numérique de beaucoup de données spatiales facilite l'utilisation de procédures statistiques, à la différence des données taxonomiques, et on pu développer à leur sujet nombre de méthodes de vérification de qualité des données (voir le document associé sur les *Principes et méthodes de nettoyage des données*).

Ceci ne signifie pas pour autant que toutes les composantes spatiales des données (le « domaine des données de terrain » de Dalcin 2004) sont faciles à numériser ou sont précises. De nombreuses collections historiques dans les muséums et les herbiers n'ont que des description textuelles très sommaires des lieux de collecte, et cela demande un effort important de les convertir en coordonnées géographiques. La nature même de ces collections peut augmenter la difficulté : si par exemple les spécimens ont été collectés à une époque où les collecteurs ne disposaient pas de cartes détaillées, où de nombreux noms de localités ne figurent plus dans les gazetiers ou sur les cartes que l'on publie de nos jours. L'adjonction d'une information de géo référencement aux données historiques, surtout là où l'on ne dispose pas de bons gazetiers historiques, peut demander beaucoup de temps avec un résultat souvent peu précis.

Un certain nombre d'outils ont été développés pour aider les utilisateurs à géo référencer leurs données, y compris des outils et des manuels en ligne. Ceux-ci ont été traités en détail dans le document associé sur les *Principes et méthodes de nettoyage des données*. En outre, la plupart des collecteurs se servent maintenant de GPS (« Global Positioning System » : Système mondial de localisation) pour enregistrer les coordonnées géographiques lors de la collecte. La précision des GPS est analysée au chapitre « *Saisie des données* ».

On peut rechercher les erreurs éventuelles dans les géo référencements inclus dans les données :

- en les confrontant à d'autres informations internes à l'enregistrement ou entre enregistrements dans la base de données – par exemple, les noms des régions administratives : Etats, Régions, Départements, etc.
- en les confrontant une référence externe : est-ce que les coordonnées sont cohérentes avec les lieux de récolte du collecteur ? est-ce que les coordonnées indiquent un point sur terre ou en mer ?
- en cherchant des valeurs aberrantes sur le plan géographique, ou
- en cherchant des valeurs aberrantes sur la plan environnemental.

Toutes ces méthodes seront détaillées dans le document associé sur les *Principes et méthodes de nettoyage des données*.

[6] http://www.ec-gis.org/inspire/
[7] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35

Précision spatiale

Comment mesure-t-on la précision de la localisation des données ?

Pour la plupart des couches de SIG[8] (topographie, cours d'eau, limites administratives, etc.) la précision est généralement aisée à déterminer du fait qu'il y a habituellement des sources externes de plus haute précision pour ces informations : repères géodésiques, intersections de rues ou de routes, etc. (Chrisman 1991). Cependant, beaucoup de ces tests ne sont pas simples, et la documentation – comme le « US National Map Accuracy Standard » (Norme nationale des Etats-Unis sur la précision des cartes) est compliquée. Traditionnellement, la précision spatiale est déterminée par comparaison avec un certain nombre de points bien définis selon des niveaux d'erreur acceptables spécifiés et exprimés sous forme d'écart type (Chrisman 1991). La notion d'écart type n'est pas aisée à appliquer à des points individuels ; elle s'applique mieux sur des jeux de données entiers ou des cartes numérisées. Dans le cas de points individuels, il est plus facile d'exprimer la distance à la véritable localisation à l'aide d'un rayon autour du point indiqué (Wieczorek *et al*. 2004) ou quelque chose de similaire. Deux facteurs comptent : la précision de la mesure du point de référence et la précision de la mesure effectuée par rapport à ce point. Par exemple, si la position d'une intersection est connue à 100 mètre près, la position d'un point de collecte mesurée par rapport à cette intersection sera connue avec une incertitude liée à cette mesure augmentée des 100 mètres de rayon autour de l'intersection (voir les commentaires dans Wieczorek 2001).

Le Comité fédéral sur les données géographiques des Etats Unis (en anglais « Federal Geographic Data Committee » : FGDC) a publié la norme de précision de la localisation géo spatiale en 1998. Ces documents contiennent une section sur les réseaux géodésiques et une dédiée à la précision des données spatiales (FGDC 1998).

- Le NSSDA[9] utilise l'écart type[10] pour exprimer la précision sur les positions, écart type calculé par rapport à des points de référence correspondants dont on connaît la position avec une grande précision.
- On parle aussi de niveau de confiance : la précision pour des distances au sol est généralement de 95 %. Un niveau de confiance de 95 % signifie que 95 % des positions dans le jeu de données auront une erreur par rapport aux vraies valeurs inférieure ou égale à la précision associée à ces positions. Cette précision prend en compte toutes les incertitudes, y compris celle liée à la précision des coordonnées des points de référence, à la transcription, et au calcul des coordonnées.

Citons un exemple d'indication de précision cartographique en Australie qui utilise de telles méthodes :

- *« La précision moyenne de cette carte est de ±100 mètres pour la position dans le plan horizontal des détails fins et de ±20 mètres pour l'altitude. » (Division de la cartographie nationale, feuille SD52-14, Edition 1, 1 :250.000).*

Ces précisions doivent être mentionnées pour chaque détermination du géo référencement d'un spécimen fondé sur un article ou une carte numérisée. Comme il y a toujours une

[8] Système d'Information Géographique

[9] « National Standard for Spatial Data Accuracy » : norme nationale des Etats-Unis sur la précision des données spatiales (NdT).

[10] L'écart type est une notion de base utilisée en statistique, la plus simple après la moyenne. Comme la moyenne, cette notion s'applique à un ensemble donné de valeurs. Il s'agit de la racine carrée de la somme des carrés des différences entre chaque valeur et la moyenne de ces valeurs.

incertitude sur la précision des données spatiales, on ne peut se contenter d'une indication globale de précision et il est important d'inscrire les précisions connues. En effet, les erreurs se propagent au long de la chaîne de l'information et contribuent à une incertitude sur les résultats finaux, qu'il s'agisse d'une carte produite à partir d'un Système d'Information Géographique (SIG) ou d'une carte de répartition potentielle d'espèce produite à l'aide d'un logiciel de modélisation de la distribution (Heuvelink 1998).

Projet BioGeomancer

Un projet intitulé BioGeomancer a été récemment financé par la Fondation Gordon et Betty Moore pour contribuer à améliorer le référencement des données primaires sur les espèces et à évaluer , améliorer et documenter sa précision. Ce projet devrait rendre son rapport et donner accès aux outils développés courant 2006.

Fausses précision et exactitude

Un autre facteur à prendre en considération est celui de la fausse précision et de la fausse exactitude. De nombreux utilisateurs de SIG n'ont pas conscience des problèmes liés à la précision, les erreurs et l'incertitude dans données spatiales, et ils supposent souvent que leurs données sont parfaitement précises. Ils indiquent souvent des niveaux de précision qui sont incompatibles avec la provenance de leurs données. Beaucoup d'institutions s'appuient maintenant sur des SIG pour le géo référencement et, en zoomant à des niveaux que ne supportent pas les données (et en utilisant des degrés décimaux) elles peuvent aboutir à une précision irréaliste. Il faut ajouter qu'en utilisant un GPS pour enregistrer la localisation d'une collecte, celle-ci est souvent consignée à 1 ou 2 mètres près, alors qu'en réalité beaucoup de GPS portables en usage ne fournissent qu'une précision à environ 10 mètres près ou plus. Ceci est particulièrement vrai lorsqu'on utilise les GPS pour déterminer l'altitude (voir les commentaires sur ce sujet dans le chapitre « Saisie des données » ci-dessous).

Données sur les collecteurs et les collectes

L'information sur le collecteur et la collecte (le « domaine des données de collecte » de Dalcin 2004) comprend des renseignements sur l'acte de collecte lui-même (le collecteur, la date), et des informations supplémentaires sur l'habitat, le sol, les conditions météorologiques, l'expérience des observateurs, etc. Ces informations peuvent se décliner de la manière suivante (adapté de Conn 1996, 2000) :

- Nom du collecteur et numéro de collecte
- Expérience des observateurs
- Date ou période de la collecte
- Méthode de collecte (particulièrement important pour les données d'observation ou d'inventaire)
- Données associées

Beaucoup de ces notions varient considérablement selon le type de données collectées - que ce soit pour un spécimen déposé dans un muséum, une observation, ou les résultats d'un inventaire fin. Dans le cas d'une collection statique comme dans les muséums, ce sont le nom du collecteur, le numéro de collecte et la date sont des attributs clés, qui peuvent être associés à des données concernant le type morphologique, l'habitat, etc., et peut-être la méthode de capture (pour les animaux). Pour ce qui est des données d'observation, ce sont des informations comme la durée de l'observation, l'étendue couverte, le moment dans la journée (heures de début et fin de l'observation en sus de la date), et des données associées telles que les conditions météorologiques, le sexe de l'animal observé, son activité, etc. Avec les données d'inventaire, ce sont des informations sur les méthodes d'inventaire, la taille (de la grille, de l'étendue), l'intensité de l'effort d'inventaire, les conditions météorologiques, la fréquence, l'indication de la collecte éventuelle de spécimens et leurs numéros de référence, etc. ainsi que nombre des informations mentionnées pour les observations.

Exactitude des attributs

Les problèmes qui peuvent influer sur la qualité des données dans le domaine de l'information sur la collecte concernent notamment la manière dont les noms des collecteurs, les numéros de collecte, les initiales, etc. sont enregistrés (Koch 2003), l'exactitude de l'enregistrement de la date et de l'heure, la cohérence de l'enregistrement des données associées au moment de la collecte, comme le type morphologique, l'habitat, le sol, le type de végétation, la couleur des fleurs, le sexe,ou les espèces associées.

Un exemple de problème récurrent dans les données de collecte, est celui de la numérotation des collectes par le collecteur : en effet, certains collecteurs n'utilisent pas une numérotation unique pour l'ensemble de toutes leurs collectes. Ceci peut induire une perte de qualité des données puisque ces numéros sont parfois utilisés pour aider à retrouver les localisations des collectes, les identifications, les doubles distribués entre plusieurs collections, etc.

Cohérence

La cohérence de la terminologie utilisée pour renseigner une collecte est très variable, et il est rare que les données associées, en particulier, soient notées de manière cohérente au sein d'un même jeu de données, sans parler de la cohérence entre divers jeux de données.

Complétude

La complétude de l'information sur les collectes est elle aussi très variable. Les informations sur l'habitat, le numéro de collecte, le fait que la plante soit ou non en fleur, etc. sont souvent absentes de nombreux enregistrements. Ceci rend par exemple difficile d'étudier les habitats à partir des seules collections.

Données descriptives

Les bases de données descriptives sont de plus en plus utilisées à la fois pour stocker de l'information et comme méthode de publication, souvent en lieu et place des publications traditionnelles. La morphologie, la physiologie et la phénologie sont des exemples de domaines de ces données descriptives, qui sont souvent utilisées en analyse cladistique, ou pour générer automatiquement des descriptions et des outils d'identification.

Le groupe de travail sur les bases de données taxonomiques (en anglais, « Taxonomic Databases Working Group » : TDWG, prononcé habituellement « tadouigue ») a une longue histoire dans le développement et la promotion de normes dans le domaine des données descriptives. D'abord avec son soutien de la norme DELTA (Dallwitz et Paine 1986) et plus récemment avec la mise en place du groupe de travail sur la « structure des données descriptives » (en anglais, « Strucure of Descriptive Data » : SDD ; http://160.45.63.11/Projects/TDWG-SDD/).

La qualité des données descriptives peut varier, et bien que les données élémentaires soient souvent mesurées, en réalité leur exactitude peut être limitée dans les cas où les données ne sont pas observables (comme dans le cas des données historiques), ne sont pas pratiques à observer (par exemple parce que cela serait trop coûteux) et/ou correspondent à une perception plus qu'à une réalité (c'est le cas de l'évaluation subjective d'une couleur, d'une abondance, etc.).

Dans la plupart des cas, les données descriptives sont stockées au niveau d'une espèce plutôt qu'au niveau d'un spécimen, et elles sont de ce fait généralement moyennées ou indiquées sous forme d'intervalle de valeurs. Comme le fait remarquer Morse (1974, rapporté par Dalcin 2004), l'information taxonomique est intrinsèquement moins fiable que celle provenant de l'observation d'un spécimen. Nonobstant, il y a actuellement une tendance croissante à enregistrer l'information descriptive, au moins pour certaines données, au niveau spécimen, avec une augmentation de la qualité en conséquence.

Complétude

Au niveau spécimen, la complétude des enregistrements de données descriptives dépend de la qualité du spécimen, de l'époque de l'année, etc. Par exemple, il peut ne pas être possible d'enregistrer les caractéristiques du fruit ou de la fleur du même spécimen. De ce fait, certains champs devront être laissés en blanc. Dans d'autres cas, l'attribut peut ne pas être pertinent pour le caractère et tous les attributs ne seront pas enregistrés.

Cohérence

Il peut y avoir des problèmes de cohérence entre deux données liées. Par exemple, les deux caractéristiques suivantes peuvent être enregistrées pour une même espèce (Dalcin 2004) :

- TYPE MORPHOLOGIQUE = HERBACE et
- USAGE = BOIS

Une représentation incohérente du même attribut peut aussi affecter la qualité, surtout quand de mauvaises définitions des attributs sont utilisées ou quand des normes cohérentes ne sont pas strictement appliquées. Par exemple (Dalcin 2004) :

- COULEUR DE LA FLEUR = CARMIN, et
- COULEUR DE LA FLEUR = POURPRE

L'utilisation de terminologies standard peut aider à réduire considérablement le degré d'erreur et d'interprétation erronée. Des terminologies standard sont en cours de développement dans un grand nombre de domaines et de disciplines, et la tendance récente à développer des bases de données descriptives fédératives a accru la cohérence de la mise en œuvre de ces terminologies. Le développement de la norme du TDWG pour la structure des données descriptives (en anglais, « Structure of Descriptive Data » : SDD) (TDWG 2005) ne peut que favoriser ces progrès.

Récolte des données

Il y a diverses manières de récolter les données primaires sur les espèces et les données sur les occurrences d'espèces, chacune ayant ses propres niveaux de précision et d'exactitude, en même temps que ses propres sources d'erreur et d'incertitude. Chacun de ces aspects a un impact différent sur « qualité d'usage » résultante des données. Plusieurs méthodes parmi les plus couramment utilisées pour les données relatives aux espèces sont examinées brièvement ci-dessous.

Opportuniste

La plupart des données d'occurrence d'espèce ont été collectées de manière opportuniste. Beaucoup de ces collectes sont maintenant stockées sous forme de spécimens dans les muséums et les herbiers. La plupart des données historiques ne contenaient qu'une référence de localisation textuelle, comme « à 5 km au Nord-Ouest de telle ville », et on rarement été pourvues d'un géo-référencement au moment de leur collecte. L'ajout du géo-référencement a généralement été effectué ultérieurement, et habituellement par quelqu'un d'autre que le collecteur (Chapman et Busby 1994). Beaucoup de données d'observation (Atlas ornithologique, etc.) ont aussi été récoltées de manière opportuniste.

Ces données sont habituellement numérisées par lots, et le géo-référencement en est généralement effectué par référence à des cartes imprimées. Il est le plus souvent peu précis. La majorité de ces données ne peuvent pas être considérées comme ayant une précision supérieure à 2 voire 10 km.

Inventaire de terrain

Les données issues d' inventaires de terrain incluent généralement une référence spatiale, souvent sous la forme d'un couple longitude – latitude ou d'une référence UTM. Ces références spatiales ont généralement une précision de 100 à 250 mètres. Il faut cependant faire attention à ce à quoi se réfère la localisation – il peut ne pas s'agir de l'observation elle-même, mais par exemple du milieu d'un transect, ou du coin (ou du centre) d'une maille carrée, et ceci n'est pas toujours explicite. De plus, comme il y a rarement des spécimens associés, l'exactitude taxonomique n'est pas toujours bonne. C'est particulièrement le cas lorsque les données sont anciennes et que la taxonomie a évolué.

Observations à grande échelle

Certains inventaires biologiques enregistrent des données à l'échelle globale d'une maille ou d'un périmètre particuliers. Par exemple, un inventaire des espèces au sein d'un parc national, ou bien les observations des oiseaux effectuées sur des mailles de 10 minutes d'arc de côté (exemple : Oiseaux d'Australie 2001, 2003). La précision spatiale de tels enregistrements ne peut être que de 1 à 10 km ou plus.

Système mondial de localisation (GPS)

Le système mondial de localisation (en anglais « Global Positioning System » : GPS) est de plus en plus utilisé pour la collecte de données sur les espèces. Non seulement pour les

données d'inventaire, mais aussi pour les données de récolte opportuniste et pour les données d'observation.

La technologie du GPS utilise la triangulation pour déterminer la localisation d'un point de la surface du globe. La distance mesurée est celle entre le récepteur GPS et les satellites du système (Van Sickle 1996). Comme la position dans l'espace des satellites GPS est connue, on peut calculer celle du point sur la Terre. Il faut au moins quatre satellites GPS pour déterminer une position sur la surface terrestre (McElroy *et al*. 1998, Van Sickle 1996). Ceci ne constitue généralement pas une limitation aujourd'hui, car on peut souvent recevoir 7 satellites ou plus dans la plupart des endroits du globe ; toutefois dans les débuts, le nombre des satellites que l'on pouvait recevoir n'était pas toujours suffisant. Antérieurement à Mai 2000, la plupart des récepteurs GPS civils étaient bridés en précision par une « accessibilité sélective » (aux satellites). La suppression de cette restriction a grandement amélioré la précision que l'on peut généralement attendre (NOAA 2002).

Avant le retrait de l'« accessibilité sélective », la précision des récepteurs GPS utilisés par la plupart des biologistes et observateurs sur le terrain était de l'ordre d'environ 100 mètres au mieux (McElroy *et al.* 1998, Van Sickle, 1996, Leick 1995). Depuis, cependant, la précision des récepteurs GPS s'est améliorée et aujourd'hui, la plupart des fabricants de GPS portables promettent une erreur inférieure à 10 mètres en terrain découvert en utilisant 4 satellites ou plus. La précision peut être améliorée encore en faisant la moyenne des résultats de plusieurs mesures de position au même point (McElroy *et al.* 1998), et certains GPS modernes qui incluent des algorithmes de calcul de moyenne peuvent monter la précision à environ 5 mètres ou parfois mieux ;

L'utilisation de GPS différentiels peut améliorer la précision considérablement. Cette méthode différentielle s'appuie sur une station GPS de référence (habituellement un point de contrôle d'inventaire) de position connue pour calibrer le récepteur. La station de référence et le récepteur portable mesurent leur position par rapport aux satellites en même temps, et cela réduit l'erreur liée aux conditions météorologiques. De cette manière, en comparant la position mesurée par la station de référence à sa position connue, le récepteur manuel peut en déduire les corrections à appliquer à la position qu'il a mesurée pour lui-même. Selon la qualité des récepteurs que l'on utilise, on peut attendre une précision entre 1 et 5 mètres. Cette précision décroît à mesure que la distance à la station de référence s'accroît. Là aussi, le calcul d'une moyenne peut encore améliorer ces résultats (McElroy *et al.* 1998).

Le « système d'augmentation de large zone (en anglais « Wide Area Augmentation System » : WAAS) est un système de navigation et d'atterrissage fondé sur le GPS développé pour le guidage précis d'aéronefs (Administration fédérale de l'aviation des Etats-Unis 2004). WAAS utilise des antennes au sol dont les positions connues avec précision permettent d'accroître la précision du GPS. Des technologies similaires telles que le système d'augmentation de zone locale (en anglais « Local Area Augmentation System » : LAAS) sont aussi développées afin d'améliorer encore la précision.

Il est même possible d'atteindre de plus grandes précisions encore à l'aide soit du GPS différentiel en temps réel (McElroy *et al.* 1998) soit du GPS statique (McElroy *et al.* 1998, Van Sickle 1996). Le GPS statique utilise des instruments de haute précision et des techniques de spécialistes, et il n'est généralement utilisé que par les agents en charge d'inventaires. Des inventaires conduits en Australie et utilisant ces techniques ont rapporté des précisions de l'ordre du centimètre. Il est peu probable que ces techniques soient utilisées

à grande échelle dans la collecte de données biologiques à cause de leur coût et du fait qu'une telle précision n'est généralement pas demandée.

Afin d'obtenir des niveaux de précision tels que ceux décrits ci-dessus, le récepteur GPS doit être placé dans une zone libre d'obstacles au-dessus du récepteur et dépourvue de surfaces réfléchissantes, et il doit avoir un bon champ de vision vers l'horizon (par exemple, les récepteurs ne fonctionnent pas bien sous une épaisse canopée). Le récepteur doit être capable d'enregistrer les signaux d'au moins quatre satellites GPS dans un arrangement géométrique approprié. Le meilleur arrangement consiste à avoir « *un satellite directement au-dessus et les trois autres uniformément répartis autour de l'horizon* » (McElroy *et al.* 1998). Le récepteur doit aussi être paramétré selon la zone et ce paramétrage doit être enregistré.

Quant à la mesure de la hauteur par GPS, la plupart des biologistes en savent peu sur ce sujet. Il faut noter que la hauteur affichée par un récepteur GPS est en fait mesurée par rapport à la surface d'un ellipsoïde modèle de la Terre et non par rapport au niveau de la mer ou par rapport à une hauteur de référence comme celle utilisée sur le territoire australien. En Australie par exemple, la différence entre la hauteur affichée par un récepteur GPS et l'altitude peut varier entre -35 et +80 mètres, et cette variation tend à être imprévisible (McElroy *et al.* 1998, Van Sickle 1996).

Saisie des données

> *La saisie des données est intrinsèquement sujette à des erreurs simples ou complexes* (Maletic et Marcus 2000)

Saisie simple

La première étape dans la saisie des données est habituellement la saisie de l'information portée sur une étiquette de spécimen, une revue, un carnet de terrain, un registre ou un ensemble de fiches. Ceci peut être effectué manuellement par des personnes qualifiées ou non, ou en scannant le document. Le niveau d'erreur dû à la saisie peut souvent être réduit par le biais d'une double saisie, ou en utilisant un logiciel adaptatif pour le scan, et en faisan appel à des experts et des superviseurs pour tester les données saisies sur la base d'un échantillonnage (voir le manuel MaPSTeDI[11] cité pus bas).

Interfaces utilisateur

Le développement d'une interface de saisie spécifique peut aussi être un moyen de diminuer le taux d'erreur. De nombreuses institutions confient la saisie à des personnels non qualifiés ou à des bénévoles, et le développement d'une interface simple (non technique) avec laquelle les opérateurs se sentent à l'aise peut améliorer l'exactitude de la saisie. Une telle interface peut faciliter la saisie en permettant de parcourir rapidement les référentiels, les saisies précédentes, les bases de données connexes, et même d'utiliser des moteurs de recherche comme Google qui peuvent aider un opérateur à décider de l'orthographe ou de la terminologie correcte quand il a du mal à lire l'étiquette ou à déterminer ce qui doit ou ne doit pas être placé dans certains champs. Dans certains cas la base de données peut être adaptée à ces utilisations, en incorporant les référentiels et les menus déroulants qui évitent à des personnels non qualifiés des décisions quant aux noms, au localités, ou aux habitats.

Géo-référencement

Les cartes sont l'un des moyens les plus efficaces pour communiquer l'information, et cela seul suffit à justifier l'augmentation récente de la numérisation et du géo-référencement des spécimens dans les muséums et les herbiers, ainsi que des observations. La capacité accrue de manipulation des données par les cartes permet de mieux étudier, identifier, visualiser, documenter et corriger les erreurs et les incertitudes (Spear et al. 1996). Elle fournit aussi une méthode puissante pour visualiser et communiquer les incertitudes intrinsèques aux données, et ainsi permettre aux utilisateurs de déterminer la qualité ou pertinence d'usage des données.

Saisir les données et les géo-référencer peut être difficile et fastidieux. Les résultats du projet MaPSTeDI (Université du Colorado 2003) suggère qu'un opérateur compétent peut géo-référencer un enregistrement toutes les cinq minutes. D'autres études (Armstrong 1992, Wieczorek 2002) ont montré que le géo-référencement peut demander nettement plus longtemps – par exemple, la base de données MANIS suggère un rythme d'environ 9 par heure pour des localités des Etats-Unis, de 6 par heure en Amérique du nord hors Etats-Unis et 3 par heure hors d'Amérique du Nord (Wieczorek 2002).

[11] « Initiative informatique de la base de données spatio-temporelles des montagnes et des plaines » (en anglais « Mountains and Plains Spatio-Temporal Database Informatics Initiative ») de l'Université du Colorado (NdT).

MaNIS/HerpNet/ORNIS
Manuel de géo-référencement
http://manisnet.org/manis/GeorefGuide.html

MaPSTeDI
Géo-référencement dans MaPSTeDI
http://mapstedi.colorado.edu/geo-referencing.html

Un certain nombre d'excellents manuels et méthodes ont été développés pour aider les gestionnaires de données à effectuer le géo-référencement. Le manuel développé par John Wieczorek au Muséum de zoologie des vertébrés de Berkeley (Wieczorek 2001) et le manuel MaPSTeDI (Université du Colorado 203) sont deux des études les plus complètes qui ont été conduites sur le sujet à ce jour, et je renvoie le lescteur à ces manuels. Ils traitent de la détermination de l'exactitude et de la précision d'un point à partir de la toponymie, des incertitudes qui découlent de l'usage des référentiels géodésiques différents, des conséquences de l'usage d'échelles différentes, etc. Ce sont des ouvrages très complets sur ce sujet et j'espère que les lecteurs du présent manuel les considèreront comme des parties adjointes.

Il existe aussi de nombreux outils en ligne qui peuvent aider à déterminer le géo-référencement – par exemple pour des lieux situés dans une direction donnée à une distance donnée d'une localité connue. Il sera traité de ces outils plus en détail dans le document associé « *Principes et méthodes de nettoyage des données* ».

BioGeoMancer
(Muséum d' Histoire Naturelle « Peabody »)
http://www.biogeomancer.org/

geoLoc
(Centre de référence sur l'information environnementale)
http://splink.cria.org.br/tools/

Erreur

Les outils mentionnés ci-dessus sont efficaces pour réduire le taux d'erreur et augmenter la qualité. Mais aucune méthode de géo-référencement ne peut totalement éliminer les erreurs. Comme indiqué dans le manuel MaPSTeDI :

> « *Bien que le géo-référencement ne soit pas une science exacte, et qu'aucune collection ne puisse être géo-référencée correctement à 100%, la vérification de la qualité améliore considérablement le pourcentage de la collection qui est correctement géo-référencé. Chaque projet devrait en tenir compte dans la programmation de son géo-référencement.* » (Université du Colorado 2003).

Une source commune d'erreur de géo-référencement est l'utilisation non critique des gazetiers électroniques. Dans certains cas ces gazetiers ont été développés par des projets afin de produire des cartes papier, et l'emplacement d'un point donné dans le gazetier est le coin

en bas à gauche du rectangle dans lequel le nom devait être imprimé sur la carte, et non l'emplacement du point en question (c'est le cas par exemple du gazetier d'Australie d'avant 1998 développé par le Groupe d'information sur le territoire australien). On peut espérer que la plupart des gazetiers ont été corrigés, mais il peut déjà y avoir des coordonnées géographiques qui ont été saisies dans les bases de données de muséums ou d'herbiers sur la base de ces valeurs. L'exactitude de tels enregistrements devrait être évaluée à l'aide de vérifications aléatoires de l'emplacement des localités, en se référant à des gazetiers valides ou à des cartes précises à grande échelle.

Il est souvent plus rapide et plus efficace d'effectuer le géo-référencement comme une activité séparée après la saisie de l'information portée sur l'étiquette. Ceci permet d'utiliser la base de données pour trier les spécimens par localité, collecteur, date, etc. et de mieux utiliser les cartes pour en tirer le géo-référencement. Ceci évite aussi de saisir plusieurs fois les coordonnées géographiques de la même localité.

Documenter les données

> « *Les métadonnées sont des données sur les données. Il s'agit d'une description des caractéristiques des données qui est établie dans un but précis*. » (ANZLIC 1996a)

Il y a besoin d'une bonne documentation à la fois au niveau du jeu de données et de la donnée unitaire.

Les métadonnées fournissent de l'information sur les jeux de données : contenu, étendue, accessibilité, tenue à jour, complétude, aptitude à l'usage et pertinence. Quand les métadonnées sont fournies l'utilisateur peut se faire une idée de la qualité du jeu de données et de sa pertinence avant de l'utiliser. De bonnes métadonnées permettent d'améliorer l'échange, le recherche et l'extraction des données. Les métadonnées traitent habituellement de l'ensemble du jeu de données, toutefois certains considèrent la documentation de la donnée au niveau de l'enregistrement (par exemple la documentation de la précision) comme une métadonnées au niveau unitaire. Quel que soit le nom qu'on lui donne, il est important d'avoir une bonne documentation au niveau à la fois du jeu de données et de l'enregistrement.

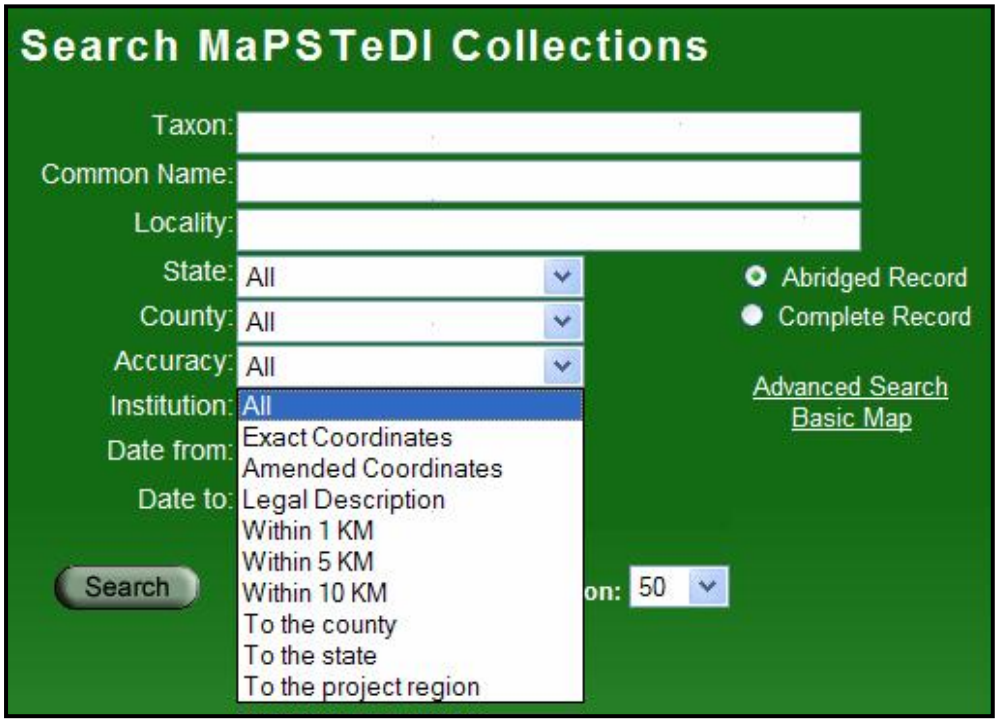
Toute données contient des erreurs – on ne peut y échapper ! ce qui compte c'est d'avoir connaissance de ces erreurs et de savoir si elles restent dans des limites acceptables au regard de ce que l'on veut en faire. C'est là que les métadonnées prennent toute leur importance au niveau global du jeu de données, et de fait c'est dans le domaine du développement des métadonnées que l'expression « aptitude à l'usage » est devenu prééminent. L'importance du concept d'aptitude à l'usage n'a été pleinement reconnue quant à l'information spatiale qu'au début des années 90, et ce n'est qu'au milieu des années 90 qu'il a fait son apparition dans la littérature dans ce contexte (Agumya et Hunter 1996).

Mais l'utilisateur ne peut toujours se contenter d'une information au niveau global du jeu de données. Il peut être extrêmement important, surtout avec les données sur les espèces, d'avoir une information sur l'erreur au niveau de la donnée unitaire, afin de déterminer l'aptitude à l'usage de chaque donnée. Lorsque cette information est disponible, un utilisateur peut demander, par exemple, uniquement les données qui ont une précision spatiale inférieure à 5.000 mètres. Il est aussi important que les outils de géo-référencement automatisé incluent un calcul de la précision spatiale dans les champs de sortie.

Les utilisateurs doivent aussi comprendre ce que recouvre le concept d'aptitude à l'usage. Les données d'occurrence d'espèces sont trop souvent extraites d'une base de données sous la forme « numéro d'enregistrement, x, y » sans qu'il soit tenu compte de l'information sur la précision qui est parfois présente. Les coordonnées sont toujours considérées comme représentant un point, alors qu'elles se réfèrent rarement, sinon jamais, véritablement à un point. Certains enregistrements peuvent avoir été saisis avec les coordonnées d'un point arbitraire (par exemple des spécimens dont l'étiquette mentionnait seulement « Amérique du Sud »), et s'être vu de ce fait attribuer une précision de 5.000.000 mètres. Certaines bases de données sont comme cela ! Extraire l'enregistrement et utiliser son point arbitraire sera source de grande confusion. Lorsqu'il y en a un, les utilisateurs doivent informés de l'existence d'un champ « précision », et il faut leur indiquer comment l'utiliser. Dans les cas où les fournisseurs de données établissent des rapports normalisés sur les données, ils devraient rendre obligatoire d'inclure le champ « précision » dans les extractions de données.

Les données doivent être documentées avec des métadonnées assez détaillées pour que des tiers puissent les utiliser sans se référer au fournisseur initial des données.

Fig. 6. *Exemple de recherche de données utilisant l'outil MaPSTeDI* **http://www.geomuse.org/mapstedi/client/textSearch.html**. *L'exemple montre la possibilité de rechercher des données ayant une précision particulière* ***en utilisant la documentation au niveau unitaire.***



La documentation de l'exactitude, de la précision et de l'erreur dans les données spatiales est essentielle pour que les utilisateurs soient capables de déterminer la qualité de ces données vis à vis de leurs objectifs. Cette documentation devrait inclure (au moins) :

- titre du jeu de données
- source des données
- historique des données (actions effectuées sur les données depuis leur collecte ou leur obtention)
- précision (spatiale, temporelle, et d'autres attributs)
- cohérence logique
- date et durée de vie des données (validité temporelle et statut, fréquence de mise à jour)
- définition des champs de données
- méthodologie de collecte
- complétude
- conditions et contraintes d'utilisation (par exemple, copyright, restrictions lies aux droits, etc.)

- information sur la conservation et contact

Il n'est pas inutile de définir certains de ces termes, car les gestionnaires de données n'en sont pas tous familiers. Beaucoup de ces termes se réfèrent globalement à une collection plutôt qu'aux enregistrements individuels.

Précision spatiale

La précision spatiale se réfère à la proximité entre la description des coordonnées d'un objet comparée à sa véritable localisation (Planification du Minesota 1999). Lorsque c'est possible et qu'il est connu, le système géodésique utilisé pour déterminer les coordonnées devrait être noté.

Il est aussi recommandé que les bases de données incluent un champ pour mémoriser la précision spatiale de chaque enregistrement. Il y a plusieurs manières de faire cela. Certaines bases de données utilisent un code, cependant il est préférable d'utiliser une simple valeur métrique pour représenter la précision estimée de l'enregistrement (Chapman et Busby 1994, Conn 1996, 2000, Wieczorek *et al*. 2004). Ceci peut être important pour les utilisateurs qui extraient des données dans un but particulier – par exemple, ils peuvent ne vouloir que des données d'une précision supérieure à 2.000 mètres. Quelquefois, il peut être intéressant d'inclure un champ au niveau de l'enregistrement individuel pour noter comment l'information spatiale a été déterminée. Par exemple :

- à l'aide d'un GPS différentiel
- à l'aide d'un GPS limité par l'« accessibilité sélective » (récolte antérieure à 2002)
- sur une carte au 1 :100 000, par triangulation en s'appuyant sur des repères facilement identifiables
- sur une carte, par estimation de proche en proche le long d'un cheminement
- sur une carte, à distance (par exemple depuis un hélicoptère)
- automatiquement, à l'aide d'un logiciel fondé sur la méthode « point-rayon »
- avec un gazetier, en indiquant le nom, la date et la version du gazetier

Précision des attributs

La précision des attributs recouvre l'évaluation de la fidélité de la description des paramètres constitutifs de la données par rapport à leurs valeurs dans la réalité. Idéalement, il faudrait une liste des attributs et une information sur la précision de chacun ; par exemple,

> *Les enregistrements sont fournis par des observateurs expérimentés. Une précision additionnelle s'obtient en testant l'exactitude des attributs par rapport aux spécimens déposés dans le muséum ou l'herbier aux fins de vérification par les experts. Environ 40 % des enregistrements sur les plantes sont vérifiés à l'aide des spécimens, 51 % pour les amphibiens, 12 % pour les mammifères, 18 % pour les reptiles, et 1 % pour les oiseaux* (Service de l'environnement et de la planification d'Afrique du Sud 2002).

Historique

L'historique se réfère aux sources des données ainsi qu'aux traitements successifs appliqués au jeu de données pour l'amener à son état actuel. Il peut inclure la méthode de collecte (par exemple : « données collectées sur un maillage de 10 mètres sur 10 ») et des informations sur

les tests de validation qui ont été effectués. L'historique des étapes de traitement peut comprendre :

- la ou les méthode(s) de récolte des données
- tout traitement intermédiaire
- les méthodes utilisées pour générer le produit final
- toute action de validation effectuée sur les données

Par exemple :

> *Les données ont été récoltées selon des quadrats fixes de 20 mètres sur 20. Le décompte de toutes les espèces, la structure et autres données sur l'habitat, ont aussi été collectés. Les données ont été classées selon la méthode Twinspan*[12] *en groupes comprenant des groupes d'espèces similaires.*

Cohérence logique

La cohérence logique fournit une évaluation succincte des relations logiques entre les différents paramètres de la donnée. Bien que pour beaucoup de données (muséum et herbier) certains de ces paramètres puissent ne pas être pertinents, ils peuvent l'être cependant pour certaines données d'observation (check-lists des espèces dans un Parc National, ou dans une biorégion, etc.) et pour certaines données d'inventaire. Pour les données spatiales numériques, des tests de cohérence logique peuvent être effectuer automatiquement, par exemple :

- est-ce que tous les points, lignes et polygones sont étiquetés, et y a-t-il des étiquettes dupliquées ?
- est-ce que les lignes se croisent aux emplacements prévus ?
- est-ce que toutes les frontières des polygones sont fermées ?
- est-ce que tous les points, lignes ou polygones sont reliés topologiquement ?

La notion de cohérence logique peut aussi s'appliquer à des jeux de données où l'on trouve d'autres relations logiques entre paramètres et entre objets au sein du jeu de données. Dans de tels cas une description de tout test qui aurait été effectué sur ces relations devrait être incluse. Il peut s'agir par exemple de dates qui apparaissent dans différents champs : si la date donnée dans un champ indique que le projet a été conduit entre les années « a » et « b », mais que la date d'enregistrement d'un attribut dans un autre champ est en dehors de cette période, on a une incohérence logique. De même si des enregistrements sont en dehors de l'étendue géographique indiquée par ailleurs : si un champ indique qu'une donnée a été collectée au Brésil, alors qu'un autre champ présente les latitudes et longitudes d'emplacements au Paraguay, alors il y a incohérence entre ces deux champs. La documentation des vérifications effectuées est une partie importante des métadonnées. Ces vérifications peuvent inclure des tests de présence de points dans des polygones, classiques dans les SIG. Voir plus de détail sur ces méthodes dans le document associé « *Principes et méthodes de nettoyage des données* ».

[12] « Two Way Indicator Species Analysis » : il s'agit d'une technique d'analyse en agrégats (« clusters ») qui classe les sites selon les espèces que l'on y attrape en utilisant une « matrice site par espèces » (Jongman *et al.* 1995, d'après Susan King 1995 : http://www.bioscience.heacademy.ac.uk/hosted/tireragan/data/inverts.htm) (NdT)

Complétude

La notion de complétude se réfère ici à la fois à la couverture temporelle et spatiale des données ou du jeu de données, comme partie de l'étendue totale que les données peuvent concerner. La documentation de la complétude est une composante essentielle dans la détermination de la qualité. Par exemple :

- Zones au Nord de 30 ° de latitude Sud entièrement couvertes ; données ponctuelles entre 30 ° et 40 ° de latitude Sud ;
- Jeu de données couvrant seulement les enregistrements antérieurs à 1995, collectés largement de manière opportuniste, principalement en Nouvelles Galles du Sud, mais incluant quelques enregistrements d'autres Etats.

Du point de vue d'un utilisateur, la complétude se réfère à « toutes les données dont il a besoin » (English 1999). L'utilisateur a besoin de savoir si la base de données contient tous les champs nécessaires à son analyse, et il faut qu'il connaisse le degré de complétude de ces champs. Par exemple, l'utilisateur peut vouloir effectuer une étude comparative de certains attributs au cours du temps, et si la base ne contient pas de données postérieures à telle date, elle peut ne pas convenir à cette étude (voir le deuxième exemple cité ci-dessus).

Accessibilité

Pour avoir une valeur quelconque vis à vis des utilisateurs, les données doivent être accessibles. Les données ne son,t pas toutes disponibles en ligne et pour accéder à certaines données, il se peut que l'utilisateur doive contacter leur conservateur et demander la permission d'y accéder, ou obtenir une copie sur un CD de ce dont il a besoin.

La documentation sur les conditions d'accès (et d'usage) est importante pour que les utilisateurs soient en mesure d'accéder aux données : c'est donc un aspect de la qualité des données. Sur ce point, la documentation peut comprendre :

- Les adresses des contacts pour les données
- Les conditions d'accès
- La méthode d'accès (si les données sont disponibles sous forme numérique
- Le format des données
- Les précautions à prendre
- L'information sur le Copyright
- Les coûts éventuels
- Les restrictions d'usage éventuelles

Précision temporelle

La précision temporelle concerne la précision de l'information relativement au temps ; Par exemple : « *données précises seulement au mois près* ». Ceci peut avoir de l'importance dans des bases de données où il n'a pas été prévu que le champ « jour » puisse prendre une valeur « vide » et où, dans les cas où la précision au jour près n'est pas disponible, ce champ prend automatiquement la valeur « 1 » par exemple. Ceci peut donner une impression fausse sur la précision temporelle. C'est encore plus important lorsque l'information temporelle n'est connue qu'à l'année près alors que la base de données enregistre dans ce cas automatiquement la date du 1er Janvier. Si un utilisateur veut étudier par exemple la période de floraison des plantes, ou encore les schémas de migration des oiseaux, alors il doit être informé de ce

manque de précision afin d'exclure ces enregistrements puisqu'ils sont alors, au regard de cet objectif, de qualité dégradée, et pas « aptes à l'usage ».

Documentation des procédures de validation

Une des clés pour connaître l'erreur qui entache les données est la documentation des procédures de validation. Il est de peu d'intérêt pour quiconque que des vérifications de la qualité des données et des corrections soient effectuées, si ceci n'est pas documenté. C'est particulièrement important lorsque ces vérifications ont été conduites par des tiers et non par le créateur des données. Il est toujours possible que des erreurs perçues n'en fussent pas du tout, et que les changements effectués aient ajouté de nouvelles erreurs. Il est aussi important de ne pas refaire plusieurs fois les mêmes vérifications. On ne peut pas se permettre de gâcher ainsi des ressources. Par exemple, des vérifications de la qualité des données effectuées par un utilisateur peuvent avoir permis d'identifier certains enregistrements suspects. Ces enregistrements peuvent ensuite s'avérer parfaitement corrects et correspondre à de véritables valeurs exceptionnelles. Si cette information n'est pas documentée dans les enregistrements, il se peut que plus tard, quelqu'un d'autre répète les vérifications de qualité et identifie à nouveau ces mêmes enregistrements comme suspects. Cette personne peut être amenée à exclure ces enregistrements de son analyse, et en plus aura perdu un temps précieux à refaire les mêmes vérifications. Il s'agit d'un principe de base de la gestion du risque, qui devrait être appliqué systématiquement par tous les conservateurs et utilisateurs des données. On ne peut trop insister sur la valeur et sur la nécessité d'une bonne documentation. Elle aide les utilisateurs à connaître ce que recouvrent les données, quelle en est la qualité, et à quels objectifs elles sont adaptées. Elle aide aussi les gestionnaires des données à en assurer la traçabilité et à en suivre la qualité, et à ne pas gaspiller leurs ressources en vérifications répétées d'erreurs supposées.

Documentation et conception des bases de données

L'un des moyens de s'assurer que l'erreur est pleinement documentée est de prévoir la documentation dès les premières étapes de la conception et de la construction de la base de données.

Des champs additionnels dédiés à la qualité et à la précision des données peuvent alors être incorporés. Des champs concernant la précision de la position ou du géo-référencement, la source de l'information sur le géo-référencement et sur l'altitude, la personne qui a ajouté cette information. Par exemple, les coordonnées ont elles été ajoutées par le collecteur en utilisant un GPS, ou ultérieurement par un opérateur de saisie à l'aide d'une carte à un certaine échelle, l'altitude a-t-elle été générée automatiquement à l'aide d'un modèle numérique de terrain, et dans ce cas d'où provenait modèle, quelles étaient sa date et son échelle, etc. Toute cette information sera précieuse pour déterminer ensuite si les données sont adaptées ou non à un usage particulier, afin que l'utilisateur puisse en décider en connaissance de cause.

> « *les utilisateurs des données doivent prendre des précautions lorsqu'ils fondent des évaluations biologiques sur des jeux de données taxonomiques qui ne présentent pas une documentation sur au moins certaines caractéristiques de performance* »
> (Stribling *et al*. 2003)

Stockage des données

Le stockage des données peut avoir un impact sur la qualité des données, et ceci de plusieurs manières, dont beaucoup ne sont pas évidentes mais doivent être prises en compte à la fois dans la conception de la base et comme un maillon de la chaîne de qualité des données.

Le thème de la conception et du développement d'une base de données est trop vaste pour être couvert ici, et il devrait faire l'objet d'une étude séparée. Une étude commandée par le GBIF a examiné les logiciels de gestion des collections (Berendsohn *et al.* 2003) et je recommande au lecteur de s'y référer.

Cette section traite de certains des grands principes du stockage des données en relation avec leur qualité.

Sauvegarde des données

La sauvegarde régulière des données permet d'assurer la cohérence des niveaux de qualité. Il est important que les organismes en charge de données mettent en place des procédures de sauvegarde et de récupération en cas de pannes subites des supports électroniques. A chaque fois que des données sont perdues ou corrompues, il y a une perte concomitante en qualité.

Archivage

L'archivage des données (y compris la suppression des données obsolètes) est un domaine de la gestion des données et du risque qui requiert une attention particulière. L'archivage des données, particulièrement dans les universités, les ONGs et chez les particuliers devrait être une priorité de la gestion des données. Les universités ont un taux important de mouvement de personnel et les données de la recherche sont souvent stockées de manière distribuée – habituellement dans les ordinateurs personnels ou les armoires des chercheurs. Si elles ne sont pas solidement documentées, de telles données peuvent devenir très rapidement inutilisables et inaccessibles. Plus souvent qu'à leur tour, ces données sont supprimées ou jetées après que le chercheur a quitté l'organisme, car personne ne sait de quoi il s'agit ou ne fait l'effort de les maintenir. C'est pourquoi les universités particulièrement ont besoin de solides stratégies en termes de documentation et d'archivage.

Les chercheurs isolés qui travaillent en dehors d'une grande institution doivent s'assurer que leur données sont maintenues et/ou archivées après leur décès, ou après qu'il ont cessé de s'y intéresser.

De la même manière les ONGs qui peuvent ne pas avoir de financement à long terme pour le stockage de leurs données, ont besoin de passer des accord avec les organismes appropriés qui disposent d'une stratégie de gestion des données à long terme (y compris en matière d'archivage) et qui sont susceptibles de s'intéresser à leurs données.

L'archivage des données est devenu plus aisé ces dernières années avec le développement des protocoles GIGIR/Darwin Core et BioCASE/ABCD[13]. Ceux-ci fournissent une manière

[13] http://www.tdwg.org; http://www.gbif.org/links/standards

simple pour une institution, un Département d'université ou un individu, d'exporter leur base de données sous l'un de ces formats et de les stocker au format XML, soit sur leur propre site, soit en les transmettant à une institution hôte. C'est une manière simple de stocker durablement les données et/ou de les rendre accessibles via des procédures de recherche distribuée comme dans le portail de données du GBIF.

Le nettoyage, la suppression et l'archivage des données concernent aussi les sites Web qui sont abandonnés par leurs créateurs, ou qui contiennent des données anciennes et obsolètes, et qui jonchent le cyberespace de débris numériques. Les organismes ont besoin d'une stratégie d'archivage des données intégrée dans leur chaîne de gestion de l'information. L'archivage physique des données est un sujet trop vaste pour être traité ici, toutefois un document récent sur l'archivage des données à l'aide de CDs et DVDs a été publié par le Conseil sur l'information et les ressources des bibliothèques, et par l'Institut national des normes et de la technologie des Etats-Unis (Byers 2003). Il constitue une synthèse précieuse sur cette technologie et le lecteur peut souhaiter s'y référer.

On ne devrait pas détruire ou mettre en danger les données qui ne sont plus requises (pour des raisons légales ou autres), sans explorer toutes les autres possibilités – y compris l'archivage (NLWRA[1] 2003).

[1] « National Land and Water Resources Audit », programme australien d'évaluation des ressources naturelles. conduit entre 1997 et 2008.

Intégrité des données

L' intégrité des données se réfère au fait que les données n'ont pas été altérées ou détruites de manière non autorisée, accidentellement ou par malveillance (du fait d'une saute de courant ou d'un virus).

Les données changent souvent – par exemple lorsque l'information taxonomique d'un enregistrement est mise à jour à la suite d'une nouvelle détermination – mais les utilisateurs s'attendent à ce que le système informatique maintienne l'intégrité des données et ne provoque pas lui-même une modification inopinée et incorrecte d'une valeur. La corruption des données se produit lorsque le maintien de l'intégrité n'est pas assuré et qu'un changement inopiné et incorrect intervient.

Gestion, stockage, sauvegarde et archivage doivent être de qualité pour préserver l'intégrité des données

Typologie des erreurs

Les bases de données taxonomiques ou sur les occurrences d'espèces, comme toutes les bases de données, sont sujettes à certains types d'erreur. English (1999) a identifié les erreurs typiques suivantes, qu'il a appelées défauts des données. Dalcin (2004) les a adoptés dans le domaine des bases de données taxonomiques.

Les valeurs citées ici sont de English (1999à avec des exemples cités de Chapman (1991) et de bases de données de l'Herbier virtuel australien[14] et du système brésilien « speciesLink[15] » :

- **Redondance des domaines de valeurs** : des valeurs non normalisées ou synonymes existent et deux ou plusieurs valeurs ou codes ont la même signification. La redondance est typique des données descriptives lorsque les terminologies normalisées ne sont pas respectées, ou quand la compilation de données issues de plusieurs sources est mal contrôlée.
- **Valeurs manquantes** : un champ de données n'est pas renseigné. Qu'il ait été ou non prévu de saisir ce champ, celui-ci est nécessaire aux traitements ultérieurs. Il peut par exemple s'agir de coordonnées géographiques.
- **Valeurs incorrectes** : les causes peuvent être diverses : mauvaise frappe sur le clavier, erreur de champ, mauvaise compréhension de la signification des données saisies, impossibilité de lire ce qui est écrit sur l'étiquette, ou champ dont la saisie est obligatoire alors que l'opérateur ne sait pas quoi y mettre. Les valeurs incorrectes sont les erreurs les plus évidentes et communes, et elles peuvent affecter toute valeur dans tout champ. L'orthographe erronée d'un nom scientifique est une erreur répandue dans les bases de données taxonomiques et nomenclaturales, de même que l'inscription d'un zéro dans les champs de géo-référencement.
- **Valeurs composées** : c'est le cas lorsqu'on inscrit plus d'un fait dans le même champ (par exemple genre, espèce et auteur ; ou bien rang et nom infra-spécifique). Ce type d'erreur résulte habituellement d'une mauvaise conception de la base de données ; il peut causer de réelles difficultés dans l'intégration des données.

Genre	Espèce	Infra-espèce
Eucalyptus	globulus	subsp. bicostata
Famille	**Espèce**	
Myrtaceae	Eucalyptus globulus Labill.	

Tableau 4. *Exemples de valeurs composées.*

- **Schizophrénie du domaine** : cela se produit lorsqu'on inscrit dans un champ des informations de nature différente et/ou pour lesquelles le champ n'avait pas été conçu

Famille	Genre	Espèce
Myrtaceae	Eucalyptus	globulus?
Myrtaceae	Eucalyptus	? globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?
Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	à déterminer

Tableau 5. *Exemples de schizophrénie du domaine*

- **Occurrences dupliquées** : c'est le cas lorsque plusieurs enregistrements représentent une seule entité. Cela se produit typiquement avec les noms et les variantes orthographiques ou les variantes nomenclaturales valides. Celles-ci peuvent entraîner

[14] http://www.cpbr.gov.au/avh/
[15] http://specieslink.cria.org.br/

des difficultés pour les utilisateurs lorsque ceux-ci recherchent un nom, ou lorsqu'ils essaient de combiner des données issues de plusieurs bases de données. Exemples :

- Phaius tancarvilleae
- Phaius tankervilliae
- Phaius tankarvilleae
- Phaius tankervilleae
- Phaius tankervillae
- Brassicaceae/Cruciferae (équivalents exacts ; tous deux sont autorisés par le Code de Botanique).

- **Valeurs incohérentes** : cela se produit lorsque des données, dans des bases de données reliées, sont mises à jour de manière incohérente ou à des moments différents dans les deux bases. Par exemple, entre les bases de la collection vivante et de l'herbier, ou entre la base de la collection de muséum et celle des images correspondantes.
- **Contamination de la qualité de l'information** : elle résulte de la combinaison de données exactes et précises avec des données inexactes ou imprécises. Par exemple, la combinaison des données renseignées au niveau sous-espèce avec d'autres issues d'une base qui ne contient qu'une information précise au niveau espèce.

Données spatiales

Le stockage de données spatiales concerne l'information textuelle sur la localisation aussi bien que les coordonnées géographiques. De nombreuses bases de données commencent à inclure diverses données de localisation normalisées, comme la distance et la direction du toponyme le plus proche, en plus de la description en texte libre. Plusieurs projets en cours visent à améliorer le codage normalisé des données de localisation en texte libre afin de créer ces divers champs qui aident au processus du géo-référencement. Le projet BioGeomancer[16] financé récemment par la Fondation Gordon et Betty Moore est l'un de ces projets.

Le géo-référencement (ou coordonnées géographiques) est généralement saisi dans les bases de données sous forme de latitude et de longitude (système de coordonnées sphériques) ou de coordonnées UTM (ou similaire) (système de coordonnées planimétriques). Un système de coordonnées sphériques comme les latitude et longitude encercle le globe, et doivent être projetées sur un plan afin d'être représentées sur une carte papier. Les systèmes de coordonnées sphériques ne sont pas homogènes en termes d'aire, et la distance correspondant à un degré de latitude, par exemple, peut varier considérablement selon que l'on est proche d'un pôle ou de l'équateur. Les systèmes de coordonnées planimétriques sont plus homogènes en termes d'aire, et peuvent être utilisés pour mesurer ou calculer des surfaces.

De nombreuses institutions commencent maintenant à saisir les coordonnées en degrés, minutes et secondes, ou en degrés et « minutes décimales » (comme ce qu'indiquent les GPS), et à laisser le soin aux systèmes de conversion intégrés dans les bases de données transformer ces coordonnées en degrés décimaux pour le stockage. Pour transférer ces coordonnées et les utiliser dans un SIG, il est généralement mieux de les stocker sous forme de degrés décimaux qui bénéficient de dispositifs de transfert d'utilisation aisée et offrent la plus haute précision possible.

[16] http://www.biogeomancer.org/

Le stockage de coordonnées au format UTM est souvent utilisé dans des institutions qui ne travaillent que sur une zone UTM donnée. Il a l'avantage d'être homogène en termes d'aire comme indiqué plus haut, c'est-à-dire que chaque maille est un carré (ou un rectangle) et permet une représentation aisée sur une carte plate, ou facilite le calcul de distances et de surfaces. Il est cependant important dans ce cas de mémoriser aussi de quelle zone UTM il s'agit, sinon il peut être difficile de combiner les données avec d'autres provenant d'autres zones ou institutions.

Degrés décimaux

Le stockage des coordonnées sous forme de degrés décimaux, utilisé dans de nombreuses bases, peut conduire à une fausse précision, comme mentionné plus haut. Il faut se préoccuper de la précision avec laquelle les données sont stockées ou fournies. La base de données ne devrait pas permettre d'afficher une précision supérieure à celle de la donnée la plus précise dans la base. Dans le cas des données biologiques, ceci sera de l'ordre de 10 mètres.

Systèmes géodésiques

Il y a plusieurs systèmes géodésiques. La terre n'est pas parfaitement sphérique ; c'est un ellipsoïde, et l'adaptation d'un système de coordonnées à la surface de cet ellipsoïde n'est pas évidente (Chapman *et al*. 2005). C'est pour résoudre cette difficulté que le concept de système géodésique a été imaginé. Il s'agit d'un ensemble de points qui sont utilisés pour repérer la position d'un point de la sphère sur l'ellipsoïde. Historiquement, plusieurs systèmes de référence différents ont été générés pour différentes parties du monde, et ce n'est qu'avec l'avènement des satellites qu'un système géodésique de référence vraiment global a pu être mis au point, lorsque les satellites ont permis de positionner précisément le centre de la Terre. La différence entre les coordonnées d'une position donnée dans différents systèmes géodésiques peut conduire à des écarts de 400 mètres et plus (Wieczorek 2001).

Du fait de cette différence, il est important de noter dans les bases de données le système géodésique utilisé ; sans quoi les combinaisons de données pourraient être entachées de fortes erreurs.

Manipulation des données spatiales

Il y a de nombreuses manières de manipuler les données spatiales. Beaucoup n'ont pas d'effet sur leur précision, mais certaines en ont un. Des exemples de méthodes qui affectent la précision des données de localisation sont développés ci-dessous.

Conversion des données d'un format à un autre

Les conversions de données qui sont peut-être les plus communément effectuées par ceux qui gèrent et utilisent les données de collections et d'occurrences d'espèces sont les conversions des coordonnées géographiques, depuis des couples –longitude, latitude- en degrés/minutes/secondes ou des coordonnées UTM, vers des degrés décimaux. Il y a aussi des conversions de miles en kilomètres ou de pieds en mètres, et d'autres encore.

Toutes ces conversions sont relativement simples, mais elles peuvent conduire à une fausse impression de la précision de la localisation, du fait de la mauvaise utilisation de la précision du codage des données. Par exemple un spécimen pour lequel l'altitude indiquée est de 250 pieds (ce qui pour le collecteur pouvait signifier qu'elle était entre 200 et 300 pieds) donnera en mètres la valeur de 76,2 (avec une décimale) ou peut-être de 76 si elle est arrondie. Il vaudrait mieux enregistrer la valeur convertie comme étant 80 mètres, et même encore mieux noter que l'altitude est de 80 mètres ± 20 mètres. Une fausse précision est en fait une perte de qualité.

Systèmes géodésiques et projections

La conversion de données d'un système géodésique à un autre peut conduire à des erreurs importantes du fait qu'elle n'est pas uniforme (voir Wieczorek 2001 pour une analyse des systèmes géodésiques et de leurs impacts sur la qualité des données). De nombreux pays ou régions convertissent maintenant la plupart de leurs données vers un système géodésique de référence pour leur région. Soit le système géodésique mondial (en anglais « World Geodetic System » : WGS84) ; soit un système qui en est une bonne approximation locale : par exemple, le système géodésique australien (en anglais « Australian Geodetic Datum » : AGD84), qui en Australie varie par rapport au WGS84 d'environ 10 cm ; ou en Europe le EUREF89, qui peut s'écarter du WGS84 d'environ 20 cm. La conversion d'un système à un autre n'est pas toujours nécessaire : elle ne l'est pas, par exemple, si la précision des données n'est en fait que de 5 à 10 km. En revanche, si l'on a affaire à des données d'une précision supérieure, par exemple, 10 à 100 mètres, les différences entre systèmes peut être significative : dans certaines régions du monde, elle peut en effet atteindre jusqu'à 400 mètres (Wieczorek 2001).

De manière analogue, lorsque les données sont localisées à l'aide de polygones (par exemple, les données recueillies dans un parc national), il faut être conscient des erreurs qui peuvent résulter de la conversion d'une projection à une autre. Des formules types sont disponibles pour calculer l'erreur résultant de telles conversions, et les métadonnées qui accompagnent les données devraient refléter cette information.

Maillages

Lorsque des données vectorisées sont converties en pixels ou en maillage, on perd en exactitude et en précision. Ceci vient de la taille des mailles ou des pixels utilisés pour approximer les valeurs vectorisées (Burrough et McDonnel 1998). Précision et exactitude ne peuvent être retrouvées en convertissant à nouveau les données en vecteurs. Pour une analyse plus poussée des problèmes rencontrés en utilisant et en convertissant des données pixélisées, et des problèmes d'échelle, voir Chapman *et al.* (2004).

Intégration des données

Les jeux de données spatialisés sont difficiles à intégrer lorsqu'il y a des incohérences entre eux. Ces incohérences peuvent concerner autant les aspects géographiques que les autres, et elles peuvent nécessiter la mise en oeuvre de mesures correctives diverses et souvent fastidieuses (Shepherd 1991). Ces incohérences peuvent résulter de :

- Différences entre les techniques de mesure et d'enregistrement (concernant par exemple la taille de la zone ou la périodicité des observations), entre les méthodes d'inventaire (concernant par exemple la taille des mailles, ou la largeur des transects) ou entre les catégories de données (par exemple utilisation de définitions différentes pour les catégories)
- Erreurs dans les mesures ou dans les méthodes d'inventaire (erreurs de transcription, de saisie, d'identification)
- Différences de résolution (spatiale, temporelle ou d'autres attributs)
- Définitions vagues et imprécises
- Flou des objets observés (dans les catégories de sol ou de végétation, dans les identifications lorsque certaines sont effectuées au niveau espèce, et d'autres jusqu'au niveau sous-espèce ou seulement au niveau genre)
- Différences d'utilisation ou d'interprétation de terminologie ou de nomenclature (différentes taxonomies par exemple)
- Différences dans les paramétrages des GPS (système géodésique, système de coordonnées, etc.)

De tels problèmes d'intégration sont accrus lorsque les données sont :

- De types différents (par exemple des données de collections mélangées à des données d'inventaire ou d'observation)
- Issues de différentes juridictions (par exemple lorsque les méthodes d »inventaire pratiquées ne sont pas les mêmes)
- Obtenues de sources multiples
- A des échelles multiples
- De natures différentes (cartes, spécimens, images, etc.)
- Recueillies à des époques différentes
- Stockées dans des SGBD[17] différents ou avec des modèles de données différents, ou sur des supports différents (par exemple, certains SGBD ne prévoient les valeurs « vides »)

[17] Système de Gestion de Base de Données (Access, Oracle, MySQL, etc.) (NdT)

- Décomposées différemment : par exemple lorsqu'une base de données enregistre le nom scientifique entier dans un seul champ, alors qu'une autre l'éclate entre des champs différents pour le genre, l'épithète d'espèce, etc.)

L'intégration de données donne de meilleurs résultats lorsque les gestionnaires des données ont utilisé des méthodes de stockage des données normalisées et cohérentes.

Représentation et Présentation

> *Des méthodes devraient toujours être développées afin de faire l'usage le plus efficace des données existantes, quelle que soit leur qualité. Toutefois, pour que les données soient fiables, elles doivent aussi être validées ou accompagnées d'une information qui indique leur niveau de fiabilité* (Olivieri et al. 1995)

Les scientifiques et les institutions scientifiques sont de mieux en mieux reconnues comme fournisseurs de données pour comprendre, expliquer, quantifier et évaluer la biodiversité. Cette reconnaissance repose sur leur capacité à fournir une information fiable et utilisable par les décideurs, les gestionnaires, le grand public, et autres. Une information ambiguë, confuse, incomplète, contradictoire ou erronée, comme pourrait en produire une base de données mal gérée, peut affecter leur réputation comme fournisseurs de données et autorités scientifiques (Dalcin 2004).

Un objectif clé de la manipulation de données numériques en sciences biologiques est de fournir aux utilisateurs une méthode efficace d'interrogation et d'analyse de l'information. Pour que cela fonctionne encore faut-il que les données représentent correctement le monde biologique. Mais le monde biologique est infiniment complexe, et il doit être généralisé, approximé et abstrait pour être représenté et compris (Goodchild *et al.* 1991). Ceci peut se faire à m'aide de SIG, de modèles environnementaux et de systèmes d'aide à la décision. Avec ces outils, il est cependant essentiel que toute variation soit échantillonnée et mesurée, et que l'erreur et l'incertitude soit décrite et visualisée. Et nous sommes encore loin, dans ce domaine, d'avoir atteint un excellent niveau de savoir faire.

La biologie est l'une des premières disciplines à avoir développé des techniques de production de rapports d'erreurs, avec une visualisation graphique des erreurs (barres d'erreur, en anglais « error bars ») et diverses mesures et estimations statistiques. La production de rapports d'erreurs n'était pas perçue comme une faiblesse car l'estimation des erreurs fournit une information cruciale pour interpréter correctement les données (Chrisman 1991). Dans la fourniture de données sur les espèces, il faut développer et mettre en œuvre les mêmes techniques afin que les utilisateurs puissent eux aussi interpréter et exploiter correctement les données.

Des programmes de qualité des données efficaces aident à éviter des situations embarrassantes pour les organismes comme pour les individus – à la fois en interne et en public.

Déterminer les besoins de utilisateurs

La détermination des besoins des utilisateurs n'est pas chose facile, et il est aussi difficile d'élaborer des spécifications détaillées que de concevoir ensuite la structure de données adaptée à ces spécifications. Il est en revanche important d'identifier des utilisateurs clés avec qui travailler sur l'analyse des besoins. De bonnes spécifications peuvent conduire à une amélioration globale de la collection, de sa gestion et de la qualité d'ensemble des données.

Pertinence

La notion de pertinence est ici étroitement liée à celle de « qualité » : il s'agit de la pertinence des données relativement à ce que l'on veut en faire. Les problèmes de pertinence peuvent être liés à quelque chose de simple, comme par exemple lorsqu'on essaie d'utiliser une Flore dans une zone qu'elle ne couvre pas mais pour laquelle rien d'autre n'est disponible, ou lorsque des données sont disponibles dans une projection différente de celle qui est requise et qu'elles nécessitent un travail considérable pour être utilisables et « pertinentes ».

Crédibilité

La crédibilité correspond au degré de confiance de l'utilisateur dans les données (Dalcin 2004). Elle dépend de la perception ou de l'évaluation de l'utilisateur quant au fait que les données soient conformes à son objectif et elle peut être fondée sur l'expérience passée ou la comparaison à des normes communément acceptées (Pipino *et al*. 2002). La réputation d'un jeu de données dépend parfois de la crédibilité (et de l'adéquation à l'usage) perçue par les utilisateurs, mais elle peut souvent être améliorée par une bonne documentation.

L'article de Wang *et al.* (1995) inclut un diagramme qui met en relation plusieurs de ces sujets dans une représentation hiérarchique, et il montre les liens entre des notions comme la crédibilité et la réputation.

Accepter l'incertitude dans les données spatiales

L'incertitude, particulièrement en ce qui concerne les données spatiales, est inévitable ; mais elle est souvent mal documentée, et elle n'est pas toujours évidente pour les utilisateurs. La prolifération des systèmes de cartographie bureautiques faciles à utiliser a permis à des gens qui ne sont pas des professionnels des SIG de visualiser et d'analyser aisément les relations spatiales entre leurs données, mais ils utilisent souvent des échelles inappropriées (Chapman *et al*. 2005), et ne tiennent pas compte de l'erreur et de l'incertitude spatiales intrinsèques des données (Chapman 1999). Dans certains cas ceci peut entraîner une utilisation dangereusement erronée des données, avec parfois des conséquences tragiques (Redman 2001). Ces dernières années, les services de cartographie en ligne, simples à utiliser, se sont multipliés. Ils permettent de visualiser et analyser les données spatiales comme dans un SIG bureautique traditionnel, mais les éditeurs contrôlent les couches de données et l'échelle des jeux de données disponibles. Dans un proche avenir cela va s'étendre encore avec le développement de service de cartographie sur le Web (en anglais, « Web Mapping Services », ou WMS). Le contrôle des couches de données et de l'échelle par les éditeurs de la carte (par exemple, le fait d'autoriser différentes couches à être automatiquement incluses ou exclues lorsque l'utilisateur fait un zoom avant) réduit la fréquence de certaines erreurs simples de manipulation.

Il est essentiel de documenter l'incertitude, d'abord en utilisant de donnes métadonnées, et ensuite grâce à la visualisation et la présentation. L'un des domaines de recherche où l'effort doit être poursuivi en ce qui concerne les données sur les espèces ou les occurrences d'espèces, et celui du développement des techniques pour visualiser l'incertitude – par exemple pour afficher des marqueurs de la précision. Au lieu de représenter l'enregistrement d'un spécimen comme un point (un couple longitude – latitude), il faut inclure la précision associée à l'enregistrement en présentant la localisation sous la forme d'une zone (cercle, ellipse, etc.), et peut-être même inclure aussi des niveaux de probabilité (Chapman 2002).

Il est important que ceux qui connaissent les données et leurs limitations en ce qui concerne la précision de la localisation ou d'autres attributs, aident les utilisateurs en documentant les données et en rendant cette information disponible afin de les guider dans leur évaluation de l'aptitude à l'usage des données selon leurs objectifs.

Visualisation de l'erreur et de l'incertitude

Il y a encore beaucoup à faire pour obtenir de bonnes méthodes de visualisation de l'erreur dans le domaine des données sur les espèces, bien que nombre de méthodes nouvelles et prometteuses soient en cours de mise au point (par exemple, Zhang et Goodchild 2002). La voie peut-être la plus aisée concerne l'utilisation d'une couche « erreur » intégrée comme une couche supplémentaire dans les SIG. De telles techniques ont été utilisées dans le monde de la cartographie, où une couche peut être traduite visuellement par une coloration d'intensité variable qui reflète la fiabilité relative des différentes parties de la carte. D'autres techniques pourraient impliquer des symboles (une ligne de points, une ligne continue, des points de taille et d'intensité variables, etc. qui indiquent des données de plus ou moins grande qualité ou précision). La mise en œuvre de telles couches peut aussi souvent donner une idée de l'origine des erreurs et peut donc être un outil précieux pour la validation et la vérification des données.

L'utilisation d'une matrice pour afficher les classifications erronées avec résultats attendus sur les lignes, et les résultats observés sur les colonnes, est intéressante lorsque ces calculs statistiques sont possibles. Dans ces cas là, les erreurs dans les lignes sont des erreurs par omission et les erreurs dans les colonnes sont des erreurs commises (Christman 1991). De telles méthodes ne sont généralement pas applicables aux données d'occurrence d'espèces, mais elles peuvent être utiles, par exemple dans le cas de données d'inventaire où l'absence/présence est observée sur une longue période.

Evaluation du risque

Les décideurs préfèreraient un climat de certitude, mais les systèmes naturels sont intrinsèquement variables et se conforment rarement à ce souhait. Les techniques d'évaluation du risque fournissent de plus en pus aux décideurs et aux gestionnaires de l'environnement des estimations de l'incertitude et du risque, de sorte que les décisions en matière environnementale peuvent être prises sur des bases plus solides. Dans le cas des espèces, où la connaissance exacte de leur localisation est souvent parcellaire, des zones de « présence vraisemblable » peuvent être utilisées faute de mieux. Toutefois, à l'intérieur de larges zones de « présence vraisemblable », il peut y avoir des endroits plus « vraisemblables » que d'autres (Chapman 2002).

On peut généralement distinguer deux éléments dans le concept de risque : la vraisemblance et l'amplitude d'un événement, et ses conséquences si celui-ci a lieu (Beer et Ziolkowski 1995). Dans le contexte des données sur les espèces, l'évaluation du risque peut aller du risque d'un incendie qui détruiraient les données si une procédure de sauvegarde extérieure n'a pas été mise en place, jusqu'au risque lié à une décision environnementale erronée due à des données de mauvaise qualité. On peut donner comme exemple le coût d'interdire un projet d'aménagement du fait d'une information erronée sur la présence d'espèces menacées dans la zone. Dans certaines situations environnementales, les gouvernements cherchent de

plus en plus à appliquer le *principe de précaution* lorsqu'ils prennent des décisions importantes.

Responsabilités légales et morales

Il y a de nombreux domaines où des responsabilités légales et morales peuvent être engagées en ce qui concerne la qualité et la présentation des données sur les espèces. On peut citer :

- Les droits de copyright et de propriété intellectuelle
- Le respect de la vie privée
- La vérité dans l'étiquetage
- La restriction de la qualité de la présentation pour les taxons sensibles
- Les droits des communautés autochtones
- La fiabilité
- Les précautions et les mises en garde

Dans la plupart des cas, les ***droits de copyright et de propriété intellectuelle*** sur les données peuvent être couverts par la documentation qui accompagne les données. Lorsque ceux-ci sont susceptibles de varier d'un enregistrement à l'autre, ils devraient être notés au niveau de l'enregistrement, sinon ils peuvent être couverts par les métadonnées.

Un certain nombre de pays ont récemment introduit une législation sur le respect de la vie privée et les gestionnaires de données devraient la prendre en compte. Ceci est particulièrement pertinent lorsque les données traversent les frontières ou sont accessibles sur Internet. Dans certains pays, les données sur les individus ne peuvent pas être stockées dans une base de données et mises à disposition sans autorisation express. Il n'est pas toujours évident de déterminer comment cette législation affecte l'information associée aux données sur les espèces, mais les conservateurs devraient en avoir conscience et prendre les dispositions éventuellement nécessaires.

De bonnes mesures de contrôle qualité ainsi que de bonnes métadonnées permettent en général de respecter les concepts de « ***vérité de l'étiquetage*** ». Jusqu'ici, dans la législation au moins, la « vérité sur l'étiquetage » a été plutôt restreinte aux produits alimentaires. Elle est cependant mentionnée dans des articles traitant du développement d'une infrastructure mondiale pour les données spatiales (Nebert et Lance 2001, Lance 2001), d'une infrastructure nationale pour les données spatiales aux Etats-Unis (Nebert 1999) et d'une infrastructure de l'Australie et de la Nouvelle-Zélande pour les données spatiales (ANZLIC 1996b). L'article sur l'infrastructure mondiale pour les données spatiales (Lance 2001) recommande qu'une structure de gestion des données spatiale soit dotée « *d'une méthode de libre publication pour fournir un accès mondial aux données détenues selon le principe de la vérité de l'étiquetage* », et, selon le document de l'Australie et de la Nouvelle-Zélande :

> « *les normes de qualité des données géographiques et sur les territoires peuvent être descriptives, prescriptives, ou les deux. Une norme descriptive est fondée sur le concept de 'vérité de l'étiquetage', demandant aux producteurs de données d'informer sur la qualité de leurs données. Ce qui permet aux utilisateurs d'avoir un jugement éclairé sur 'la pertinence selon l'objectif' des données.* »

La ***restriction de la qualité de la présentation pour les espèces sensibles*** peut être réalisée avec une information « floue » sur la localisation : par exemple pour limiter la connaissance de la position exacte d'une espèce menacée, d'une espèce commercialisée sensible, etc. C'est

une réduction de la qualité publiée des données, et dans ce cas cette restriction doit être clairement documentée afin que les utilisateurs sachent ce qu'ils obtiennent, et puissent décider si les données leur sont alors utiles ou non.

Le respect des ***droits des communautés autochtones*** peuvent aussi affecter la qualité des données, lorsque certaines informations sont considérées comme sensibles par ces communautés. Cette restriction et sa raison doivent alors être documentées.

En 1998, Epstein *et al.* a étudié le problème de la responsabilité légale en relation avec l'usage de l'information spatiale. Voici quelques-uns des points clés soulevés :

- *L'erreur dans l'information spatiale recèle maintenant un '**potentiel considérable**' de litige et de perte de réputation et d'intégrité pour les individus comme pour les organismes.*
- *Les clauses traditionnelles limitant la responsabilité peuvent ne pas suffire en cas de litige.*
- *Afin de limiter leur responsabilité, les organismes devraient maintenir un haut niveau de documentation sur la qualité, qui certifie de manière appropriée et honnête que leurs produits sont conformes à ce qu'ils peuvent faire de mieux '**selon leur compétence et leur connaissance**'.*

Avertissements et clauses de non responsabilité sont une partie importante de la documentation de la qualité des données. Iles devraient être rédigés de telle manière que non seulement ils couvrent l'organisme qui héberge les données, mais aussi ils fournissent à l'utilisateur une idée de leur qualité et de ce qu'il peut en attendre.

La plupart des agences et des équipes impliquées dans la production des données seront jugées sur la facilité avec laquelle données et information seront accessibles, et sur la qualité de l'information. Ceux qui sont capables de publier, partager, accéder, intégrer et utiliser l'information sont les plus grands bénéficiaires (NLWRA 2003).

Certification et accréditation

Est-ce que les données d'occurrence d'espèces peuvent et doivent être certifiées ? Avec l'accroissement du volume de données disponibles auprès de nombreuses agences, les utilisateurs veulent savoir lesquelles sont fiables, et lesquelles suivent des procédures de contrôle qualité documentées. Devraient-ils ses reposer seulement sur les institutions les plus connues, ou y en a-t-il de moins connues qui ont aussi des données fiables ? Quelles données obtenues auprès d'institutions bien connues sont fiables, et lesquelles ne le sont pas ? La *réputation* seule peut être le facteur décisif quant au choix des sources de données, mais la réputation est un concept subjectif et une caractéristique fragile pour fonder actions et décisions (Dalcin 2004). Est-ce cela que nous voulons dans notre discipline ? De bonnes métadonnées et une bonne documentation sur les procédures du contrôle qualité peuvent souvent transformer un facteur subjectif comme la réputation en quelque chose sur lequel les utilisateurs peuvent fonder une évaluation plus scientifique et rationnelle. Nous devrions peut-être développer un processus de certification et d'accréditation qui informe les

utilisateurs sur les organismes qui respectent un niveau minimum de normes et de procédures de documentation du contrôle qualité.

Le développement d'une certification consensuelle de qualité pourrait entraîner une amélioration dans la qualité globale des données et une confiance accrue des utilisateurs dans la valeur des données. Et cela pourrait améliorer le financement des organismes certifiés. Dalcin (2004) suggère *« qu'une certification de qualité des données taxonomiques pourrait impliquer trois aspects : les sources de données primaires (la matière première), la chaîne de l'information (le processus) et la base de données (le produit). »*

Revue par les pairs des bases de données

Un système de revue par les pairs des bases de données pourrait être introduit dans le domaine des données sur les espèces. Un tel processus de revue par les pairs pourrait s'insérer dans une procédure de certification comme analysé plus haut, et peut couvrir les aspects des procédures de contrôle qualité, la documentation et les métadonnées, les mécanismes de mise à jour et de retour d'information, etc.

Conclusion

> *Un des buts de tout spécialiste de l'information est d'éviter les erreurs inutiles. En reconnaissant directement les erreurs, il est possible de les limiter à un taux acceptable. Il n'en reste pas moins que les erreurs ne peuvent pas toujours être évitées facilement et sans coûts élevés.*
>
> (Christman 1991).

On ne peut trop insister sur l'importance de la qualité des données et de la vérification des erreurs. Comme cela a été répété tout au long de ce document, il est essentiel que les données aient une réelle valeur pour la production de résultats qui améliorent les décisions et la gestion de l'environnement. La qualité est un aspect important de tout données, qu'elle provienne de collections de muséums ou d'herbiers, d'observations, d'inventaires, ou de check lists. Il y a une exigence commune de nombreux gouvernements pour avoir des données de grande qualité et mieux documentées. Par exemple :

- Les gouvernements australiens des Etats, du Territoire du Nord et fédéral encouragent fermement à améliorer les services et à faire un usage plus efficace des ressources, y compris des sources de données et d'information.
- Il y a une reconnaissance croissante de ce que les données collectées aux frais du contribuable doivent être gérées correctement afin de les rendre accessibles à tous publics pour en tirer le meilleur profit et justifier les coûts considérables de leur production et de leur maintenance.
- Il y a une pression croissante de la part des utilisateurs en faveur d'un accès plus aisé et plus rapide aux données et informations pertinentes à un coût réduit ou nul.
- Il y a une attention accrue des gouvernements sur le besoin de rationaliser et de combiner les données afin d'améliorer l'efficacité et la valeur ajoutée.
- Il y a une exigence croissante quant à la pertinence des données. Ceci concerne autant les nouvelles collectes et les nouveaux inventaires que la gestion et la publication des données.

Le besoin de données de qualité est incontestable, mais de nombreux gestionnaires de données supposent que les données qu'il hébergent dans leur système sont parfaites et entachées d'aucune erreur – ou que ces erreurs ne sont pas importantes. Mais erreur et incertitude sont inhérentes à toute donnée, et toute erreur affecte les usages finaux que l'on fait des données. Les processus d'acquisition et de gestion des données susceptibles d'améliorer leur qualité sont une partie essentielle de leur gestion. Toutes les étapes de la chaîne de qualité de l'information doivent être examinées et améliorées par les organismes en charge des données d'occurrence d'espèces, et leur documentation est un élément clé pour que les utilisateurs puissent connaître et comprendre les données et déterminer leur « aptitude à l'usage », et donc leur qualité.

> *Le facteur humain est potentiellement la plus grande menace qui pèse sur la précision et la fiabilité de l'information spatiale. C'est aussi le facteur qui peut assurer à la fois la fiabilité et la compréhension des faiblesses inhérentes à tout jeu de données spatialisé* (Bannerman 1999).

Remerciements

Texte

Références bibliographiques

Agumya, A. and Hunter, G.J. 1996. Assessing Fitness for Use of Spatial Information: Information Utilisation and Decision Uncertainty. *Proceedings of the GIS/LIS '96 Conference*, Denver, Colorado, pp. 359-70

ANZLIC. 1996a. *ANZLIC Guidelines: Core Metadata Elements Version 1, Metadata for high level land and geographic data directories in Australia and New Zealand*. ANZLIC Working Group on Metadata, Australia and New Zealand Land Information Council. http://www.anzlic.org.au/metaelem.htm. [Accessed 14 Jul 2004]

ANZLIC 1996b *Spatial Data Infrastructure for Australia and New Zealand. Discussion Paper.* www.anzlic.org.au/get/2374268456. [Accessed 1 Jul 2004].

Armstrong, J.A. 1992. The funding base for Australian biological collections. *Australian Biologist* 5(1): 80-88.

Bannerman, B.S., 1999. *Positional Accuracy, Error and Uncertainty in Spatial Information.* Australia: Geoinovations Pty Ltd. http://www.geoinnovations.com.au/posacc/patoc.htm [Accessed 14 Jul 2004].

Beer, T. & Ziolkowski, F. (1995). *Environmental risk assessment: an Australian perspective.* Supervising Scientist Report 102. Canberra: Commonwealth of Australia. http://www.deh.gov.au/ssd/publications/ssr/102.html [Accessed 14 Jul 2004]

Berendsohn, W.G. 1997. A taxonomic information model for botanical databases: the IOPI model. *Taxon* 46: 283-309.

Berendsohn, W., Güntsch, A. and Röpert, D. (2003). Survey of existing publicly distributed collection management and data capture software solutions used by the world's natural history collections. Copenhagen, Denmark: Global Biodiversity Information Facility. http://circa.gbif.net/Members/irc/gbif/digit/library?l=/digitization_collections/contract_2003_report/ [Accessed 16 Mar. 2005].

Birds Australia. 2001. *Atlas of Australian Birds. Search Methods*. Melbourne: Birds Australia. http://www.birdsaustralia.com.au/atlas/search.html [Accessed 30 Jun 2004].

Birds Australia. 2003. *Integrating Biodiversity into Regional Planning – The Wimmera Catchment Management Authority Pilot Project*. Canberra Environment Australia. http://www.deh.gov.au/biodiversity/publications/wimmera/methods.html. [Accessed 30 Jun 2004].

Brigham, A.R. 1998. Biodiversity Value of federal Collections **in** Opportunities for Federally Associated Collections. San Diego, CA, Nov 18-20, 1998.

Burrough, P.A., McDonnell R.A. 1998. *Principals of Geographical Information Systems*: Oxford University Press.

Byers, F.R. 2003. *Care and Handling of CDs and DVDs. A Guide for Librarians and Archivists*. Washington, DC: National Institute of Standards and Technology and Council on Library and Information Resources. http://www.itl.nist.gov/div895/carefordisc/CDandDVDCareandHandlingGuide.pdf [Accessed 30 Jun 2004].

CBD. 2004. *Global Taxonomic Initiative Background.* Convention on Biological Diversity. http://www.biodiv.org/programmes/cross-cutting/taxonomy/default.asp [Accessed 13 Jul 2004].

Chapman, A.D. 1999. Quality Control and Validation of Point-Sourced Environmental Resource Data pp. 409-418 **in** Lowell, K. and Jaton, A. eds. *Spatial accuracy assessment: Land information uncertainty in natural resources*. Chelsea, MI: Ann Arbor Press.

Chapman, A.D. 2002. Risk assessment and uncertainty in mapped and modelled distributions of threatened species in Australia pp 31-40 **in** Hunter, G. & Lowell, K. (eds) *Accuracy 2002 – Proceedings of the 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Melbourne: Melbourne University.

Chapman, A.D. 2004. Environmental Data Quality – b. Data Cleaning Tools. Appendix I to *Sistema de Informação Distribuído para Coleções Biológicas: A Integração do Species Analyst e SinBiota. FAPESP/Biota process no. 2001/02175-5 March 2003 – March 2004.* Campinas, Brazil: CRIA 57 pp. http://splink.cria.org.br/docs/appendix_i.pdf [Accessed 14 Jul. 2004]

Chapman, A.D. and Busby, J.R. 1994. Linking plant species information to continental biodiversity inventory, climate and environmental monitoring 177-195 **in** Miller, R.I. (ed.). *Mapping the Diversity of Nature*. London: Chapman and Hall.

Chapman, A.D., Muñoz, M.E. de S. and Koch, I. 2005. Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context. *Biodiversity Informatics* **2:** 24-41.

Chrisman, N.R. 1983. The role of quality information in the long-term functioning of a GIS. *Proceedings of AUTOCART06*, 2: 303-321. Falls Church, VA: ASPRS.

Chrisman, N.R., 1991. The Error Component in Spatial Data. pp. 165-174 **in**: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

Conn, B.J. (ed.) 1996. *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data.* Version 3. Sydney: Royal Botanic Gardens.

Conn, B.J. (ed.) 2000. *HISPID4. Herbarium Information Standards and Protocols for Interchange of Data.* Version 4 – Internet only version. Sydney: Royal Botanic Gardens. http://plantnet.rbgsyd.nsw.gov.au/Hispid4/ [Accessed 30 Jun. 2004].

Cullen, A.C. and Frey, H.C. 1999. *Probabilistic Techniques in Exposure Assessment. A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. New York: Plenum Press, 335 pages.

CRIA 2005. *species*Link. Dados e ferramentas – Data Cleaning. Campinas, Brazil: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/dc/ [Accessed 4 Apr. 2005].

Dalcin, E.C. 2004. Data Quality Concepts and Techniques Applied to Taxonomic Databases. Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton. November 2004. 266 pp. http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf [Accessed 7 Jan. 2004].

Dallwitz, M.J. and Paine, T.A. 1986. *Users guide to the DELTA system*. CSIRO Division of Entomology Report No. 13, pp. 3-6. *TDWG Standard*. http://biodiversity.uno.edu/delta/ [Accessed 9 Jul 2004].

Davis R.E., Foote, F.S., Anderson, J.M., Mikhail, E.M. 1981. *Surveying: Theory and Practice*, Sixth Edition: McGraw-Hill.

DeMers M.N. 1997. *Fundamentals of Geographic Information Systems*. John Wiley and Sons Inc.

English, L.P. 1999. Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. New York: John Wiley & Sons, Inc. 518pp.

Environment Australia. 1998. *The Darwin Declaration*. Canberra: Australian Biological Resources Study. http://www.biodiv.org/programmes/cross-cutting/taxonomy/darwin-declaration.asp [Accessed 14 Jul 2004].

Epstein, E.F., Hunter, G.J. and Agumya, A.. 1998, Liability Insurance and the Use of Geographical Information: *International Journal of Geographical Information Science* 12(3): 203-214.

Federal Aviation Administration. 2004. Wide Area Augmentation System. http://gps.faa.gov/Programs/WAAS/waas.htm [Accessed 15 Sep. 2004].

FGDC. 1998. *Geospatial Positioning Accuracy Standards.* US Federal Geographic Data Committee. http://www.fgdc.gov/standards/status/sub1_3.html [Accessed 14 Jul. 2004].

Foote, K.E. and Huebner, D.J. 1995**.** *The Geographer's Craft Project*, Department of Geography, University of Texas. http://www.colorado.edu/geography/gcraft/contents.html [Accessed 14 Jul 2004].

Gad, S.C. and Taulbee, S.M. 1996. *Handbook of data recording, maintenance, and management for the biomedical sciences*. Boca Raton: CRC Press.

Goodchild, M.F., Rhind, D.W. and Maguire, D.J. 1991. *Introduction* pp. 3-7 In: Maguire D.J.,Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

Heuvelink, G.B.M. 1998. *Error Propagation in Environmental Modeling with GIS*: Taylor and Francis.

Huang, K.-T., Yang, W.L. and Wang, R.Y. 1999. *Quality Information and Knowledge*. New Jersey: Prentice Hall.

Juran, J.M. 1964. *Managerial Breakthrough.* New York: McGraw-Hill.

Knapp, S., Lamas, G., Lughadha, E.N. and Novarino, G. 2004. Stability or stasis in the names of organisms: the evolving codes of nomenclature. Phil. Trans: Biol. Sci. 359(1444): 611-622.

Koch, I. (2003). *Coletores de plantas brasileiras.* Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 26 Jan. 2004].

Lance, K. 2001. Discussion of Pertinent Issues. pp. 5-14 in *Proceeedings USGS/EROS Data Center Kenya SCI Workshop, November 12 2001*. http://kism.iconnect.co.ke/NSDI/proceedings_kenya_NSDI.PDF [Accessed 1 Jul 2004].

Leick, A. 1995. *GPS Satellite Surveying*: John Wiley and Sons, Inc: New York.

Library of Congress. 2004. *Program for Cooperative Cataloging.* Washington, DC. US Library of Congress. http://www.loc.gov/catdir/pcc/ [Accessed 26 Jun 2004].

Lunetta, R.S. and Lyon, J.G. (eds). 2004. Remote Sensing and GIS Accuracy. Boca Raton, FL, USA:CRC Press.

Maletic, J.I. and Marcus, A. 2000. Data Cleansing: Beyond Integrity Analysis pp. 200-209 in *Proceedings of the Conference on Information Quality (IQ2000).* Boston: Massachusetts Institute of Technology. http://www.cs.wayne.edu/~amarcus/papers/IQ2000.pdf [Accessed 21 November 2003].

Mayr, E. and Ashlock, P.D. 1991. *Principles of systematic zoology*. New York: McGraw-Hill.

McElroy, S., Robins, I., Jones, G. and Kinlyside, D. 1998. *Exploring GPS, A GPS Users Guide*: The Global Positioning System Consortium.

Minnesota Planning. 1999. *Positional Accuracy Handbook. Using the National Standard for Spatial data Accuracy to measure and report geographic data quality.* Minnesota Planning:

Land Management Information Center. http://www.mnplan.state.mn.us/pdf/1999/lmic/nssda_o.pdf [Accessed 14 Jul. 2004]

Morse, L.E. 1974. Computer programs for specimen identification, key construction and description printing using taxonomic data matrices. *Publs. Mich. St. Univ. Mus., biol.* ser. 5, 1–128.

Motro, A. and Rakov, I. 1998. Estimating the Quality of Databases. *FQAS 1998*: 298-307

Naumann, F. 2001. *From Database to Information Systems – Information Quality Makes the Difference*. IBM Almaden Research Center. 17 pp.

Nebert, D. and Lance, K. 2001. Spatial Data Infrastructure – Concepts and Components. *Proceeedings JICA Workshop on Application of Geospatial Information and GIS. 19 March 2001, Kenya*. http://kism.iconnect.co.ke/JICAWorkshop/pdf/Ottichilo.pdf [Accessed 1 Jul 2004].

Nebert, D. 1999. *NSDI and Gazetteer Data*. Presented at the Digital Gazetteer Information Exchange Workshop, Oct 13-14, 1999. Transcribed and edited from audiotape. http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/session3/nebert.htm [Accessed 1 Jul 2004].

NLWRA. 2003. *Natural Resources Information Management Toolkit*. Canberra: National Land and Water Resources Audit. http://www.nlwra.gov.au/toolkit/contents.html [Accessed 7 Jul 2004].

NOAA. 2002. Removal of GPS Selective Availability (SA). http://www.ngs.noaa.gov/FGCS/info/sans_SA/ [Accessed 15 Sep 2004].

Olivieri, S., Harrison, J. and Busby, J.R. 1995. Data and Information Management and Communication. pp. 607–670 **in** Heywood, V.H. (ed.) *Global Biodiversity Assessment*. London: Cambridge University Press. 1140pp.

Pipino, L.L., Lee, Y.W. and Wang, R.Y. 2002. Data Quality Assessment.

Communications of ACM 45(4): 211-218.

Pullan, M.R., Watson, M.F., Kennedy, J.B., Raguenaud, C., Hyam, R. 2000. The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. *Taxon* **49**: 55-75.

Redman, T.C. 1996. *Data Quality for the Information Age*. Artech House, Inc.

Redman, T.C. 2001. *Data Quality: The Field Guide*. Boston, MA: Digital Press.

SA Dept Env. & Planning. 2002. *Opportunistic Biological Records (OPPORTUNE).* South Australian Department of Environment and Heritage. http://www.asdd.sa.gov.au/asdd/ANZSA1022000008.html [Accessed 14 Jul. 2004].

SEC 2002. *Final Data Quality Assurance Guidelines*. United States Securities and Exchange Commission. http://www.sec.gov/about/dataqualityguide.htm [Accessed 26 Jun 2004].

Shepherd, I.D.H. 1991.Information Integration and GIS. pp. 337-360 **in:** Maguire D.J.,Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

Spear, M., J.Hall and R.Wadsworth. 1996. *Communication of Uncertainty in Spatial Data to Policy Makers* **in** Mowrer, H.T., Czaplewski, R.L. and Hamre, R.H. (eds) Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium, May 21-23, 1996. Fort Collins, Colorado. USDA Forest Service Technical Report RM-GTR-277.

Stribling, J.B., Moulton, S.R. II and Lester, G.T. 2003. Determining the quality of taxonomic data. *J. N. Amer. Benthol. Soc*. **22(4):** 621-631.

Strong, D.M., Lee, Y.W.and Wang, R.W. 1997. Data quality in context. *Communications of ACM* 40(5): 103-110.

Taulbee, S.M. 1996. *Implementing data quality systems in biomedical records* pp. 47-75 in Gad, S.C. and Taulbee, S.M. Handbook of data recording, maintenance, and management for the biomedical sciences. Boca Raton: CRC Press.

TDWG. 2005. TDWG Working Group: Structure of Descriptive Data (SDD). Taxonomic Databases Working Group (TDWG). http://160.45.63.11/Projects/TDWG-SDD/ [Accessed 4 Apr. 2005].

University of Colorado. 2003. MaPSTeDI. *Georeferencing in MaPSTeDI*. Denver, CO: University of Colorado. http://mapstedi.colorado.edu/georeferencing.html [Accessed 30 Jun. 2004].

USGS. 2004. *What is SDTS?* Washington: USGS. http://mcmcweb.er.usgs.gov/sdts/whatsdts.html [Accessed 30 Jun. 2004].

Van Sickle, J. 1996. *GPS for Land Surveyors*: Ann Arbor Press, Inc: New York.

Wang, R.Y. 1998. A Product Perspective on Total Data Quality Management. *Communications of the ACM* 41(2): 58-65.

Wang, R.Y., Storey, V.C., Firth, C.P., 1995. A frame-work for analysis of data quality research, *IEEE Transactions on Knowledge and Data Engineering* 7: 4, 623-640.

Wieczorek, J. 2001. *MaNIS: GeoreferencingGeo-referencing Guidelines*. Berkeley: University of California, Berkeley - MaNIS http://manisnet.org/manis/GeorefGuide.html [Accessed 26 Jan. 2004].

Wieczorek, J. 2002. *Summary of the MaNIS Meeting. American Society of Mammalogists, McNeese State University, Lake Charels, LA, June 16, 2002*. Berkeley: University of California, Berkeley - MaNIS. http://manisnet.org/manis/ASM2002.html [Accessed 30 Jun. 2004].

Wieczorek, J., Guo, Q. and Hijmans, R.J. (2004). *The point-radius method for georeferencing locality descriptions and calculating associated uncertainty*. International Journal for GIS 18(8): 754-767.

Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: John Wiley & Sons.

Zhang, J. and Goodchild, M.F. 2002. Uncertainty in Geographic Information. London: Taylor and Francis.

Index