PRINCIPES ET MÉTHODES DE NETTOYAGE DE DONNÉES

DONNÉES PRIMAIRES D'ESPÈCES ET LOCALISATION D'ESPÈCES

Arthur D. Chapman[1]

Error qui non resistitur, approbatur

(Ref. *Doct. & Stud.* c. 770).

GBIF
www.gbif.org

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

Citation suggérée:

Chapman, A. D. (2005). *Principes et méthodes de nettoyage de données*, version 1.0. Trad. Roosen, A. Copenhague: Global Biodiversity Information Facility, 90 pp. Disponible en ligne sur http://links.gbif.org/gbif_nettoyage_donnees_manual_fr_v1.pdf

ISBN/DoI: non disponible.
Persistent URI: http://links.gbif.org/gbif_nettoyage_donnees_manual_fr_v1.pdf
Langage du document: Français (Titre original: *Principles and Methods of Data Cleaning*)
Date de publication de la traduction en Français: Juillet 2011.
Date de publication de la version originale en Anglais: Juillet 2005
Copyright © Global Biodiversity Information Facility, 2011
Licence:

Cette travail est publié sous une licence Creative Commons Paternité 3.0 http://creativecommons.org/licenses/by/3.0/deed.fr

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. Helophilus pendulus (Linnaeus, 1758).

En 2004 le programme DIGIT du GBIF a commandé à le Dr. Arthur Chapman l'élaboration de ce document pour mettre en valeur l'importance de la qualité des données au niveau des donnés primaires d'occurrence d'espèces. Nous admettons que ce sujet et les outils disponibles évoluent rapidement. C'est pour ça que nous voyons ce document comme un document de discussion intermédiaire sur l'état de ce sujet en 2004. Par conséquent, nous nous attendons qu'il y aura des versions futures de ce document et nous remercierions les apports par les fournisseurs des données et par la communauté d'utilisateurs.

Nettoyage des données

Introduction 1
Définition : nettoyage de données 1
La nécessité du Nettoyage des Données 2
Où sont les erreurs? 2
Prévenir les erreurs 3
Erreurs spatiales 4
Erreurs de Nomenclature et de Taxonomie 4
Fusionner les Bases de Données 5
Principes du Nettoyage des Données 6
La préparation est essentielle (Développer une Vision, une Politique et une Stratégie) 6
Organiser les données augmente l'efficacité 6
Mieux vaut prévenir que guérir 7
Responsabilité partagée (collectionneurs, détenteurs de données et utilisateurs) 7
Le Partenariat améliore l'efficacité 7
Priorétiser réduit les Duplications. 8
Etablir des objectifs et des Mesures de Performance 8
Minimiser les duplications et le remaniement des données 8
Le feedback doit être à double sens 9
L'éducation et la formation améliorent les techniques 9
Responsabilité, Transparence et Contrôlabilité 10
Documentation 10
Méthodes pour le Nettoyage des Données 11
Introduction 11
Les Données de Taxonomie et de Nomenclature 13
Certitude d'Identification 13
Orthographe des noms 15
Données spatiales 31
Saisie de données et Géoréférencement 32
Vérification des géocodes et validation 47
Données descriptives 66
Documentation des erreurs 68
Visualisation de l'Erreur 70
Outils Cités 72
1. Logiciels 72
2. Ressources en ligne 75
3. Normes et Directives 76
Conclusion 78
Remerciements 79
TexteRéférences bibliographiques 79
Références bibliographiques 80
Index 86

Introduction

Le nettoyage des données est une partie essentielle de la Chaîne de la Gestion des Informations comme mentionné dans le document associé, *Principes de la Qualité des Données* (Chapman 2005a). Comme cela y est mis en valeur, la prévention des erreurs est une tâche d'importance bien supérieure à l'identification et au nettoyage, parce qu'il est moins coûteux et plus efficace d'éviter les erreurs que d'essayer de les trouver et de les corriger ultérieurement. Peu importe à quel point le procédé d'encodage de donnée est efficace, des erreurs surviendront toujours et c'est pourquoi la validation et la correction des données ne doivent pas être ignorées. L'identification des erreurs, la validation et le nettoyage jouent un rôle-clé, en particulier lorsqu'il s'agit de données anciennes (ex: données de musée ou d'herbier rassemblées au court des 300 dernières années), et par conséquent, la prévention des erreurs ainsi que le nettoyage des informations devraient être intégrés à la politique de gestion des informations au sein d'un organisme.

Un résultat importante du nettoyage des données est de permettre d'identifier les causes basiques des erreurs détectées, et, en utilisant cette information afin d'améliorer la saisie d'information, d'éviter que ces erreurs se reproduisent.

Ce document examinera les méthodes de prévention ainsi que d'identification et de nettoyage des erreurs dans les bases de données de recueils biologiques primaires. Il traitera de directives, de méthodologies et d'outils qui peuvent aider les musées et les herbiers à appliquer de meilleurs procédés lors de la numérisation, la documentation ou la validation des informations. Mais tout d'abord, il établira un ensemble de principes simples qui devraient être suivis dans n'importe quel exercice de nettoyage d'information.

Définition : nettoyage de données

Un procédé utilisé pour déterminer des informations inexactes, incomplètes ou déraisonnables, pour ensuite améliorer la qualité en corrigeant les erreurs détectées et les oublis. Le procédé peut inclure la vérification du format, de la complétude, de la cohérence, des limites, une analyse des données pour identifier des aberrations (géographiques, statistiques, temporelles ou environnementales) ou d'autres erreurs, et l'évaluation des données par des experts dans des domaines donnés (ex: des experts en taxonomie). Le plus souvent, ces procédés ont pour résultat le signalement *(« flagging »),* la documentation et la vérification et la correction ultérieure de rapports suspects. La vérification avant validation peut également inclure de s'assurer de la conformité aux standards, règles et conventions en vigueur.

La structure générale du nettoyage de données (après Maletic & Marcus 2000) est :

- Définir et déterminer les erreurs types ;
- Chercher et identifier les occurrences d'erreurs ;
- Corriger les erreurs ;
- Documenter les cas d'erreurs et d'erreurs types ;
- Modifier le procédé d'entrée de données afin de réduire les erreurs futures.

Voici ci-dessous un certain nombre de termes utilisés par différentes personnes pour parler plus largement du même procédé. Le choix du mot employé relève de la préférence. Ces termes sont :

- Vérification des erreurs ;

- Identification des erreurs ;
- Validation des données ;
- Nettoyage des données *(« Data cleaning »)* ;
- Assainissement des données *(« Data Cleansing »)* ;
- Nettoyage en processus automatisé des données *(« data scrubbing »)* ;
- Correction des données.

J'ai pour intention d'utiliser le terme Nettoyage de Données pour englober trois sous-procédés, à savoir :

- La vérification des données et détection des erreurs ;
- La validation des données ;
- La correction des erreurs.

Un quatrième item - amélioration du procédé de prévention des erreurs- pourrait peut-être être ajouté.

La nécessité du Nettoyage des Données

Le besoin d'un nettoyage de données englobe l'amélioration de la qualité des données afin de les rendre "prêtes à l'emploi" par les utilisateurs en réduisant les erreurs dans les données et en améliorant leur documentation et leur présentation (voir le document associé sur les *Principes de Qualité des Données*- Chapman 2005a). Les erreurs de données sont fréquentes et parfois inévitables. Redman (1996) suggérait qu'à moins qu'une précaution toute particulière ait été employée, un champ d'erreur de l'ordre d'1 à 5% était à prévoir. Le point de vue général sur les erreurs est qu'elles sont mauvaises, mais une bonne compréhension de ces erreurs et de leur propagation peut mener à un contrôle de qualité actif et permettre d'améliorer la qualité des données dans leur ensemble (Burrough et McDonnell 1998). Les erreurs de localisation spatiale (géocodes) et d'identification sont deux des causes principales d'erreur en matière de données liées au recensement d'espèces (données d'occurrence), et c'est le nettoyage de ces erreurs qui est couvert dans ce dossier. Corriger les erreurs de données et éliminer les informations erronées peut être un travail long et ennuyeux (Williams *et al.* 2002) qui ne doit toutefois pas être ignoré. Il est important cependant que les erreurs ne soient pas simplement supprimées mais que les corrections soient documentées et les changements retraçables. Comme mentionné dans le document joint sur les *Principes de la Qualité des Données*, il est préférable d'ajouter des corrections à la base de données tout en conservant les données originales dans un fichier séparé, de telle sorte qu'il y ait toujours une possibilité de retourner à l'information de base.

Où sont les erreurs?

Les principales données d'espèce englobent une grande variété de données -des données de musées et d'herbiers en passant par les données d'observation (basées sur un point, une région, une zone géographique, qu'elles soient systématique ou faites grâce à des grilles de repérage), afin d'étudier des données systématiques ou non (Chapman 2005a). A cause de la nature historique de beaucoup de pièces de musées et d'herbiers (souvent nommés donnés anciennes ou '*legacy data'*), de nombreux enregistrements ne comportent que peu d'information géographiques différentes d'une simple description générale du lieu où les pièces avaient été trouvées (Chapman et Milne 1998). Dans le cas de données historiques, les codes géographiques et les coordonnées lorsqu'ils sont mentionnés ne sont souvent pas très précis (Chapman 1999) et ont généralement été ajoutés ultérieurement et par une personne autre que celui qui avait rassemblé les données (Chapman 1992).

Beaucoup de ces données comportent des inconvénients quand il s'agit de les utiliser pour des études relatives à la distribution des espèces. Les données d'enquête et d'observation sont également des informations d'une grande valeur pour beaucoup d'études et le géoréférencement d'information peut être assez fiable, mais comme les spécimens de référence matériels (« voucher ») sont rarement conservés, les informations de taxonomie et de nomenclature sont généralement moins fiables que lorsqu'il s'agit de pièces de musée documentées. Le géoréférencement des données d'enquête et d'observation peut néanmoins toujours inclure des erreurs ou des ambiguïtés, par exemple lorsqu'il s'agit de savoir si le code géographique concerne le centre de la grille ou un coin d'une grille de marquage.

Une grande partie des données (de musée ou d'observation) ont été rassemblées de manière plus opportuniste que méthodique (Chapman 1999, Williams et al. 2002) et cela peut avoir pour résultat de nombreuses 'biais' (détermination incontrôlée de l'interprétation) spatiaux -par exemple, des spécimens fortement corrélés par un réseau routier ou fluvial (Margules et Redhead 1995, Chapman 1999, Peterson et al. 2002, Lampe et Riede 2002). Les données de musées et d'herbiers et la plupart des données d'observation ne fournissent généralement que des informations sur la présence d'une entité à un moment donné et ne renseignent aucunement sur ses absences dans d'autres lieux ou à une autre époque (Peterson et al.1998). Cela réduit leur usage dans certains modèles environnementaux mais demeure le seul ensemble d'informations biologiques complètes couvrant remontant à plus loin que les 200 dernières années. Il serait trop onéreux de remplacer ces données par de nouvelles études. Il est inhabituel pour une simple étude d'excéder 1 million de dollars (Burbidge 1991). De plus, de par le fait qu'elles avaient été rassemblées au fil du temps, elles fournissent des données de références d'une grande valeur sur la diversité biologique à une époque pendant laquelle les humains ont eu sur elle un impact immense (Chapman et Busby 1994). Ce sont des ressources essentielles dans tout effort de conservation de l'environnement, puisqu'elles fournissent les seuls résultats complètement documentés d'occurrence d'espèce dans des lieux qui ont pu subir un changement d'habitat pour cause de déforestation à des fins agricoles, d'urbanisation, de changement climatique, ou pour tout autre raison (Chapman 1999).

Prévenir les erreurs

Comme précédemment souligné, la prévention des erreurs est préférable à leur correction ultérieure, et de nouveaux outils sont en cours de développement afin d'assister les institutions dans la prévention de ces erreurs.

Des outils sont développés pour assister le procédé d'addition d'information de géoréférencement aux bases de données. De tels outils incluent eGaz (Shattuck 1997), geoLoc (CRIA 2004a), BioGeomancer (Peabody Museum *n.dat*.), GEOLocate (Rios et Bart *n.dat*.) et la Calculatrice Géoréférentielle, the *Georeferencing Calculator* (Wieczorek (2001b). Un projet fondé en 2005 par la fondation Gordon et Betty Moore et comprenant une collaboration internationale rassemble désormais ces outils dans le but de les rendre plus accessibles à la fois lorsqu'il s'agit de sources ouvertes de logiciel individuelles et de Web Services (services en ligne). Le besoin pour toujours plus d'outils de validation variés ne peut néanmoins par être oublié. Ces outils seront abordés plus en détail plus tard dans ce dossier.

Des outils sont aussi en cours de développement afin d'assister la réduction des erreurs en matière de données de taxonomie et de nomenclature. Il y a deux raisons principales aux erreurs de ces données; les identifications inexactes ou mauvaises identifications (taxonomie) et les fautes

d'orthographe (nomenclature). Les outils assistant l'identification des taxons incluent des taxonomies améliorées, des faunes et flores (à la fois régionales et locales), des codes automatisés et gérés par ordinateur pour accéder aux taxons et l'imagerie numérique des types et espèces de spécimens. Quant à l'orthographe des noms, des listes de noms globales et régionales sont développées, ce qui permet de constituer des référentiels taxonomiques et des check-lists pour la saisie dans les bases de données, afin de réduire les erreurs.

Il est possible que la meilleure façon de prévenir de nombreuses erreurs soit de proprement concevoir les bases de données dès le départ. En mettant en place une philosophie et une démarche conception de bases de données relationnelles intelligentes, toute information qui serait fréquemment répétée telle que les noms d'espèces, les lieux et les institutions, n'aurait besoin d'être saisie qu'une seule fois et vérifiée à la sortie. L'intégrité référentielle garantit donc l'exactitude des entrées à venir.

Erreurs spatiales

Lorsqu'il s'agit de déterminer la qualité des données d'espèce d'un point de vue spatiale, il y a un certain nombre de problèmes à examiner. Ceux-ci incluent l'identité du spécimen ou de l'objet de l'observation - une erreur d'identification peut placer la donnée enregistrée en dehors du groupe de taxon attendu et de ce fait apparaître comme une erreur spatiale- des erreurs d'encodage géographique (latitude et longitude), et des aberrations spatiales dans le ramassage des données. L'emploi de critères spatiaux peut aider à trouver les taxons mal identifiés ou mal placés géographiquement. Le problème des erreurs spatiales- très évidentes dans le cas d'herbiers et de musées (ex: trouvés le long d'une route) est plus un problème pour les collectes futures et la conception des études à venir que pour l'exactitude des données relatives à un spécimen de plante ou d'animal en particulier (Margules et Redhead 1995, Williams *et al*. 2002). En effet, les erreurs de séries ('*collection bias*') sont liées à la totalité des spécimens d'une espèce données dans son ensemble, plutôt qu'à un spécimen individuel. Afin d'améliorer la couverture spatiale et taxonomique globale des données biologiques au sein d'une zone donnée, et donc de réduire les erreurs spatiales (*'spatial bias'*), des données historiques existantes peuvent être utilisées afin de déterminer les emplacements les plus intéressants écologiquement parlant pour les études à venir, en utilisant par exemple des critères environnementaux (climat etc...) mais aussi géographiques. Neldner *et al.* 1995).

Erreurs de Nomenclature et de Taxonomie

Les noms constituent la principale clé permettant d'accéder aux informations primaires des bases de données d'espèce *(« primary species database »)*. Si le nom est erroné, l'accès à l'information sera donc plus difficile, voire impossible. En dépit de l'existence des règles de nomenclature biologique datant de plus de 100 ans, les informations d'une base de données concernant la nomenclature et la taxonomie (le *Classification Domain* de Dalcin 2004) sont souvent celles où il est le plus difficile de détecter les erreurs. C'est aussi le domaine qui cause le plus d'angoisses et de perte de confiance auprès des utilisateurs de bases de données primaires d'espèce. Ceci est souvent due au fait que ces utilisateurs ignorent la nécessité de changements taxonomiques et sur de nomenclature, mais c'est aussi en partie dû aux taxonomistes eux même qui ne documentent ou ne leurs expliquent pas suffisamment ces changement, compliquant ainsi la relation entre les noms et les taxons et laissant place à la confusion en matière de concepts taxonomiques qui ne sont souvent pas bien couverts dans les bases de données primaires d'espèces (Berendsohn 1997).

Parmi ces erreurs, la plus simple à rectifier est celle relative aux données sur la nomenclature - les fautes d'orthographe. Les listes de noms (et synonymes) sont l'outil clé permettant d'aider dans cette tâche. De nombreuses listes existent par région et/ou par groupes taxonomiques, et sont progressivement intégrées dans des listes plus globales (Froese and Bisby 2002). Il existe encore de nombreuses régions du monde où les groupes taxonomiques n'ont cependant pas encore de listes fiables.

L'erreur taxonomique - une identification inexacte ou l'échec d'identification - est l'erreur la plus difficile à détecter et à corriger. Les musées et les herbiers avaient traditionnellement en service un système *determinavit* au moyen duquel les experts travaillant dans des groupes taxonomiques données examinent de temps en temps les spécimens et déterminent ainsi leur circonscription ou leur identification. C'est une méthode qui a fait ses preuves, mais qui prend beaucoup de temps et qui laisse une place importante au hasard. Il est cependant assez peu probable de trouver un moyen de contourner ce problème, car l'identification automatisée par ordinateur ne risque pas de se concrétiser dans un futur proche ou lointain. Il y a cependant de nombreux outils disponibles pour aider dans ce processus. Ils comprennent à la fois les publications taxonomiques traditionnelles avec lesquelles nous sommes tous familiers, et des nouveaux outils électroniques. Les outils traditionnels incluent des publications telles que des révisions taxonomiques, des faunes et flores nationales et régionales et des check-lists illustrées. Les outils plus modernes incluent des clés de détermination taxonomique automatiquement générées par ordinateur, des publications électroniques interactives illustrées, des descriptions, des codes, et des glossaires illustrés. Des bases de données organisées autour de critères précis, des outils d'imagerie, des bases de données d'images scientifiques, des images de séries méthodiquement rassemblées et des images facilement accessibles en ligne (qu'elles aient été vérifiées scientifiquement ou non).

Fusionner les Bases de Données

La fusion de deux bases de données ou plus identifiera les erreurs (là où il y aura des différences entre les deux bases de données) et créera aussi de nouvelles erreurs (duplication d'enregistrements par exemple). Les entrées doubles devraient être balisées et signalées lors de la fusion afin qu'elles puissent être identifiées et exclues de l'analyse dans les cas où des données doubles pourraient biaiser les résultats, etc.., mais ne devraient généralement pas être supprimées. Bien qu'ayant l'air de doublons, dans de nombreux cas les enregistrements de chaque base de données peuvent inclure des informations uniques à chacune d'elles, et donc se contenter d'en supprimer un des deux (connu sous le nom de 'Fusionner et Epurer' (Maletic et Marcus 2000)) n'est pas toujours la meilleure option puisque cela peut mener à une perte de données d'une grande valeur.

Un autre problème qui peut survenir lors du croisement des données est le mélange d'informations basées sur différents critères, par exemple différentes sortes de concepts taxonomiques, différents sortes de suppositions ou d'unités de mesure et différentes mécanismes de qualité de contrôle. La source des données individuelles doit toujours être documentée lors de type de fusion de telle sorte que le procédé de nettoyage des données puisse être mené sur ces données de différentes façons selon les différentes sources. Autrement, il pourrait être plus dur de nettoyer efficacement la nouvelle base de données et de documenter efficacement les changements.

Principes du Nettoyage des Données

Un grand nombre des principes de nettoyage de données chevauchent les principes généraux de qualité des données couverts dans le document associé sur les *Principes de Qualité des Données* (Chapman 2005a). Les principes clés incluent:

La préparation est essentielle (Développer une Vision, une Politique et une Stratégie)

Une bonne préparation est une part essentielle d'une bonne politique de gestion des données. La Chaîne de Gestion de l'Information (figure 1) (Chapman 2005a), place le Nettoyage des Données en chaînon centrale nécessitant d'être incorporé dans la politique et la vision de qualité des données de l'organisme. Une stratégie visant à mettre en place le nettoyage des données et leur validation dans la culture de l'organisme améliorera la qualité d'ensemble des données de cet organisme et améliorera également sa notoriété à la fois tant auprès des utilisateurs que des fournisseurs.

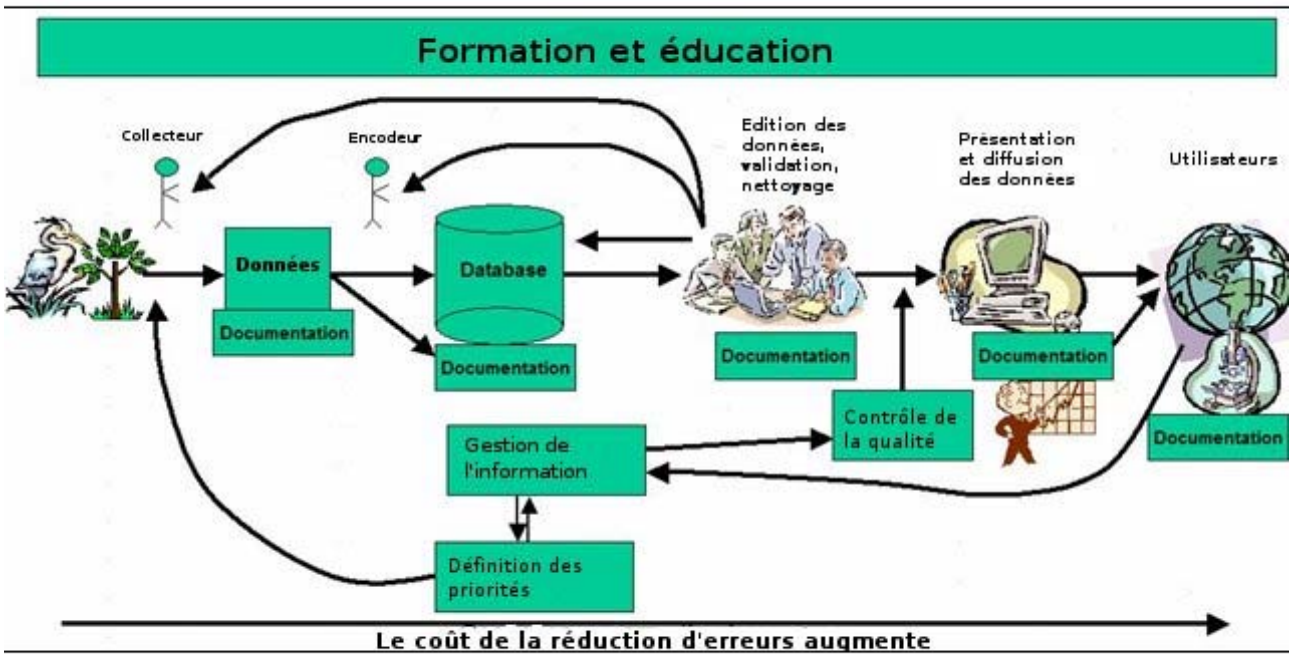


Fig. 1. *Chaîne de Gestion de l'Information montrant que le coût de la correction d'erreur augmente au fur et à mesure que l'on se déplace dans la chaîne. L'Education, la Formation et la Documentation sont inhérents à toutes les étapes (d'après Chapman 2005a).*

Organiser les données augmente l'efficacité

Organiser les données avant la vérification, la validation et la correction des données peut augmenter l'efficacité et considérablement réduire le temps et le coût du nettoyage des données. Par

exemple, en triant les données par emplacement, il est possible de gagner en efficacité en vérifiant tous enregistrements dépendants d'un emplacement au même moment donné, plus qu'en faisant des allers-retours entre les mots clés. De même, en triant les enregistrements par collectionneur et par date, il est possible de détecter des erreurs là où un enregistrement serait placé à un emplacement improbable pour ce collectionneur à cette date. Les fautes d'orthographe dans toutes sortes de champs peuvent également être détectées de cette manière.

Mieux vaut prévenir que guérir

Comme mis en valeur précédemment (Chapman 2005a), il est beaucoup moins cher et beaucoup plus efficace de prévenir une erreur que de la rechercher et de la corriger plus tard. Il est aussi important que lorsque les erreurs sont détectées un mécanisme de feedback assure qu'elles ne se reproduiront plus lors de la saisie de données, ou tout du moins qu'il y ait beaucoup moins de chances qu'elles se reproduisent. La bonne conception de bases de données garantit que la saisie de données est contrôlée de telle sorte que les entités comme les noms des taxons, les emplacements et les personnes ne soient entrées qu'une seule fois et qu'ils soient vérifiées à la fin de la saisie. Ceci peut être effectué grâce à des menus déroulant ou par le biais d'une identification par clavier dans le cas de données déjà existant dans un domaine.

Responsabilité partagée (collectionneurs, détenteurs de données et utilisateurs).

La responsabilité en matière de nettoyage de données appartient à tous. La responsabilité principale de la portion de la Chaîne de Gestion de l'Information concernant le nettoyage des données (figure 1) relève en tout logique du détenteur de ces données – la personne ou l'organisme dont la responsabilité principale est la gestion et l'organisation des données. Le collecteur, lui aussi, a sa part de responsabilité et a besoin de répondre aux questions du détenteur lorsque celui ci trouve des erreurs ou des ambiguïtés qui pourraient se rapporter à l'information originale prodiguée par le collectionneur. Cela pourrait traiter d'ambiguïté liée à l'étiquetage, d'erreurs de date ou d'emplacement, etc. Comme il deviendra évident plus tard dans ce document, l'utilisateur a aussi une part de responsabilité importante puisqu'il se doit de renvoyer les informations aux détenteurs des données lorsqu'il rencontre des erreurs ou des omissions, ceci incluant des erreurs de documentation associée aux données. C'est souvent l'utilisateur lui-même qui, en analysant ou en regardant ces données dans un autre contexte, remarquera des erreurs ou des valeurs extrêmes de données qui seraient autrement passés inaperçues. Un musée peut ne rassembler à lui seul qu'une petite partie d'un ensemble de données (d'un état ou d'une région par exemple), et ce n'est que lorsque les données sont combinées avec celles d'autres sources que des erreurs peuvent alors apparaître plus clairement. Nombre des outils élaborés dans ce document sont beaucoup plus efficaces lorsque l'on s'attaque à la globalité des données d'une espèce, d'un collectionneur ou d'une expédition plutôt qu'à une petite partie.

Le Partenariat améliore l'efficacité

Le partenariat peut se révéler une méthode très efficace de gestion du nettoyage des données. Comme mentionné précédemment, l'utilisateur est souvent le mieux placé pour identifier les erreurs de données. Si le détenteur de ces données peut développer des partenariats avec ces utilisateurs-clé alors ces erreurs ne seront pas ignorés. En développant de tels partenariats, beaucoup de procédés

de validation de données ne devront pas être dupliqués, les erreurs seront documentées et corrigées de manière plus systématique et les nouvelles erreurs ne seront pas incorporées grâce à la "correction" par inadvertance de données suspectes ne comportant pourtant pas d'erreur. Il est important de créer de tels partenariats avec les utilisateurs à l'intérieur des organisations tout comme à l'extérieur, comme il en est fait mention dans le document associé sur les *Principes de la Qualité des Données*.

Priorétiser réduit les Duplications.

Lorsqu'il s'agit d'organisation et de tri, prioriser aide à réduire les coûts et augmente l'efficacité. Il est souvent judicieux de se concentrer sur les enregistrements comportant des données détaillées pouvant être nettoyées à un prix réduit. Par exemple, celles pouvant être examinées en utilisant des traitements par lot ou des méthodes automatisées, avant de travailler sur les enregistrements plus compliquées. En se concentrant sur ces données qui ont plus de valeur aux yeux des utilisateurs, il y a également une plus grande probabilité que les erreurs soient détectées et corrigées. Cela améliore la relation client/fournisseur et les notoriétés, et fournit une plus grande motivation à améliorer la qualité des données à la fois pour les fournisseurs de données pour ceux qui les utilisent, de par son usage immédiat.

Etablir des objectifs et des Mesures de Performance

Les mesures de performance sont aussi un ajout de grande valeur aux procédures de contrôle de qualité, et sont énormément utilisées avec les métadonnées géographiques. Elles aident également un organisme à gérer le procédé de nettoyage des données. De même que cela fournit aux utilisateurs des informations sur les données et leur qualité, de telles mesures peuvent également être utilisées par les gestionnaires et conservateurs pour détecter les parties de bases de données qui pourraient nécessiter une attention particulière. Ces mesures peuvent inclure des vérifications statistiques des données (par exemple, 95% de toutes les enregistrements ont une précision inférieure à 5000 mètres comparée à leur position rapportée), des vérifications du niveau de contrôle de la qualité (par exemple, 65% de toutes les enregistrements ont été vérifiés par un taxonomiste qualifié dans les 5 dernières années contre 90% dans les 10 dernières années), et des vérifications de complétude (par exemple, tous les carrés d'une grille géographique de 10 minutes ont été échantillonnés) (Chapman 2005a).

Minimiser les duplications et le remaniement des données

La duplication est un facteur clé de nettoyage de données dans la plupart des organismes. De nombreux organismes effectuent le géoréférencement en même temps qu'ils entrent les enregistrements dans les bases de données. Comme les enregistrements sont rarement triés géographiquement, cela signifie que les localisations identiques ou similaires seront cherchés un bon nombre de fois. En effectuant le géoréférencement des collections comportant uniquement des informations géographiques sous forme textuelle et non pas des informations de coordonnées en tant qu'opération spéciale, les enregistrements provenant de localisations similaires peuvent être triés et localisés sur la carte ou dans le dictionnaire géographique (*'gazetteer'*) approprié. Certaines institutions utilisent également la base de données elle-même pour aider à réduire les duplications en cherchant à voir si la localisation a déjà été géoréférencé (voir *Saisie de Données et Géoréférencement*, ci-dessous).

La documentation des procédures de validation (préférablement dans un format standardisé) est également importante lorsqu'il s'agit de réduire le remaniement des données. Par exemple, les vérifications de qualité des données effectuées par un utilisateur peuvent identifier un certain nombre d'enregistrements suspects. Ces enregistrements pourraient alors être vérifiés et s'avérer être valides et de réelles valeurs extrêmes. Si cette information n'est pas documentée dans l'enregistrement, plus tard, quelqu'un d'autre pourrait mener à bien encore plus de contrôles de qualité qui identifieraient à nouveau l'enregistrement comme étant suspect. Cette personne pourrait alors passer un temps plus précieux à revérifier les informations et à retravailler les données. Lors de la conception des bases de données, un champ ou des champs doivent être inclus et indiquer si les données ont été vérifiées, par qui, quand et quel en avait été le résultat.

L'expérience dans la sphère des affaires a montré que l'utilisation de la chaine de gestion de l'information (voir figure 1) peut réduire les duplications et le remaniement des données et mener à la réduction du taux d'erreurs à plus de 50% et à une réduction du coût liés à l'utilisation de bases de données de mauvaises qualité à plus de deux tiers (Redman 2001). Ceci est largement dû à un gain d'efficacité par l'affectation de responsabilités claires dans la gestion des informations et le contrôle de qualité, minimisant les embouteillages et les temps d'attente, minimisant les duplications au travers de différentes équipes refaisant les vérifications de contrôle de qualité et en améliorant l'identification de meilleures méthodes de travail (Chapman 2005a).

Le feedback doit être à double sens

Les utilisateurs de ces données mèneront inévitablement à bien des détections d'erreurs, et il est important qu'ils fassent part de leurs remarques aux détenteurs des données. Comme il a déjà été mentionné, les utilisateurs ont beaucoup plus de chance de détecter certaines erreurs types par le biais de la combinaison des données de sources différentes que les détenteurs de données individuels travaillant indépendamment. Il est essentiel que les détenteurs de données encouragent le feedback chez les utilisateurs de leurs données, et qu'ils tiennent compte des remarques qui leur sont faites (Chapman 2005a). Des mécanismes standards de feedback ont besoin d'être développé, et des procédures de réceptions de feedback doivent faire l'objet d'un accord entre les détenteurs de données et les utilisateurs. Les détenteurs de données doivent également transmettre les informations sur les erreurs aux collecteurs et fournisseurs de données quand cela est pertinent. De cette manière, il y a beaucoup plus de chances que la fréquence d'erreurs futures soit réduite et que la qualité générale des données soit améliorée.

L'éducation et la formation améliorent les techniques

Une mauvaise formation, principalement lors des étapes de collecte des données et de saisie de la Chaine de Qualité de l'Information, est la cause d'une grosse proportion des erreurs dans les données primaires d'espèce *(« primary species data »)*. Les collecteurs de données ont besoin d'être éduqués quant aux besoins des détenteurs des données et des utilisateurs de telle sorte que les bonnes informations soient rassemblées (exemple, toutes les stade de développement et l'âge des spécimens), que les séries soient bien documentées- exemple, l'information relative à l'emplacement a bien été enregistrée (par exemple, est ce que 10km NW de la ville "y" signifie 10km sur route ou à vol d'oiseau?), que les standards soient appliqués lorsque cela est pertinent (exemple: la même échelle est utilisée pour les études liées), et que les étiquettes soient claires et lisible et préférablement disposées avec cohérence afin de faciliter la saisie des données.

La formation à la saisie de données est aussi très importante comme relevé dans le manuel de géoréférencement MaPSTeDI (Université du Colorado Regents 2003a). Une bonne formation à des encodeurs peut considérablement réduire les erreurs associées à la saisie, réduisant ainsi les coûts et améliorant la qualité générale des données.

Responsabilité, Transparence et Contrôlabilité

La responsabilité, la transparence et la contrôlabilité sont des éléments essentiels du nettoyage des données. Le hasard et les nettoyages de données non prévus sont des exercices inefficaces et généralement improductifs. Au sein des stratégies et des politiques de qualité des données, des lignes claires de responsabilité des données ont besoin d'être établies. Afin d'améliorer l' "aptitude à l'emploi" *('fitness for use')* des données et donc leur qualité, le processus de nettoyage des données nécessite une transparence et une bonne documentation incluant de bons protocoles de contrôle visant à réduire les duplicatas et à assurer qu'une fois corrigées, les erreurs ne surviennent plus.

Documentation

La documentation est la clé d'une bonne qualité des données. Sans une bonne documentation, il est difficile pour les utilisateurs de déterminer si les données sont prêtes à l'emploi et pour leur détenteurs de savoir par qui ou par quoi les contrôles de qualité des données ont été menés. La documentation est généralement de type double et elle devrait être prise en compte dès la conception de la base de données. Le premier aspect est attaché à chaque enregistrement individuel et enregistre quelles vérifications de données ont été faites et quels changements ont été apportés et par qui. Le deuxième aspect est celui des métadonnées qui enregistrent les informations décrivant de l'extérieur l'ensemble du jeu de données. Les deux sont importants, et sans eux, la bonne qualité des données est compromise.

Méthodes pour le Nettoyage des Données

Introduction

Les musées et les herbiers à travers le monde commencent de plus en plus à entrer leur collections dans des bases de données, et commencent à rendre moins une partie des informations disponible sur l'Internet. Le taux de saisie en base de données a considérablement augmenté ces dernières années avec le développement d'outils et de méthodologies pouvant assister ce processus, et les publications ont augmenté depuis la création de le Système Mondial d'Information en Biodiversité, *Global Biodiversity Information Facility* (GBIF) qui a pour but de "rendre les informations mondiales en matière de biodiversité librement et universellement accessibles via l'Internet" (GBIF 2003a).

De même que pour les bonnes pratiques, (voir documents associés : *Principes de Qualité des Données* et *Principes du Nettoyage des Données*), il y a un grand besoin en outils utiles et puissants qui automatisent ou assistent grandement le procédé de nettoyage. Les méthodes automatisées peuvent n'être qu'une part de cette procédure et il y a un besoin continu dans le développement de nouveaux outils d'assistance, ainsi que dans l'intégration de leur utilisation au quotidien dans des pratiques optimalisées. Le nettoyage de données manuel est laborieux et demande beaucoup de temps, et est de par sa nature sujet aux erreurs (Maletic et Marcus 2002), mais il continuera de jouer un rôle important avec les bases de données de données primaires d'occurrence. Si nécessaire, il devrait représenter le dernier recours pour les jeux de données comportant peu de données, ou là où les autres contrôles n'ont laissé que quelques erreurs qui ne peuvent être vérifiées autrement.

Certaines des techniques qui ont été développées incluent l'utilisation de modèles climatiques pour identifier les valeurs extrêmes d'espace de climat (Chapman 1992, 1999, Chapman et Busby 1994), d'espaces géographiques (CRIA 2004b, Hijmans *et al*. 2005, Marino *et al.* in prep.), l'utilisation d'outils de géoréférencement automatisés (Beaman 2002, Wieczorek et Beaman 2002) et bien d'autres encore. La plupart des institutions de collections n'ont pas un niveau élevé d'expertise dans les techniques de gestion des données ou dans le Système d'Information Géographique, *Geographic Information Systems* (GIS). Ce dont ces institutions ont besoin est un simple et peu coûteux jeu d'outils pour assister la saisie de données et d'informations, ce qui inclut le géoréférencement des informations, et des outils également similaire et peu coûteux pour la validation des données sans forcément intégrer un logiciel GIS onéreux. Certains outils ont été développés pour assister la saisie de données -des outils tels que Biota (Colwell 2002), BRAHMS (Université d'Oxford 2004), Specify (Université du Kansas 2003a), BioLink (Shattuck etFitzsimmons 2000), Biótica (Conabio 2002), et d'autres qui fournissent une gestion des bases de données et des entrées de données associées (Podolsky 1996, Berendsohn *et al.* 2003); et eGaz (Shattuck 1997), geoLoc (CRIA 2004a, Marino *et al.* in prep.), GEOLocate (Rios and Bart *n.dat*.) et BioGeoMancer (Peabody Museum *n.dat*.), qui assistent dans le géoréférencement des collections. Il y a également un certain nombre de directives documentées disponibles sur l'Internet qui peuvent assister les institutions dans la mise en place et la gestion de leur programmes de saisie, par exemple les directives de Géoréférencement MaNIS (Wieczorek 2001a) et les directives de géoréférencement MaPSTeDI (University of Colorado Regents 2003a) et HISPID (Conn 1996, 2000).

Il y a de nombreuses méthodes et techniques qui peuvent aider au nettoyage des erreurs en matière de données primaires d'espèce et d'occurrence des espèces dans les bases de données. Elles varient de méthodes ayant été mises en œuvre en musées et herbiers depuis des centaines d'années à des

méthodes automatisées qui sont toujours très largement non testées. Ce document regarde en détail une variété de méthodes de nettoyage de bases de données d'espèces, et quand cela est possible, fournit des exemples. Il ne s'agit en aucune façon d'une liste exhaustive puisque de nombreuses institutions ont développé leurs propres techniques et méthodes.

De par la nature historique des collections, il n'est pas possible que toutes les informations de géoréférencement soient extrêmement précises, ou qu'il existe un niveau de précision consistant à l'intérieur d 'une base de données. Les données avec une précision très faible ne sont cependant pas nécessairement de mauvaise qualité. La qualité n'est à considéré que lorsque la donnée est utilisée et cela n'est pas une propriété de la donnée *per se* (voir le document sur *les Principes de la Qualité des Données*- Chapman 2005a). La qualité est plutôt un facteur rendant l'information prête à l'emploi ou un à emploi potentiel, et c'est un terme relatif. Ce qui est important est que les utilisateurs des données soient capables de déterminer à partir de la donnée elle-même si cette donnée est susceptible de correspondre à l'application nécessaire. Le degré de précision de chaque géocode donné devrait donc être contenu dans la base de données. Je préfère que cela soit entré sous une forme non-catégorique, enregistré en mètres, cependant beaucoup de bases de données ont développé des codes de classes et de catégories dans ce but. Quand cette information est accessible, un utilisateur peut demander, par exemple, seulement les données qui sont plus grandes qu'une certaine valeur métrique; par exemple, plus grand que 5000 mètres (voir l'exemple en utilisant les codes d'extraction des données à l'Université du Colorado à Regents 2003b). Il y a un certain nombre de manières de déterminer l'exactitude des enregistrements géoréférencés. La méthode du point-radius (Wieczorek *et al.* 2004), est pou moi la plus simple et la plus pratique, et est celle dont l'emploi est déjà recommandé en Australie (Chapman et Busby 1994). Il est aussi important que des outils de géoréférencements automatisés calculent l'exactitude en tant que champ à la sortie. Le geoLoc geoLoc (CRIA 2004a, Marino *et al.* in prep.) et le BioGeomancer (Peabody Museum *n.dat.*) sont des outils, qui sont encore en développement, incluant cette caractéristique.

Au fil du temps, on espère que les ressources en matière de données de collection d'espèces s'amélioreront puisque que les institutions utiliseront des instruments plus précis (comme le GPS) pour enregistrer les emplacements de nouveaux enregistrements et puisque les enregistrements historiques seront corrigés et améliorés. Il est également important que les collecteurs fassent le meilleur usage possible de ces outils qui sont à leur disposition et qu'ils ne se contentent pas simplement d'un GPS pour enregistrer les données avec une résolution d'un arc -minute à cause de raisons historiques – puisque c'est la mesure la plus précise qu'ils avaient pu enregistrer avant l'invention du GPS. Si malgré tout ils procèdent ainsi, ils devraient alors s'assurer que l'exactitude appropriée est ajoutée à la base de données, autrement quoi on pourrait penser que comme un GPS a été utilisé, l'exactitude est de 10 mètres en opposition aux 2000 mètres réels. La prévention des erreurs est préférable à leur détection, mais l'importance de la détection des erreurs ne doit pas être diminuée puisque la prévention seule ne peut jamais suffire à garantir que toutes les erreurs seront évitées.

Les Données de Taxonomie et de Nomenclature

Les noms, qu'ils se présentent sous forme binomiale scientifique ou de noms communs, fournissent le premier point d'entrée de la plupart des bases de données d'espèces. Les erreurs de noms peuvent survenir d'un certain nombre de façons: l'identification est fausse, le nom est mal orthographié ou le format est faux (ou pas celui attendu par l'utilisateur). Le premier de ces problèmes n'est pas facile à vérifier et à rectifier sans d'immenses efforts, et nécessite les services d'un taxonomiste expert. Les autres par contre sont plus facilement remarqués avec une base de données bien conçue et des méthodes d'assistance à la saisie des données, de telle sorte que ces erreurs ne se reproduisent pas ou qu'elles sont minimes.

Certitude d'Identification

Les musées et les herbiers avaient traditionnellement un système d'opération *determinavit* par lequel les experts travaillant dans des groupes taxonomiques examinent de temps en temps les spécimens et déterminent leurs identifications. Cela peut avoir lieu lors d'une révision taxonomique générale, ou bien lors de la visite d'un expert externe dans une institution qui, pendant sa présence, vérifie les corrections. C'est une méthode qui a fait ses preuves, mais qui prend beaucoup de temps et qui laisse une place importante au hasard. Il est assez peu probable de trouver un moyen de contourner ce problème, mais l'identification automatisée par ordinateur ne risque pas d'être une option dans un futur proche ou lointain.

Conception de la base de données

Une option pourrait être d'incorporer un champ dans la base de données qui fournirait certaines indications relatives à la certitude de l'identification lorsqu'elle a été faite. Il y a un certain nombre de moyen par lesquels cela pourrait être fait, et il est peut-être nécessaire d'en discuter pour développer des standards de méthodologie. Cela serait un code contenu dans un champ pouvant apparaître comme:

- identifié par un expert mondial dans le taxon considéré avec une haute certitude
- identifié par un expert mondial dans le taxon considéré avec une certitude raisonnable
- identifié par un expert mondial dans le taxon considéré avec quelques doutes
- identifié par un expert régional dans le taxon considéré avec une haute certitude
- identifié par un expert régional dans le taxon considéré avec une certitude raisonnable
- identifié par un expert régional dans le taxon considéré avec quelques doutes
- identifié par un non-expert dans le taxon considéré avec une haute certitude
- identifié par un non-expert dans le taxon considéré avec une certitude raisonnable
- identifié par un non-expert dans le taxon considéré avec quelques doutes
- identifié par un collecteur avec une haute certitude
- identifié par un collecteur avec une certitude raisonnable
- identifié par un collecteur avec quelques doutes

Le système de hiérarchisation de ces catégories et le fait de savoir s'il s'agit des meilleures catégories possibles devraient être ouvert à la discussion. Il y a apparemment des institutions qui ont déjà des champs de cette sorte. Le HISPID Standard Version 4 (Conn 2000) inclut en effet une

version simplifiée- la Table de Vérification en cinq étapes (Tableau 1).

De nombreuses institutions ont également déjà une forme de certitude enregistrée avec l'utilisation de termes tels que: "aff.", "cf.", "*s.la*t.", *s.str.*", "?". Même si certains d'entre eux (aff., cf.) ont des définitions strictes, leur utilisation par les personnes individuelles peut varier considérablement. L'emploi du *sensu stricto* et *enso lato* incluent des variations dans les concepts taxonomiques plus que dans les dégrées de certitudes, même si ce n'est pas toujours utilisé ainsi.

0	le nom de l'enregistrement n'a pas été officiellement vérifié
1	le nom de l'enregistrement a été déterminé en comparaison avec d'autre plantes identifiées
2	le nom de l'enregistrement a été déterminé par un taxonomiste ou par toute autre personne compétente à l'aide d'herbiers et/ou de bibliothèques et/ou des sujets vivants déjà documentés
3	le nom de la plante a été déterminé par un taxonomiste engagé dans une révision systématique du groupe
4	L'enregistrement fait partie d'un type rassemblé ou propagé à partir du matériel-type avec des méthodes de reproduction asexuées

Tableau *1.* *Table de Vérification en cinq étapes de HISPID (Conn 2000).*

En tant qu'alternative dans les cas où les noms sont définis à partir d'autres sources que l'expertise taxonomique, il serait possible de faire la liste des sources des noms utilisés (d'après Wiley 1981):

- description des nouveaux taxons
- révisions taxonomiques
- classifications
- clés taxonomiques
- études sur la faune et la flore
- atlas
- catalogues
- check-lists
- manuels
- bourses d'études taxonomiques et règles de nomenclature
- analyse phylogénétique

Saisie des Données

Pendant que les données sont entrées dans la base de données, des contrôles peuvent être effectués que le nom ait été vérifié par un expert ou non, et si un ou plusieurs des champs précédents existe, qu'ils ont bien été rentrés. Si de tels champs sont utilisés, ils devraient être entrés grâce à l'utilisation d'une check-list ou d'un document faisant autorité qui réduirait les options possibles et qui donc réduirait les chances que des erreurs soient ajoutées.

Vérification des erreurs

Les méthodes de vérifications des géocodes (voir sous *Données Spatiales*, ci-dessous) peuvent

également aider à identifier les erreurs d'identification ou les identifications inexactes au travers de la détection de valeurs extrêmes dans des espaces géographiques ou environnementaux. Bien que généralement une valeur extrême trouvée par le biais de la vérification des géocodes soit une erreur soit de latitude soit de longitude, elle indique parfois que le spécimen avait mal été identifié pendant l'étude du taxon et qu'il tombe donc en dehors des de l'éventail normal des données climatiques, environnementales ou géographiques du taxon. Voir ci-dessous une discussion plus détaillée sur les techniques d'identification de valeurs extrêmes géographiques.

La méthode principale pour vérifier si une collection a été identifiée ou non de manière fiable est néanmoins que les experts vérifient l'identification en examinant les spécimens ou collections de références *(« voucher collection »)* préalablement validées lorsqu'elles existent. Les méthodes de détection de valeurs extrêmes dans le référencement géographique ne peuvent déterminer si la collection a été correctement identifiée ou non, mais peuvent aider les experts taxonomistes à identifier et vérifier les collections prioritaires. Avec les données d'observation, les experts peuvent être en mesure de déterminer, à partir de leurs connaissances personnelles, si le taxon est vraisemblable avec la zone géographique (exemple, *Birds Australia* 2004) ; mais généralement, il est difficile d'identifier une identification inexacte d'un enregistrement d'observation là où il n'y pas de spécimens de référence *(« voucher specimens »)*. De nombreuses institutions pourraient signaler des informations douteuses ou suspectes et il revient donc ensuite à l'utilisateur de décider de leur utilité à ses yeux.

Orthographe des noms

Ce dossier ne tentera pas de couvrir toutes les sortes de noms possibles qui pourraient être saisis dans une base de données primaires d'espèce. Par exemple, les hybrides et cultivars dans les bases de données de plantes, les synonymes de différentes sortes, et les concepts taxonomiques ont tous des problèmes spécifiques et leur vérification peut être problématique. Des exemples de comment de tels noms peuvent être traités peuvent être trouvés dans les Codes de Nomenclature International, ainsi qu'au TDWG qui rassemble des standards semblables à ceux de l'HISPID (Conn 1996, 2000) et Noms de Plantes dans les Bases de Données Botaniques (Bisby 1994).

Noms scientifiques

L'orthographe exacte des noms scientifiques est généralement gouvernée par un des nombreux codes de nomenclature. Cependant, des erreurs peuvent toujours se produire en les tapant, ou au travers d'ambiguïté dans les mots clés, etc. La méthode la plus simple pour s'assurer que de telles erreurs restent minimes est d'utiliser un référentiel" pendant la saisie de données. La plupart des bases de données peuvent être conçues pour incorporer soit un référentiel non modifiable, soit un référentiel qui peut être mis à jour au fur et à mesure de la saisie.

Conception des bases de données

Un des éléments-clé permettant d'être capable de maintenir une bonne qualité de données avec les noms de taxons est de décomposer le nom dans différent champs individuels (c'est à dire: atomiser les données) plutôt que de les conserver toutes dans le même champ (exemple: genre, espèces, rangs infraspécifiques, noms infraspécifiques, auteur et certitude). Les conserver dans un seul champ réduit les opportunités de bonne validation des données et de détection des erreurs et peut mener à un accroissement important des probabilités d'erreurs. Par exemple, lors de la séparation des parties genre et espèce d'un nom, chaque nom de genre n'aura besoin d'être entré qu'une seule

fois dans une base de données relationnelle (grâce à un référentiel ou un menu déroulant) et cela réduit donc les opportunités d'erreur typographiques et les fautes d'orthographe.

Les bases de données qui incluent toutes les parties du nom dans un seul champ peuvent rendre très difficile le maintien de la qualité ou le croisement avec d'autres bases de données. Cela introduit un domaine supplémentaire d'erreurs possibles et n'est pas recommandé. Certaines bases de données utilisent ces deux méthodes, un champ sous une forme complète et des champs atomisés, mais encore une fois cela fournit des opportunités d'erreurs ajoutées lors de la saisie s'ils ne sont pas automatiquement générés et si l'un est mis à jour et pas l'autre. Les champs générés automatiquement éliminent ce danger.

Deux problèmes qui doivent être pris en compte avec les données atomisées sont l'incorporation des données dans une base de données où les données sont dans un champ (exemple: importer des listes de plantes ou d'animaux), et le besoin de présenter une vue concaténée de données provenant d'une base de données atomisée, par exemple sur un site Web ou dans une publication.

Dans le premier cas- l'analyse grammaticale *(« parsing »)* de champs concaténés au sein de champs individuels n'est généralement pas un précédé aussi simple qu'il y parait. Ce n'est pas seulement un problème de noms taxonomiques, évidemment, puisque le problème survient avec des références et des informations de localité comme expliqué précédemment. A mon sens, il n'existe aucun outil simple permettant cette analyse grammaticale, mais de nombreux musées et herbiers ont déjà fait cette tâche pour leurs propres institutions, et donc des algorithmes peuvent être disponibles auprès d'elles.

Dans le second cas, la nécessité de présenter à l'extraction une vue concaténée dans les cas de présentation sur le Web ou dans des dossiers, pourrait être faite en utilisant un champ supplémentaire (généré) au sein de la base de données qui stockerait les différents champs sous forme concaténée, ou générée à la volée lors de l'extraction. Ceci est un problème qui devrait être pris en compte lors de la conception de la base de données et de son mécanisme de production de rapports . Il y a des problèmes que les communautés taxonomiques pourraient avoir besoin d'étudier plus en profondeur dans le but de développer des outils et des méthodologies simples.

Référentiels taxonomiques

Des référentiels existent pour un grand nombre de groupes taxonomiques et sont en train d'être développés par diverses institutions. Des référentiels fiables sont disponibles pour beaucoup de taxons supérieurs (Familles, Ordres, Genres), et peuvent être utilisés pour assurer l'intégrité des données dans ces champs. Il est assez peu probable qu'un référentiel détaillé couvrant tous les taxons soit créé dans un futur proche, particulièrement pour les espèces non supérieures. Cependant, il existe des référentiels (exemple: IPNI 1999, Froese et Bisby 2004) que l'on peut utiliser pour commencer. Lorsque des référentiels sont disponibles, les bases de données peuvent être conçues de telle sorte que les nouveaux noms puissent y être ajoutés. Par exemple, imaginez qu'une base de données ait un référentiel accessible sous forme de menu déroulant ou avec un système d'auto-complétion à la saisie (comme dans le cas d'EXCEL où lorsque le mot a déjà été tapé il réapparaît automatiquement lorsqu'un mot commence par les mêmes lettres).

1. Utilisez le menu déroulant pour chercher le nom
2. Il n'y est pas
3. Cliquez sur le bouton "nouveau nom"

4. Ajoutez le nouveau nom
5. La base de données pourrait vous répondre que “Ce nom est similaire à <nom>” voulez vous continuez?
6. Oui
7. Le nom est ajouté à la liste, et la prochaine fois que vous souhaitez ajouter un nom, ce nom apparaîtra alors dans le menu déroulant.

De cette manière, vous agrandissez et améliorez progressivement le fichier de référence.

Dans le cas d'une vérification supplémentaire, ces noms pourraient ensuite aller dans une liste secondaire qu’un superviseur pourrait vérifier puis approuver ou écarter. En fonction du niveau de sophistication de la base de données, la liste pourrait inclure des synonymes et si vous commencez à taper un nom, le menu pourrait vous demander si vous voulez réellement ajouter ce nom étant donné qu’il est déjà listé dans le référentiel en tant que synonyme de <nom>.

Il est recommandé que le référentiel soit utilisé partout où cela est possible. Un bon moyen de commencer est Species2000 & ITIS Catalogue of Life (Froese and Bisby 2004) disponibles sur CD en temps que référentiel actualisé chaque année. Le format de ce document est en train d’être amélioré afin de rendre plus facile l’incorporation dans les bases de données. Le référentiel est également disponible sous forme électronique pour vérifier les taxons individuels et s’ajoute aux référentiels régulièrement actualisés, qui sont également disponibles en ligne. De plus, un certain nombre de noms de bases de données existent ou sont en train d’être développées et elles peuvent former la base d’un référentiel en matière de noms. Certains exemples incluent,

Des listes globales telles que:

- Species2000 & ITIS Catalogue of Life *(Froese and Bisby 2002),*
- Ecat (GBIF 2003b),
- International Plant Name Index (IPNI 1999);
- Global Plant Checklist (IOPI 2003).

- Des listes régionales telles que:
- Integrated Taxonomic Information System (Ruggiero 2001);
- Australian Plant Name Index (Chapman 1991, ANBG 2003);
- Proyeto Anthos – Sistema de información sobre los plantas de España (Fundación Biodiversidad 2005)
- Australian Faunal Directory (ABRS 2004);
- Med Checklist (Greuter *et al.* 1984-1989).

Des listes taxonomiques telles que:

- ILDIS World Database of Legumes (Bisby *et al.* 2002);
- Fishbase (Froese and Pauly 2004);
- World Spider Catalog (Platnik 2004);
- et beaucoup d’autres

Lorsque les référentiels sont importées à partir d’une source externe telle que celles citées ci-dessus, le numéro identifiant de la source devrait être enregistré dans la base de données de telle sorte que les changements effectués entre les différentes éditions des sources autoritatives (*«authority sources »*) puissent être facilement incorporées dans la base de données et que celle ci puisse donc

être mise à jour. Heureusement, ce processus pourrait être facilité rapidement grâce à l'utilisation de *Globally Unique Identifiers (GUIDs)*[1].

Doubles saisies

Même lorsque l'on conçoit une base de donnée à partir de zéro et que l'on essaie de la normaliser autant que possible en utilisant par exemple des tableaux de référence, le problème des doubles saisie ne peut être évité, en particulier lorsque l'on importe des données à partir d'une source secondaire (ex: noms ou références). Afin de retirer (ou d'indiquer) ces doubles saisies, une interface spéciale peut être nécessaire. L'interface devrait être capable d'identifier des doublons potentiels en utilisant des algorithmes spéciaux. L'opérateur de saisie des données (ou conservateur, expert, etc.) devra alors décider parmi la liste des doublons potentiels, lesquels sont de véritables doublons et lesquels devraient être conservés. Les systèmes s'en débarrasseront et les archiveront alors, ou indiqueront les données superflues tout en conservant l'intégrité référentielle du système. Des logiciels génériques pourraient être mis en place afin de rectifier ceci, mais comme dans le cas des logiciels d'analyse grammaticale, ne semblent pas exister pour le moment. Les concepteurs de bases de données relatives à la biodiversité devraient être conscients du problème et envisager de concevoir un outil logistique générique pour ces fonctions.

Vérification des erreurs

Il est possible de mener à bien certaines vérifications automatisées sur les noms. Bien que des listes complètes de noms d'espèces n'existent pas, des listes de familles ou de genre (ex: IAPT 1997, Farr et Zijlstra n.dat.) sont beaucoup plus complètes, particulièrement pour certains taxons. Des contrôles portant sur ces champs pourraient être menés en les confrontant avec de telles listes. Avec l'épithète des espèces (la deuxième partie du binomial) il y a toute une série de tests qui peuvent être conduits. Par exemple, en regardant les noms du même genre comportant un degré élevé de similitude, les noms dont l'un des caractères a été déplacé ou dont l'un ou plusieurs des caractères est manquant, etc. Le *CRIA Data Cleaning system* (CRIA 2005) a entrepris beaucoup de ces tests de ce type sur des données distribuées obtenues grâce à *speciesLink* (CRIA 2002). Les meilleures pratiques dans ce cas seraient par détection automatique, mais pas par correction automatique. D'autres vérifications possibles incluent (à partir de English 1999, Dalcin 2004):

Valeurs de données manquantes – Ceci implique de chercher des champs vides où il devrait y avoir des valeurs. Par exemple, dans une base de données botanique si un nom infraspécifique est cité, alors une valeur doit être présente mentionnant le rang infraspécifique correspondant. Ou si un nom d'espèce est présent, alors un nom de genre correspondant devrait également être présent.

Valeurs de données incorrectes – Ceci implique de chercher des erreurs typographiques, de transpositions de mots-clés, des données entrées au mauvais endroit (ex: l'épithète d'une espèce entré dans le champ du genre du nom), et des valeurs de données entrées impérativement dans un champ qui demandait une valeur (c'est à dire: champ obligatoire) mais pour lesquelles l'opérateur de saisie ne connait pas la valeur et se contente donc juste d'entrée une valeur aléatoire. Il y a un certain nombre de façons de vérifier certaines de ces erreurs- par exemple en utilisant Soundex, (Roughton et Tyckoson 1985), Phonix (Pfeiffer *et al.* 1996), ou Skeleton-Key (Pollock et Zamora 1984). Chacune de ces méthodes utilise des algorithmes légèrement différents afin de détecter les similitudes. Un test récent sur le nombre des méthodes (incluant celles mentionnées) utilisant des noms d'espèce et un ensemble de données (Dalcin 2004) a démontré que la méthode Skeleton-Key

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

produisait la plus grosse proportion de vraies erreurs et de fausses erreurs dans les ensembles de données testés. Un exemple en ligne montrant comment utiliser ces méthodes peut être visionné sur le site du CRIA au Brésil (CRIA 2005). Ce point est développé ci-dessous.

Valeurs de données non-atomiques – Ceci implique de chercher les champs où un ou plusieurs faits ont été entrés. Par exemple "subsp. Bicostata" dans le champ infraspécifique où cela devrait être divisé en deux champs. En fonction de la conception des bases de données (voir ci-dessus), cela peut être ou ne pas être une erreur. Les valeurs des données non-atomiques se retrouvent dans de nombreuses bases de données et sont difficiles à retirer. Un premier pas essentiel est que de telles valeurs indiquent que la base de données nécessite probablement la création d'un nouveau champ. Certaines données non-atomiques peuvent être séparées dans les champs appropriés en utilisant des méthodes automatisées, mais le plus souvent ce n'est pas le cas et beaucoup sont laissées et ne peuvent être corrigées que manuellement sous le contrôle d'un expert.

Schizophrénies de domaine – Ceci implique de repérer des champs utilisés à des fins pour lesquelles ils n'avaient peut-être pas été conçus. Cela arrive souvent lorsqu'un champ sur la certitude des données n'a pas été inclus dans la base de données et que des points d'interrogation, des incertitudes telles que « cf. », « aff. » sont ajoutées dans le même champ que l'épithète des espèces, ou que des commentaires sont ajoutés (tableau 2). La nature de ces 'erreurs' dépend aussi de la conception de la base de données.

Famille	Genres	Espèces
Myrtaceae	Eucalyptus	globulus
Myrtaceae	Eucalyptus	globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?
Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	À déterminer

***Table 2**. Exemples de champs schizophrènes (Chapman 2005a).*

Cas des doublons- Ceci implique de chercher des noms pouvant référer au même objet dans le monde réel. Il y a deux types de doublons pouvant se produire ici. La première est une erreur due à des fautes d'orthographe et la deuxième se situe là il y a plus d'un nom alternatif valide comme dans le *Code of Botanical Nomenclature* (2000) qui permet d'alterner entre les noms de famille (ex: Brassicaceae/Cruciferae, Lamiaceae/Labiatae). Cette dernière peut être traitée, soit en utilisant une des alternatives valides de la base de données, soit en utilisant un synonyme lié, en fonction de la politique de l'institution. Des problèmes similaires peuvent se produire lorsque des classifications alternatives ont été suivies pour des rangs taxonomiques plus grands, ou également lorsque le niveau du genre où une espèce pourrait apparaitre validement dans plus d'un seul genre, en fonction de la classification suivie (*Eucalyptus/Corymbia*; Espèces d'albatros dans l'Hémisphère Sud, espèces de petits chats sauvages, et bien d'autres).

Valeurs de données incohérentes – Cela se produit quand deux bases de données liées n'utilisent pas les mêmes noms de listes, elles montrent des incompatibilités quand elles sont croisées (ou comparées). Par exemple, cela peut se produire dans des jardins botaniques entre les spécimens vivants et les herbiers, lorsque l'on fusionne les bases de données de deux spécialistes, ou dans les musées entre les bases de données des collections et celles des images. Corriger cela implique de vérifier et de confronter les bases de donnée entre elles afin d'identifier les incohérences.

Dalcin (2004) a mené un certain nombre d'expériences détaillées sur les méthodes de vérifications des erreurs orthographiques des noms scientifiques et a développé une série d'outils visant à vérifier les similitudes phonétiques. Je n'ai pas utilisé ou testé ces outils, mais les détails de ces méthodes et leurs résultats ainsi que les tests comparatifs entre ces méthodes peuvent être obtenus chez Dalcin (2004) pp. 92-148. De plus, le CRIA au Brésil a développé des protocoles de vérification des noms en suivant des principes similaires (CRIA 2005) qui sont traités plus en détails ci-dessous dans la section sur les méthodes-.

Noms communs

Il n'y a pas de règles fermes ou rapides pour les noms "communs" ou vernaculaires, qu'ils soient en Portugais, en Espagnol, en Anglais, en Hindi, dans une toute autre variété de langues ou même qu'il s'agisse de noms indigènes relatifs à une région. Souvent, ce qu'on appelle noms "communs" sont en réalité des noms familiers (particulièrement en botanique) et ont pu simplement être inventé à la suite d'une traduction d'un nom scientifique en latin. Dans certains groupes, par exemple chez les oiseaux (voir Christidis et Boles 1994) et chez les poissons (Froese et Pauly 2004), des conventions ont été acceptées et la recommandation de noms anglais a été mise en place. Dans de nombreux groupes le même taxon peut avoir beaucoup de noms communs, qui sont souvent spécifiques à une région, une langue ou à un peuple. Il y a l'exemple de l'espèce *Echium plantagineum* qui est différemment connue sous les noms de "Malédiction de Paterson" (*Paterson's Curse*) dans un état d'Australie et sous le nom "Salut de Jane" (*Salvation Jane*) dans un autre état, et sous d'autres noms (ex: *Viper's Bugloss, Salvation Echium*) dans d'autres langues ou pays. À l'inverse, le même nom commun peut être donné à de multiples taxons, parfois dans différentes régions, mais parfois au sein de la même région.

Il est tout simplement presque impossible de standardiser les noms communs, même au sein d'une seule langue sauf peut être lorsqu'il s'agit de petits groupes. Mais y a t'il un sens à essayer de faire cela (Weber 1995)? Les vrais noms communs sont des noms qui se sont développés et qui ont évolué à travers le temps, et c'est pour que les gens puissent communiquer qu'ils existent. Ce que je suggère ici est que les noms communs ne soient pas standardisés mais que quand ils sont placés dans les bases de données que cela soit fait de façon standardisée et que leurs sources soient documentées. De nombreux utilisateurs de données primaires en matière d'espèce veulent pouvoir accéder aux données grâce à l'utilisation de noms communs. Il est donc important de les avoir dans nos bases de données si nous voulons être sûr que nos données soient le plus utiles possibles et au plus grand nombre. En adoptant des méthodes standardisées pour recoder les noms communs, que ce soit par espèce ou par centaines, dans une langue ou dans plusieurs, en documentant la source de chaque nom, nous pouvons rendre la recherche et donc la récupération d'informations bien plus efficace et utile.

Il y a de multiples difficultés lorsqu'il s'agit d'inclure des noms communs dans les bases de données d'espèce. Ceci inclut:

- des noms dans une langue autre que l latin qui nécessitent l'utilisation d'Unicode dans la base de données lors du stockage. Des problèmes peuvent se produire :
    - quand les bases de données tentent d'enregistrer les noms phonétiquement en utilisant simplement l'alphabet latin,
    - quand les gens ne sont pas capables d'afficher correctement les caractères de leur écran ou de leur feuille,

    - en menant des recherches de données où les utilisateurs n'utilisent qu'un clavier "latin",
    - lors de la saisie de données où les noms sont un mélange de latin et d'autres langues,
- le besoin de stocker les informations relatives aux langues des noms, particulièrement lorsque les noms en plusieurs langues sont mélangés,
- le besoin de stocker des informations relatives aux facteurs régionaux- zone géographique pour laquelle le nom peut être pertinent, les dialectes, etc.
- le besoin de stocker les informations telles que les références des sources du nom.

Il ne s'agit pas d'une tâche facile et particulièrement d'une manière qui intensifie l'utilité tout en réduisant le taux d'erreurs. S'il est décidé d' inclure de tels noms, ce qui suit pourrait être utile afin de fournir un certain degré de standardisation.

Saisie de données

Lors de la saisie en base de données des noms communs, il est recommandé qu'une forme de cohérence de construction soit suivie. Il est probablement plus important que chaque base de données individuelle soit intérieurement cohérente. Le développement de standards nationaux ou régionaux est recommandé lorsque cela est possible. Il y a beaucoup trop de variations de langues et de régions pour essayer de développer des standards pour toutes les langues et tous les taxons, même si certains des concepts proposés ici pourraient former la base d'un standard de langues différentes de celles couvertes.

Pour les noms communs en anglais et en espagnol, je recommande qu'une convention similaire à celle développée dans Environment Australia (Chapman *et al.* 2002, Chapman 2004) soit suivie. Ces directives ont été développées afin de soutenir la cohérence de nombreuses bases de données de cet organisme. Ces conventions incluent de commencer chaque mot par une majuscule.

Sunset Frog[2]

Dans le cas de noms généralisés ou groupés, l'utilisation d'un trait d'union est recommandée. Le mot suivant le trait d'union ne commence généralement pas par une majuscule, sauf pour les oiseaux où c'est alors le cas s'il s'agit du nom d'un groupe plus large, comme recommandé par Christidis et Boles (1994).

Yellow Spider-orchid
Double-eyed Fig-Parrot ('Pscitaculle double-œil') ('Parrot' commence par une majuscule parce qu'il s'agit du groupe des Parrots)

Les noms portugais sont généralement écrits en minuscules, souvent avec des traits d'union entre chaque mot s'il s'agit de noms, ou séparés par un espace s'il s'agit de noms ou d'adjectifs. Lorsqu'il s'agit de noms communs portugais, il est recommandé que cette convention soit suivie ou modifiée pour se conformer aux exemples anglais et espagnols

mama-de-cadela,

[2] Sunset Frog (« Grenouille du Crépuscule » ou *Spicospina Flammocaerulea*), espèce (et genre) de grenouille endémique à l'extrême sud-ouest de l'Australie

fruta-de-cera
cedro vermelho

Il y a quelques désagrément et de la confusion quand à l'usage d'apostrophes avec les noms communs. Pour les noms géographiques, il y a une tendance grandissante à ignorer les apostrophes (ex: Smiths Creek, Lake Oneill), et c'est maintenant une pratique acceptée en Australie (ICSM 2001, Geographic Names Board 2003 Art. 6.1). Je recommande qu'une convention similaire puisse être adoptée avec les noms communs, même si cela n'est pas une obligation à l'heure actuelle.

Là où les noms sont ajoutés dans plus qu'une seule langue et/ou qu'ils varient d'une région à l'autre, alors la langue et l'information régionale devraient être inclues de telle sorte qu'elles puissent être rattachées au nom. La meilleure façon de le faire est d'utiliser une base de données relationnelle en reliant les champs régionaux et linguistiques supplémentaires. Dans certaines bases de données, lorsqu'il n'y a seulement qu'une différence de langue, la langue est souvent notifiée entre parenthèses après le nom, mais bien que cela puisse apparaitre une solution assez simple au départ, cela devient souvent de plus en plus compliqué au fil du temps, et même irréalisable sur le long terme. Il vaut mieux concevoir la base de données dans le but de gérer ces problèmes dès le départ plutôt que d'avoir à ajouter de la flexibilité par la suite.

Si des noms issus d'alphabet non-latin sont ajoutés à la base de données, alors la base de données doit être conçue pour permettre d'autoriser l'inclusion des jeux de caractères UNICODES.

Vérification des erreurs

Comme les noms communs sont souvent liés à un nom scientifique, des vérifications peuvent être menées de temps en temps pour s'assurer de la cohérence au sein de la base de données. Ceci peut être une procédure fastidieuse mais qui n'a besoin pas besoin d'être menée à intervalles réguliers. Les contrôles peuvent être effectués en extrayant tous les cas uniques et en vérifiant les incohérences, ex: les traits d'union manquant.

Une fois encore, des programmes tels que Soundex, (Roughton et Tyckoson 1985), Phonix (Pfeiffer *et al.* 1996), ou Skeleton-Key (Pollock et Zamora 1984) peuvent être utilisés pour rechercher des erreurs typographiques telles que les transpositions de caractères comme mentionné précédemment pour les *noms scientifiques* (voir Dalcin 2004).

Classement infraspécifique

L'utilisation de champs de classification infraspécifiques est un élément plus important dans les bases de données de plantes que dans celle d'animaux. Les taxonomistes animaliers n'utilisent généralement qu'un rang de classification en dessous des espèces, c'est à dire les sous-espèces (et cela est même utilisé de moins en moins), en traitant le nom de manière trinomiale :

Stipiturus malachurus parimeda.

Historiquement, cependant, certains taxonomistes animaliers ont utilisé d'autres rangs de classification que les sous-espèces, et les bases de données peuvent avoir besoin de pourvoir aux besoins que cela engendre. Si c'est le cas, alors les remarques faites à propos des bases de données végétales ci-dessous s'appliqueront également aux bases de données de noms d'animaux.

Conception de la base de données

Comme déjà mentionné, il y a d'importants avantages concernant la qualité des données lorsque l'on sépare le rang infraspécifique du nom infraspécifique. Ceci permet des vérifications aisées du rang et rend la vérification du nom infraspécifique beaucoup plus simple aussi. Le rang et le nom ne devraient jamais être inclus en tant que contenu dans le même champ.

Un problème qui devrait être pris en compte dans le cas de bases de données atomisées est le besoin dans certains cas de concaténer ces champs pour l'affichage en ligne, etc. Ceci peut généralement être automatisé, mais la façon dont cela devrait être fait (que ce soit dans la base de données en tant que champs additionnel automatiquement généré ou bien une valeur créé sur le moment) devrait être prise en considération lors de la conception de la base de données.

Saisie de données

Avec les plantes (et historiquement avec les animaux), il y a différent niveaux de classification des espèces qui peuvent être utilisées. Ces rangs infraspécifiques sont le plus souvent : ***sous-espèces, variété, sous-variété, forme et sous-forme*** (le Code Botanique n'exclut pas d'insérer des rangs de classification supplémentaires, de telle manière qu'il est possible que d'autres rangs existent déjà dans certaines bases de données). Les sous-variétés et les sous-formes sont rarement utilisées, mais nécessitent néanmoins de pouvoir être traitées dans les bases de données. À nouveau, une liste de choix devrait être constituée avec un nombre de choix limités. Si ce n'est pas fait, des erreurs pourraient alors commencer à s'insinuer, et vous verrez invariablement des sous-espèces nommées: sous-espèces (subspecies), subsp., ssp., subspp., etc. Ceci peut être un cauchemar pour quelqu'un qui essaierait d'extraire les données, ou de mener à bien une vérification d'erreurs. Il vaut mieux réduire les options au moment de la saisie que de devoir pourvoir à tous les rangs au moment de l'extraction des données, ou d'essayer de nettoyer la base de données pour appliquer une cohérence ultérieurement. Il est recommandé que ce qui suit soit utilisé:

subsp. subspecies (sous-espèce)
var. variety (variété)
subvar. subvariety (sous-variété)
f. form/forma (forme)
subf. subform (sous-forme)
cv. cultivar (culture) (souvent traité dans les bases de données en tant qu'autre rang de classification, mais voir commentaire séparé plus bas)

Dans les collections de bases de données, l'inclusion d'une hiérarchie n'est pas nécessaire quand plus d'un seul niveau pourrait exister, parce que cela ajoute simplement un couche de confusion supplémentaire, et dans le Code International pour la Nomenclature Botanique, *Code for Botanical Nomenclature,* (2000) la hiérarchie est inutile pour définir le taxon de manière non ambiguë. Si la hiérarchie est inclue, il doit être possible d'extraire seulement celle nécessaire à la désambiguïsation du taxon.

Leucochrysum albicans subsp. *albicans* var. *tricolor* (= *Leucochrysum albicans* var. *tricolor*).

Vérification des erreurs

Si la base de données a bien été conçue et qu'une liste de vérification des valeurs est utilisée, il n'y a alors aucun besoin d'effectuer d'autres contrôles. Lorsque cela n'est pas le cas, par contre, des

contrôles devraient être menés afin de s'assurer que seules les sous-ensembles des valeurs autorisées apparaissent. Le contrôle qui devrait cependant toujours être effectué est celui des valeurs manquantes.

Cultivars et Hybrides

Les cultivars et hybrides de plantes apparaissent souvent dans les bases de données mais ne sont pas souvent bien maniées. Les cultures sont des sujets avec leur propres Codes de Nomenclature (Brickell *et al.* 2004). Dans beaucoup de bases de données d'espèces de plantes, elles sont traitées simplement en tant qu'un autre rang infraspécifique ("cv") et dans certaines bases de données cela peut être tout à fait acceptable. Les hybrides sont bien plus compliqués à manier que la plupart des autres groupes. Il peut leur être donné un nom binomial et ils peuvent ensuite être traités comme n'importe quel autre taxon du même rang (précédé par un "**X**" -symbole de multiplication "fois"- pour indiquer un hybride), ou ils peuvent être traités sous fome de formule, *formula*, (un croisement entre deux taxons qui pourraient même appartenir à deux rangs différents) et indiqué par des noms taxonomique séparés par des signes de multiplication.

Conception des bases de données

Je recommande que toute personne cherchant à mettre en place une base de données de plantes qui pourrait inclure des hybrides ou des cultures consulte le standard HISPID (Conn 1996, 2000) où les hybrides sont traités comme part du Groupe d'Identification des Enregistrements, *Record Identification Group*. Quelle que soit la façon dont ils sont maniés, il est bon d'inclure un champ qui indique que le nom appartient à une cultivar ou à un hybride. De cette manière, ils peuvent être extraits séparément et traités différemment (pour le formatage, la concaténation, etc.) à l'extraction et pour la vérification des erreurs.

Vérification des erreurs

La vérification des erreurs en matière d'hybrides et de cultivars est une tâche compliquée si la base de données n'a pas été conçue pour répondre à ce besoin. Une façon de procéder pourrait être de les traiter en tant que groupe. Par exemple, extraire tous les hybrides et les trier alphabétiquement par espèce en fonction de la façon dont elles sont rangées dans la base de données. Cela est beaucoup plus simple lorsque qu'un champ séparé est inclus et identifie les enregistrements d'hybrides en tant que tels. Une erreur-clé qui risque souvent de se produire est l'inconsistance dans l'utilisation du signe '**X**'. Certaines bases de données n'autorisent peut être pas le signe "multiplié" et le remplacent souvent par un "x" ou un "X", parfois avec un espace avant le nom, mais pas toujours. Ces sortes d'inconsistances peuvent facilement être vérifiées. Je ne connais aucun système réellement fiable pour vérifier les erreurs de noms d'hybrides.

Noms non publiés

Saisie de données

Tous les enregistrements placés dans les bases de données primaires d'espèces (« primary species database ») ne correspondront pas à un taxon publié validement. Afin d'être capable de retrouver ces enregistrements dans les bases de données il est nécessaire de fournir un nom "temporaire" à cette série. Si des noms non publiés peuvent être incorporés dans une base de données dans un format standard, cela rend beaucoup plus facile le fait d'en garder une trace et de les retrouver

ultérieurement. C'est aussi mieux et moins déroutant que d'ajouter des noms non publiés qui sont des binomiaux et qui ressemblent à des noms publiés, avec ou sans des étiquettes telles que "nomen novum", "nom. nov." et "ms". Trop souvent, la partie "ms" ou "nom. nov." est laissée de côté et les utilisateurs peuvent passer beaucoup de temps à chercher la publication et des informations de référence pour un nom non publié. En utilisant cette formulation, il est évident pour tout le monde qu'il s'agit d'un nom non publié.

En Australie dans les années 80, les botanistes se sont mis d'accord sur une formulation (Croft 1989, Conn 1996, 2000) à utiliser pour les noms non publiés. Ceci était pour éviter la confusion pouvant découler de l'usage de choses comme "*Verticordia* sp.1", "*Verticordia* sp.2" etc. Une fois que les bases de données commencent à être croisées, par exemple l'Herbier virtuel australien (CHAH 2002), *species*Link (CRIA 2002), Biological Collection Access Service for Europe (BioCASE 2003), le Mammal Networked Information System (MaNIS 2001), le GBIF Portal (GBIF 2004) et bien d'autres, ce genre de nom peut créer encore plus de confusion puisqu'il n'y a aucune garantie que ce qui était nommé "sp.1" dans une institution est identique à ce qui est nommé "sp.1" dans une autre. Une manière de garder ces bases de données propres et cohérentes, et de permettre un transfert de données de l'une à l'autre en douceur, est au travers de l'utilisation d'une formulation semblable à celle adoptée par la communauté botanique australienne.

La formulation se présente sous cette forme :

> "<Genus> sp. <colloquial name or description> (<Voucher>)":
>
> *Prostanthera* sp. Somersbey (B.J.Conn 4024)

Plus tard, lorsque le taxon est décrit et nommé de façon formelle, le nom de la formulation peut être traité en tant que synonyme, comme n'importe quel autre synonyme.

L'utilisation d'une telle formulation rend une base de données plus compliquée qu'elle ne pourrait l'être, parce qu'à la place du champ d'espèce ne comportant toujours qu'un seul mot, l'utilisation de la formulation requiert l'inclusion d'une phrase. L'utilisation de "sp.1" "nom.nov." etc... comme c'est souvent le cas, mène au même problème et cette méthode laisse de la place aux ambiguïtés. Cette utilisation de cette formulation peut causer des difficultés de concaténation (par exemple sur le Web, etc...) mais néanmoins, l'expérience en matière d'utilisation de cette méthodologie en Australie (voir par exemple l'utilisation de la base de données SPRAT du Département d'Environnement Australien, *Australian Department of the Environment* (DEH 2005b)) a prouvé son efficacité. De toute autre manière, la formulation est traité comme un épithète d' "espèce" bien qu'il puisse y avoir des espaces ou des parenthèses, etc.

De par le besoin d'utiliser des noms non publiés, par exemple dans des listes légales d'espèces menacées (voir par exemple DHE 2005a), il est essentiel d'avoir une cohérence du système lors du nommage ou de l'étiquetage de ces taxons pour des publications non taxonomiques, par exemple dans le cas d'instruments législatifs. En utilisant une formulation comme celle suggérée, il y a peu de risques d'accidentellement publier un nomen nudum par erreur.

Il est recommandé que les musées et les herbiers adoptent un système similaire à utiliser dans leurs bases de données.

Vérification des erreurs

L'erreur la plus commune se produisant avec une telle formulation est liée à une mauvaise orthographe. Du fait que cette formulation inclut plusieurs mots, il est souvent facile de faire des

erreurs de citations ou de rédaction d'un reçu, etc. Le moyen le plus facile de vérifier de tels noms est de les trier au sein d'un genre (ils devraient alors être le seul nom de plus d'un mot dans le champ d'espèce ou d'infra-espèce) et de rechercher les similitudes. Ceci ne devrait pas être une tâche pénible puisqu'il y a peu de chances qu'il y en ait beaucoup au sein d'un seul genre. Des techniques semblables telles que Soundex, mentionnée plus tôt, peuvent être utilisées.

Noms d'auteurs

Les auteurs du nom d'une espèce peuvent être inclus dans certaines bases de données de spécimens, mais le plus souvent, leur inclusion peuvent mener à des erreurs puisque qu'elles sont rarement vérifiées à fond avant insertion. Ils ne sont vraiment nécessaires que lorsque le nom aurait pu être donné par inadvertance à deux taxons différents (homonymes) membres du même genre ou quand la base de données essaye d'inclure des concepts taxonomiques (Berendsohn 1997). L'insertion du nom de l'auteur à la suite de celui des espèces (ou infra-espèces) peut aider à faire la différence entre deux noms ou concepts. Si les bases de données incluent les noms d'auteurs et d'espèces, ceux-ci doit alors impérativement être inclus dans des champs séparés du nom des espèces.

La concaténation des données où le nom de l'auteur et les dates sont gardées séparément n'est souvent pas un problème majeur sauf dans le cas des plantes à autonymes (voir ci-dessous). Les bases de données mélangeant des plantes et des animaux, peuvent cependant causer quelques difficultés là où les autorités sont traitées de manière un peu différente. Cela ne devrait pas causer beaucoup de problèmes si les champs réservés à l'auteur on été prévus à cet effet mais les règles d'extraction peuvent différer pour les différent Royaumes taxonomiques.

Dalcin (2004) considère que l'autorité est un élément de l'étiquette en dessous du domaine de ses données de nomenclature.

Saisie de données

Pour les animaux, le nom de l'auteur (souvent en toutes lettres) est toujours suivi d'une année; pour les plantes, le nom de l'auteur ou l'abréviation sont donnés seuls.

Animaux:
Emydura signata Ahl, 1932
Macrotis lagotis (Reid, 1937)
(les parenthèses indiquent que Reid a attribué les espèces à un différent genre)
Plantes:

Melaleuca nervosa (Lindley) Cheel
Synonyme : *Callistemon nervosus* Lindley
(Lindley l'avait originellement décrite comme *Callistemon ;* Cheel l'a ensuite transféré au genre *Melaleuca*).

Dans le cas de plantes, les termes " ex " ou " in " peuvent occasionnellement être trouvé parmi les noms d'auteurs. L'auteur devant " ex " (« le pré-'ex' auteur ») est celui qui a fourni le nom mais n'a pas rempli les pré-requis nécessaires à une publication valide, ou qui a publié le nom avant la date de début nomenclaturale du groupe concerné. Un auteur mentionné après 'in' est celui dans le travail duquel une description ou un diagnostic fournis par d'autres auteurs est publié. Pour plus d'explications sur les auteurs mentionnées avant 'ex' et après 'in' et leur utilisation, voir les Articles 46.2 and 46.3 du Code International de Nomenclature Botanique, *International Code of Botanical*

Nomenclature (2000). Si les noms d'auteurs sont utilisés dans des bases de données ils devraient être séparés des champs des noms (voir discussion sur l'atomisation ci-dessus) et il est recommandé que ni le pré-'ex' ou le post-'in' ne soient cités.

> Green (1985) a attribué la nouvelle combinaison *Tersonia cyathiflora* à "(Fenzl) A.S. George"; puisque Green ne mentionne nulle part que George a contribué d'une façon ou d'une autre, l'auteur double doit être cité ainsi: "A.S.George ex J.W.Green" ou préférablement même, ainsi: "J.W.Green".
>
> *Tersonia cyathiflora* (Fenzl) J.W.Green
>
> Dans la deuxième édition de *Hortus Kewensis* (1813) de W.T.Aiton, beaucoup de descriptions sont signées Robert Brown, et donc il peut être considéré que Brown a décrit les espèces. L'auteur du nom est souvent cité tel que "R.Br. in Ait." Il est recommandé, cependant, qu'il soit simplement cité ainsi: "R.Br."
>
> *Acacia acuicularis* R.Br.

Avec les plantes -pour le type de sous-espèces ou de variétés, etc. - où le nom infraspécifique est le nom de l'espèce elle même (autonymes), l'auteur du nom de l'espèce est utilisé et suit l'épithète. Ce format cause généralement beaucoup de confusion pour reconstruire les noms dans les bases de données de spécimens qui incluent des noms d'auteurs, puisque cela est une exception à d'autres règles.

> *Leucochrysum albicans* (A.Cunn.) Paul G.Wilson subsp. *albicans*

Pour les plantes, l'abréviation du nom des auteurs doit suivre un standard adopté internationalement (Brummitt et Powell 1992), et cette publication peut être utilisée pour mettre en place une check-list, ou pour la saisie de données et/ou la validation après vérification.

> A.Cunn. = Allan Cunningham
> L. = Linnaeus
> L.f. = Linnaeus filius (fils de-)

Parfois, un espace est placé entre l'initiale et le nom de famille. C'est une affaire de goût.

Vérification des données

Les noms d'auteurs comme utilisés chez Brummit et Powell (1992) pourraient être utilisés pour vérifier les auteurs dans les bases de données botaniques. L'université d'Harvard a également préparé un fichier téléchargeable de noms d'auteurs et l'a rendu accessible en ligne[2]. Ceci devrait s'avérer être un fichier très utile pour la vérification des noms d'auteurs et les dates. Certaines bases de données incluent aussi des noms d'auteurs (exemple, IPNI 1999, Froese et Bisby 2002). Une fois encore, des techniques semblables à celles de Soundex comme mentionnées ci-dessus pourraient être utilisées pour chercher des similitudes entre les deux noms. C'est la combinaison du

[2] http://www.huh.harvard.edu/databases/cms/download.html

nom de l'espèce et de celui de l'auteur qui est le facteur décisif, mais cela n'est pas toujours facile à vérifier.

Si les auteurs sont utilisés, alors tous les noms publiés dans la base de données devraient avoir un auteur. Dans ces cas-là, un contrôle de valeur de données manquante, *Missing Data Values check*, devrait être mené.

Noms des collecteurs

Le nom des collecteurs sont généralement standardisés dans les collections de bases de données, même si la standardisation du nom des collecteurs de plantes se fait sous le nom de la plante dans le projet *species*Link au Brésil (Koch, 2003) et dans le *Kew Gardens* de Peter Sutton.

De longues listes de noms de collecteurs ont été publiées pour certains domaines, mais surtout pour les collecteurs botaniques (voir Steenis -Kruseman 1950, Hepper et Neate 1971, Dorr 1997, Index Herbariorum 1954-1988). Il y a également un certain nombre de ressources accessibles en ligne:

- L'université d'Harvard a récemment préparé un fiché téléchargeable de collecteurs botaniques et d'équipes de collecteurs et l'a rendu accessible en ligne. http://www.huh.harvard.edu/databases/cms/download.html

- Index Collectorum – de l'université de Göttingen http://www.sysbot.uni-goettingen.de/index_coll/default.htm

- Annuaire des Collectionneurs d'Insectes d'Afrique du Sud (Société Entomologique d'Afrique du Sud) http://www.up.ac.za/academic/entomological-society/collectr/collectr.html

- Index bio-bibliographicus notorum hominum Nonveilleriana (Société Entomologique Croate) http://www.agr.hr/hed/hrv/bibl/osobe/comentsEN.htm

Il y a également un certain nombre de copies papier de ces publications, et il y en a sous doute encore bien d'autres telles que des index disponibles dans les disciplines variées de la zoologie.

Saisie de données

Il est recommandé que les noms soient inclus dans les bases de données primaires d'espèces dans un format standard. L' HISPID Standard (Conn 2000) recommande ce qui suit:

> *Nom de famille du collecteur principal suivi d'une virgule et d'un espace (,) puis initiales (chacune en majuscule et séparées par des points). Toutes les initiales et premières lettres du nom de famille du collecteur doivent être en majuscule. Par exemple, Chambers, P.F.*

Il est recommandé que les collecteurs secondaires soient placés dans un second champ. Si ce n'est pas le cas, alors il est recommandé qu'ils soient cités en utilisant une virgule et un espace les séparant les uns des autres. Par exemple:

Tan, F., Jeffreys, R.S.

Lorsqu'il y a un risque de confusion, les autres prénoms devraient être écrits en toutes lettres. Par exemple, pour distinguer Wilson, Paul G. et Wilson, Peter G (un espace est placé après le prénom; pas de ponctuation à part pour séparer les deux noms, comme décrit précédemment).

Les titres devraient être laissés de côté.

Si le nom de famille contient une particule et un substantif, comme pour beaucoup de noms européens (ex: C.G.G.J. van Steenis), la particule est en minuscule et le substantif commence par une majuscule. Par exemple:

Steenis, C.G.G.J. van

D'autres noms de forme similaires sont de la Salle, d'Entrecastreaux, van Royen etc, il devrait être noté que, néanmoins, beaucoup de ces noms ont été anglicisés, particulièrement aux Etats-Unis, de telle sorte que les deux parties du nom de famille sont traités en substantif. Dans de tels cas, ces noms sont transférés ainsi:

De Nardi, J.C.

Les préfixes O', Mac', Mc' and M' (ex: MacDougal, McKenzie, O'Donnell) devraient tous être traités comme part du substantif et donc être transférés comme partie du nom de famille. Par exemple:

McKenzie, V.

Les prénoms comprenant un trait d'union devraient être transcrits en majuscules, avec la première et la dernière initiale séparées par un trait d'union (sans espace), et seulement la dernière doit être suivie d'un point. Par exemple:

Quirico, A-L.

Peng, C-I.

Si le collecteur des ces données est inconnu, alors le terme "Anonyme" *("Anonymous")* doit être employé.

Les informations résultant d'interprétation devraient être enfermées entre crochets, par exemple:

Anonymous [? Mueller, F.]

Vérification des erreurs

Comme mentionné ci-dessus, sans une liste standard de collecteurs, il n'est pas facile de mener à bien des vérifications d'erreurs sur les noms des collecteurs. Ceci est particulièrement vrai dans les bases de données qui ne suivent pas une pratique standard (telle que d'écrire le nom de famille en premier, comme mentionné plus haut). Si la base de donnée est standardisée, il est alors assez facile de trier tous les noms de collecteurs de la base de données et de chercher les variations légères (par exemple, un collecteur qui utilise tantôt une initiale, tantôt plusieurs). Un soin extrême doit être pris à ne pas introduire de nouvelles erreurs dans la base de données en altérant le nom du

collecteur sans certitude absolue que le changement est correct. L'exemple d'initiales ci-dessus est un cas où un changement pourrait facilement introduire de nouvelles erreurs. Les erreurs qui pourraient être corrigées sont les fautes d'orthographe dans les noms de famille, par exemple.

Une façon de développer une liste de collecteurs est de créer une liste unique des valeurs de cette base de données de la même manière que des tableaux de références sont développés pour les noms de taxons.

Les champs associés avec le nom du collecteur, *Collector-Name,* tels que "date de collection", *Date-of-Collection,* peuvent également être utilisés pour la vérification des erreurs. Des historiens ont effectué un travail énorme en développant des itinéraires d'explorateurs et de collecteurs, d'expéditions scientifiques historiques, d'itinéraires maritimes, etc. Souvent, ces expéditions n'ont pas été menées par des scientifiques, mais par des historiens, et notre science peut bénéficier grandement de ce travail (voir les ressources citées ci-dessous, avec les collections des bibliothèques (publications, journaux, etc) de beaucoup des musées et herbiers les plus vieux du monde, un travail récent de l'université de Californie à San Diego, de l'Institution d'Océanographie Scripps, *Scripps Institution of Oceanography (SIO)*, et de la Bibliothèque Nationale de Science Numérique, *National Science Digital Library*, sur la saisie et la documentation des données issues de missions maritimes dans le cadre du projet de « Bibliothèque numérique du SIO). Les liens entre ces bases de données et les bases de données primaires d'espèces peuvent conduire à corriger à la fois les incohérences et les erreurs détectées.

Données spatiales

La localisation spatiale est l'aspect le plus crucial pour être en mesure de déterminer la justesse d'emploi de nombreuses données d'occurrences d'espèces. Les études bio-géographiques liées à l'espace comprennent une des utilisations les plus larges pour ces données – des études telles que le modelage de la distribution des espèces, les études bio-géographiques, le planning et la gestion environnemental, la sélection des réserves et planning des activités de conservation, et enfin l'aide aux décisions environnementales. Pour une étude détaillée, voir le document associé sur les Emplois des Occurrences des Données Primaires d'Espèce, *Uses of Primary Species-Occurrence Data* (Chapman 2005b).

Nous pensons souvent que les principales données d'espèces *(« primary species data »)* sont des mesures de ***points*** en matière d'occurrences de plantes ou d'animaux mais ceci n'est qu'une partie de l'histoire. Rares sont les zones de collectes enregistrées précisément ou de façon suffisamment fiable pour être considérés comme des "vrais" points. L'exactitude associée à la collection signifie que le point traité représente en fait une zone ou une empreinte. Par exemple, un lieu dont le contexte d'information textuelle dit "10 km au nord de Campinas" (ville du Brésil) impose une justesse associée à la distance "10 km" (à peut être plus ou moins 500m), une justesse associée à la direction "nord" (par exemple, le nord est entre le nord-ouest et le nord-est), et une justesse associée à "Campinas" (s'agit il des limites de la ville -et donc d'un polygone-? Du centre ville? Etc). Pour une discussion plus précise, voir Wieczorek 2001a, Wieczorek *et al.,* 2004. De plus, beaucoup de données d'observations et d'études sont enregistrées à partir d'une surface (***polygone***) telles que des observatoires d'oiseaux au dessus d'une zone de 2 ha, ou au sein de Parcs Nationaux, ou à partir d'une ***grille*** de marquage classique où un carré équivaut à dix minutes pour toute l'Australie (Blakers *et al.* 1984) ou des grilles d'études de terrain de type 10 m X 10 m, ou grâce à des coupes transversales *(transect ou **ligne**)* comme par exemple le long d'une route ou d'une rivière (même si cela est généralement mieux traité en tant que polygone dérivant d'une route ou d'une rivière, en fonction de l'échelle). Pour plus d'informations voir *Visualisation des Erreurs*, ci-dessous.

Comme mentionné précédemment, un certain nombre de programmes pouvant aider à vérifier et tester les erreurs d'attribution des données géographiques attachées aux données primaires d'espèce existent déjà. D'autres outils sont accessibles pour assister l'assignation originale des coordonnées des données depuis l'information de localisation (telle que la distance et la direction à partir d'un lieu nommé).

Les tests d'erreurs pour des références géographiques ayant déjà été assignées incluent:

- Vérifier en comparant avec d'autres informations internes provenant de l'enregistrement lui-même. Par exemple, un état ou un district nommé.
- Vérifier à partir d'une référence externe en utilisant une base de données. Par exemple, voir si l'enregistrement est cohérent avec les lieux de collecte connus pour le collecteur.
- Vérifier à partir d'une référence externe en utilisant un GIS, en faisant des tests vérifiant si un point est bien situé à l'intérieur du polygone censé le contenir *(« point-in-polygon »)* - que la donnée tombe bien dans les terres et non plus pas dans la mer par exemple.
- Chercher des valeurs extrêmes *(points exagérément excentrés)* dans les espaces géographiques, ou,
- Chercher des valeurs extrêmes dans les espaces environnementaux.

Saisie de données et Géoréférencement

Comme mis en valeur tout au long de ces documents, la prévention des erreurs est préférable à leur détection, et le géoréférencement ou l'attribution de codes géographiques à ces informations est l'une des plus grandes sources d'erreur dans la mise en base de données des occurrences d'espèces. De nombreux nouveaux outils sont en voie de développement pour assister le processus d'addition de coordonnées (particulièrement en matière de longitude et de latitude) aux données primaires d'espèce. Ceci n'est cependant pas un processus facile, particulièrement parce que la plupart des données anciennes (ancienne données de musées ou d'herbiers rassemblées depuis 300 ou 400 ans) ne comportent que peu d'informations géographiques autres qu'une description générale de l'endroit où les spécimens avaient été trouvés (Chapman et Milne 1998). Ces pièces avaient souvent été rassemblées avant que des établissements de population modernes aient été construits et nommés, et avant que les routes aient été construites. Beaucoup avaient été rassemblées à dos de cheval ou par bateau à une époque ou les derniers en date des établissements et points de références étaient souvent difficiles à déterminer. Beaucoup de ces points de référence ne sont plus présents sur les cartes modernes, et dans de nombreux cas lorsqu'ils existent toujours, ils sont souvent ambigus. Quand les codes géographiques sont donnés, ils ne sont souvent pas précis (Chapman 1999) et ont souvent été ajoutés ultérieurement (géoréférencement rétrospectif -*retrospective georeferencing,* Blum 2001) par une autre personne que le collecteur des données lui-même (Chapman 1992).

Définitions:

Avant de continuer, il y a un certain nombre de terme dont l'utilisation dans ce document nécessite une définition. Certains de ces termes sont utilisés indifféremment ailleurs, et des disciplines différentes utilisent d'autres termes pour décrire un procédé similaire.

Géocode ou code géographique: Comme utilisé dans ce document, un géocode est un code (souvent des coordonnées de type x et y) qui enregistre le positionnement dans l'espace d'une information à partir d'un système de référence standard. Ici, il sert à localiser une position à la surface de la terre. Pour les données d'occurrences d'espèces, le géocode est donné en fonction d'un des différents systèmes de référence géographique standards, dont les deux plus communs sont Universal Transverse Mercator (UTM) ou le système de longitude et latitude. Il peut être enregistré de façons variées : mètres, degrés décimaux, degrés, minutes, secondes, degrés, minutes décimales, etc... Les définitions du terme géocode sont nombreuses et générales. Dans de nombreuses applications du GIS il s'agit de faire référence à une adresse et à un code postal, dans les termes du marketing il y est fait référence en tant que caractérisation démographique d'un quartier, et dans quelques cas (Clarke 2002) le géocode traite le lieu sous une forme lisible uniquement par ordinateur. Le géocode est aussi parfois appelé ***géoréférence*** ou ***coordonnées***.

Géoréférencement: dans ce document le terme géoréférencement est utilisé pour décrire le procédé d'assignation de coordonnées géographiques à des informations liées à un lieu géographique sur terre. On l'appelle parfois ***géocodage***.

Conception des bases de données

La conception d'une base de données primaires d'occurrences d'espèces devrait garantir que des champs permettant de pourvoir aux informations souvent mal placées dans le champ de la localité, tels que l'habitat ou les informations sur le port végétal ou l'habitus *(« habit »)* ou les notes

géographiques. Un exemple de champ de distribution avec des informations mélangées (Fishbase[3] pour *Perca fluviatilis*) est:

> *"A travers l'Europe et la Sibérie jusqu'à la rivière Kolyma, mais pas en Espagne, en Italie ou en Grèce; de nombreux cas d'introduction. Plusieurs pays rapportent un impact écologique défavorable après introduction dans le milieu".*

De tels croisements dans les champs rendent le traitement des informations très difficile à automatiser par des algorithmes et ne répondent pas de façon fiable à la philosophie et à la conception des bases de données relationnelles où les informations peuvent être rangées dans des longs champs textes (parfois appelés *champs mémos)*.

Il existe une série de champs additionnels qui peuvent être ajoutés aux bases de données d'occurrences d'espèces afin d'assister le nettoyage de données et qui peuvent mener à une amélioration de la documentation de la qualité des données. De tels champs incluent :

- ***Exactitude spatiale*** – un champ qui rapporte (de préférence en mètre, mais parfois sous une forme codée) l'exactitude avec laquelle le lieu de la donnée a été déterminé.

- ***Lieu nommé, Distance et Direction*** – certaines bases de données incluent "Lieu nommé", "Distance" et Direction" dans des champs séparés de même qu'un champ de localité simple. L'insertion de tels champs peut aider dans la détermination des géocodes et dans la vérification des erreurs.

- ***Méthode de géocodage*** – un champ (ou des champs) rapportant la façon dont le géocodage a été effectué (Chapman 2005a). Ceci peut inclure :
    - l'utilisation d'un GPS différentiel ;
    - GPS portatif corrompu par une disponibilité sélective (par exemple, enregistrant avant Mai 2000) ;
    - une carte à l'échelle 1:100 000 obtenue par une triangulation utilisant des caractéristiques facilement identifiables ;
    - une référence cartographique sans échelle précise (tracée à l'œil nu – *« Dead reckoning »*) ;
    - une référence cartographique obtenue approximativement (dans un hélicoptère par exemple) ;
    - ou obtenue automatiquement par l'utilisation d'un logiciel de géoréférencement utilisant une méthode point-rayon ;
    - Ou obtenue à partir d'une base de données utilisant des localités précédemment géoréférencées.
- ***Type de géocode*** – rapporte le type de description de la localité utilisé dans la détermination du géocode.

  Dans un document sur la méthode de géoréférencement point-rayon, Wieczorek et al. (2004) fournissent un tableau de neuf types de description de localités trouvées dans des collections d'histoire naturelle. Les trois premières qu'ils recommandent ne doivent pas être géoréférencées mais une annotation doit figurer pour justifier de cette absence de géoréférencement. Certaines bases de données utilisent un centroïde avec un champ de

[3] http://www.fishbase.org/

précision immense (par exemple 100,000,000 mètres). Ceci à l'inconvénient de pousser les utilisateurs à extraire les données en utilisant uniquement le géocode et non le champ de précision associé et à obtenir un point dénué du grand rayon auquel il est associé. La méthode Wieczorek vient à bout de cet inconvénient en ne fournissant pas un géocode aussi trompeur. Les neuf catégories listées par Wieczorek *et al.* (2004) sont (avec des exemples modifiés):

1. ***Suspecte** (« Dubious »)* (par exemple, 'Boca Brava?')
2. ***Ne peut être localisée*** (par exemple 'Mexico', 'endroit non localisé')
3. ***Démonstrativement inexacte*** (par exemple, contient des affirmations contradictoires)
4. ***Coordonnées*** (par exemple, latitude et longitude, coordonnées UTM)
5. ***Lieu nommé*** (par exemple, 'Alice Springs')
6. ***Décentrée*** (par exemple, '5 km en dehors de Calgary')
7. ***Décentrée le long d'un chemin*** (par exemple, 24 km N de Toowoomba le long de Darling Downs Hwy')
8. ***Décentrée dans une direction orthogonale*** (par exemple, '6 km N et 4 km O de Welna')
9. ***Décentrée à partir d'une indication de direction/mouvement (« heading »)*** (par exemple, 50 km NE de 'Mombasa')

Chacun de ces points nécessiterait une méthode de calcul différente de la précision comme il en est question dans le document (Wieczorek *et al.* 2004).

Directives de Géoréférencement

Deux excellentes directives ont été développées pour assister les gestionnaires de données dans le géoréférencement. Les Directives de Géoréférencement, *Georeferencing Guidelines*, développées par John Wieczorek au Musée de Zoologie des Vertébrés de Berkeley, *Museum of Vertebrate Zoology*, (Wieczorek 2001) et les directives du MaPSTeDI (*Mountains and Plains Spatio-Temporal Database Informatics Initiative,* University du Colorado 2003) sont deux des plus englobantes. Je crois qu'il existe également des directives développées par Conabio au Mexique (CONABIO 2005), qui sont en train d'être traduites en anglais et qui seront donc bientôt disponible en espagnol et en anglais.

Contrôles d'édition

Les contrôles d'édition incluent des règles d'usage déterminant les valeurs permises pour un champ en particulier. Une des erreurs les plus fréquente dans les bases de données spatiales est l'omission accidentelle du signe '–' (moins) dans les enregistrements des hémisphères sud ou est. Si la base de données provient de l'hémisphère sud (Australie par exemple), il devrait alors être automatique que les données soient dotées d'un signe de latitude "négative". Les bases de données avec des enregistrements mélangés sont évidemment encore plus compliquées à utiliser, mais les champs relatifs aux pays et aux états pourraient être utilisés pour vérifier par comparaison la latitude et la longitude.

Les bases de données ne sont à la base pas toutes crées correctement, et ceci rend possibles des erreurs qui auraient pu être évitées. Par exemple, les latitudes supérieures à 90° ou inférieures à -90° et les longitudes supérieures à 180° ou inférieures à -180°. Si la base de données le permet, alors sa structure doit être modifiée sous peine de devoir effectuer des contrôles régulièrement pour s'assurer que des erreurs de ce type ne se produisent pas ou sont corrigées.

Utiliser des bases de données existantes afin de déterminer des géocodes

Les informations déjà incluses dans la base de données peuvent être utilisées pour attribuer des géoréférencements à de nouvelles données qui sont en cours d'ajout. Une simple procédure de production de rapport peut être incorporée qui permet une recherche établissant si un spécimen d'une autre localité a déjà été entré en base de données et si on lui a déjà attribué un géocode.

Par exemple, vous êtes sur le point d'entrer en base de données une collection qui a une information de localisation de "10 km NO de Campinas" mais pas d'information de géoréférencement. Vous pouvez chercher dans la base de données pour "Campinas" et parcourir les collections déjà entrées en base de données pour voir si un géocode a déjà été attribué à une autre collection à "10 km NO de Campinas". Ce procédé peut être rendu beaucoup plus simple si la structure de la base de données inclue des champs pour "Lieu connu le plus proche" *(« Nearest Named Place »)*, "Distance" et "Direction" ou semblables, en plus du champ de description sous forme textuelle libre traditionnel.

Cette méthodologie pose problème si le premier géocode a été attribué avec une erreur parce qu'alors cette erreur sera répétée tout au long de la base de donnée. Cependant, cela permet une correction globale si une erreur de ce type est trouvée dans une des collections mis dans ce type de base de données. Si une telle méthode est utilisée pour déterminer le géocode, elle devrait être documentée dans le champ "Méthode de géocodage" *(« Geocode method »)* (voir ci-dessus).

Avec des bases de données liées, telles que l'Herbier virtuel australien (CHAH 2002), *species*Link (CRIA 2002), ou le portail GBIF (GBIF 2004), des procédures en ligne pourraient être mises en place afin de permettre un historique de géocodage collaboratif et utilisées de façon similaire. De telles collaborations peuvent être menées par l'usage de Web Services(Beaman *et al.* 2004, Hobern et Saarenmaa 2005). Évidemment, un inconvénient est la perte d'un certain contrôle de votre base de données et les erreurs d'une base de données peuvent être copiées par inadvertance dans votre propre base de données. Lorsque cela est le cas, le champ identifiant de la source *(« source-id »)* devrait être attachée à l'enregistrement de telle sorte que les mises à jour futures et les corrections y soient incorporées. Un bon mécanisme de feedback nécessiterait d'être développé entre les institutions pour assurer que les erreurs ne soient d'une part pas répétées par inadvertance et que d'autre part, l'information sur les erreurs détectées soient rapportées à la base de données originelle et à toutes celles qui en dépendent.

Beaucoup de collections de plantes sont distribuées en tant que "duplicatas" à d'autres institutions. Traditionnellement, cela se faisait avant le géoréférencement, et l'on peut souvent trouver exactement les mêmes collections dans un différemment nombre d'institutions, avec à chaque fois un code de géoréférencement différemment. Pour éviter ces divergences, les géocodes ont besoin d'être ajouté avant distribution, ou un arrangement collaboratif doit être mis en place entre les institutions. Comme expliqué précédemment, il coûte beaucoup de temps et d'argent d'ajouter des géocodes, et c'est un exercice extrêmement inutile si plusieurs institutions individuelles passent du temps à géoréférencer les mêmes collections. Cela est d'autant plus vrai si différents géocodes sont données à la même collection dans ces institutions séparées.

Attribution automatisée de géocodes

Les outils de géoréférencement automatisé sont basés sur la détermination de la latitude et de la longitude à partir d'informations de localité textuelles utilisant une distance et une direction à partir d'un lieu connu. Idéalement, les bases de données incluent au moins un "Lieu connu le plus

proche", "Distance" et "Direction" ou encore mieux "Dieu connu 1", "Dist 1", "Dir. 1", "Lieu connu 2", "Dist 2", "Dir. 2". Donc, "5 km E de Smithtown, 20 km NNO de Jonestown" serait placé de manière appropriée dans les six champs cités ci-dessus.

Comme la plupart des bases de données ne sont pas aussi structurées, des efforts sont fait pour développer des logiciels d'analyse grammaticale automatisés afin de changer les champs de description de localité libres en simples champs de type "Lieu connu le plus proche", "Distance" et "Direction", qui utilisent ensuite ces champs en collaboration avec les dictionnaires appropriés pour déterminer la géoréférence (voir *BioGeomancer* ci-dessous). Pendant que le géocode est déterminé de cette façon, une Exactitude du Géocode *(« Geocode Accuracy »)* devrait être rapportée dans un champ supplémentaire et quand cela serait possible, les résultats devraient être vérifiés par des experts par comparaison avec les originaux afin d'éviter des erreurs imprévues. Dans tous les cas, une telle analyse ne devrait en aucune façon altérer la donnée (ou le champ) de "localité" originale, mais devrait être une information ajouté en annexe. Cela pourrait donc toujours être utilisé pour vérifier la fiabilité de l'exercice d'analyse.

Les inconvénients de cette méthodologie incluent des possibles erreurs de dictionnaires (les dictionnaires les plus disponibles auprès du grand public ont un nombre considérable d'erreurs (voir par exemple l'image 15), les localités relatives au lieu connu le plus proche ne sont pas toujours aussi claires que celles citées précédemment, souvent des noms historiques sont utilisés et beaucoup de distances sont "par route" et non pas "à vol d'oiseau", même si cela est rarement indiqué. Des champs de fiabilité nécessitent de prendre en considération ces problèmes et les erreurs inhérentes aux vecteurs de distances – est ce que "Sud Ouest" signifie "Sud" et "West" ou entre SSW et WSW. Plus la distance avec la source augment, plus les erreurs inhérentes augmenteront rapidement (voir Wieczorek *et al.* 2004). L'utilisation d'une telle méthode en accord avec un simple GIS pourrait fournir l'opportunité à l'opérateur de voir l'information sur une carte et d'ensuite "déplacer " avec la souris le point à un endroit plus approprié- par exemple, vers la route la plus proche.

Logiciel de géocodage

Un certain nombre d'outils en ligne ou indépendants ont été développés pour assister les utilisateurs dans le géoréférencement de leur collection. Trois sont ici mentionnés, deux "en ligne" et deux "indépendants".

BioGeoMancer

BioGeoMancer est un système de géoréférencement automatisé créée pour les collections d'histoire naturelle (Wieczorek et Beaman 2002). Dans l'état présent, BioGeoMancer est un système prototype, et les commentaires ci-dessous ne prennent pas en compte les modifications prévues censées améliorer son utilisation. BioGeoMancer peut analyser grammaticalement des descriptions de noms de lieu en anglais et fournir un jeu de coordonnées de latitudes et de longitudes associées à cette description. L'analyse grammaticale des champs libres, données de localités en anglais, fournit une réponse comportant le lieu connu le plus proche, la distance et la direction sous le format suivant (Wieczorek 2001b):

- 2.4 km WNW of Pandemonium
- Springfield, 22 miles E
- Springfield, 0.5 mi. E of Pandemonium

Comme un certain nombre d'autres programmes (par exemple Diva-GIS, eGaz) il prend l'information analysée et en conjonction avec un gazetteer approprié, calcule un jeu de coordonnées. L'avantage de BioGeoMancer par rapport à d'autres programmes de géocodage est qu'il fait une analyse du texte. C'est le premier programme de géo-analyse de cette sorte accessible au public et aux chercheurs sur l'Internet.

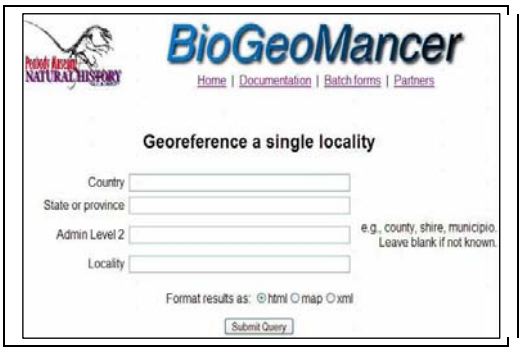


Fig. 2. *formulaire de recherche de localités simples de BioGeoMancer* http://biogeomancer.org/ *(Peabody Museum n.dat.).*

Le programme BioGeoMancer existe sous deux formes. La première est un formulaire de recherche de type Web pour une seule valeur (figure 2) qui permet aux utilisateurs de taper une localité et de trouver des géoréférences en retour. La seconde forme, un procédé groupé *(« batch »)*, accepte des données dans des interfaces HTTP/CGI dans des documents textes bruts délimités par des virgules *('CSV' : Comma Separated Values')* (figure 3) ou dans une version SOAP/XML qui fournit un fichier réponse avec des informations de géoréférencement soit sous une de document délimité par des virgules, dans un format HTML (figure 4) ou un format XML. Ce projet a récemment reçu une avancée considérable dans le financement et s'est étendu pour devenir une collaboration mondiale tentant de développer de nouveaux outils de géoréférencement améliorés.

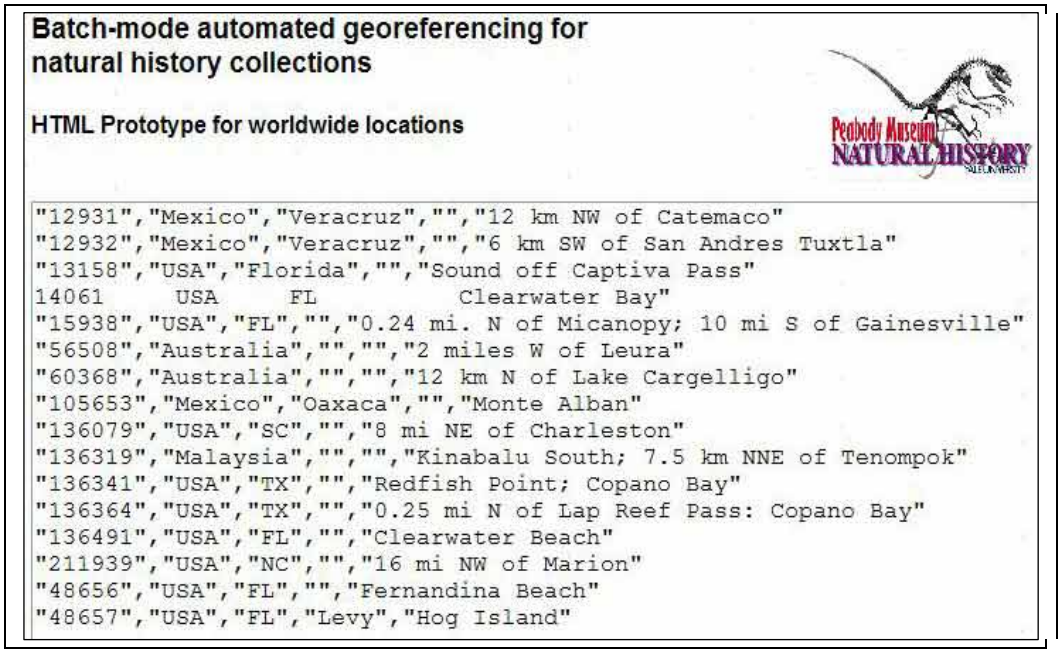


Fig. 3. *Format d'entrée des données en mode groupé de l'outil BioGeoMancer pour le Web, outil de géoréférencement automatisé pour les collections d'histoire naturelle.* http://georef.peabody.yale.edu/yu/bgm-forms/batch-int02.html *(Peabody Museum n.dat.).*

Biogeomancer Results

Summary

Query Id	Query Country	Query Adm1	Query Adm2	Query Locality	Number of records matched	Centroid Latitude	Centroid Longitude	Error radius (km)	Multipoint match	Bounding box
12931	Mexico	Veracruz		12 km NW of Catemaco	1	18.49331	-95.19701	0.0	MULTIPOINT(-95.19701 18.49331)	BOX(-95.19701 18.49331, -95.19701 18.49331)
12932	Mexico	Veracruz		6 km SW of San Andres Tuxtla	1	18.41167	-95.25682	0.0	MULTIPOINT(-95.25682 18.41167)	BOX(-95.25682 18.41167, -95.25682 18.41167)
13158	USA	Florida		Sound off Captiva Pass	1	26.60917	-82.22222	0.0	MULTIPOINT(-82.22222 26.60917)	BOX(-82.22222 26.60917, -82.22222 26.60917)
14061	USA	FL		Clearwater Bay	1	27.97222	-82.82083	0.0	MULTIPOINT(-82.82083 27.97222)	BOX(-82.82083 27.97222, -82.82083 27.97222)
15938	USA	FL		0.24 mi. N of Micanopy; 10 mi S of Gainesville	1	29.50614	-82.325	0.0	MULTIPOINT(-82.32500 29.50614)	BOX(-82.32500 29.50614, -82.32500 29.50614)
56508	Australia			2 miles W of Leura	2	-28.449995	149.925235	587.4	MULTIPOINT(149.55188 -23.18333, 150.29859 -33.71666)	BOX(149.55188 -33.71666, 150.29859 -23.18333)

Fig. 4. *Échantillon partiel du résultat sous forme tabulaire de l''outil Web BioGeoMancer, outil de géoréférencement automatisé pour les collections d'histoire naturelle. (Peabody Museum n.dat.).*

Lorsque plus d'une option est possible, elles sont toutes rapportées sous cette identité. Lorsqu'aucune option n'est évidente, l'information n'est pas retournée. La colonne *Bounding Box* fournit l'exactitude calculée. Le système fonctionne bien pour une grande partie des données, mais rencontre des difficultés avec le texte lorsqu'il n'est pas facilement décomposable dans les champs de lieu connu le plus proche, de distance et de direction. D'autres problèmes remarqués dans la version actuelle à l'heure de la rédaction de ce manuel incluent (des améliorations futures sont prévues afin de les réduire):

- Il est limité aux descriptions en langue anglaise.
- La précision n'est reportée que dans une *Bounding Box* dans la version présente, et ceci pourrait être amélioré. Déjà, un programme en rapport développé par John Wiecrorek (2001b)- La calculette de géoréférencement, the *Georeferencing Calculator*, peut délivrer ces informations http://manisnet.org/manis/gc.html et pourra être lié à BioGeoMancer dans l'avenir. Des travaux sur une méthode d'assignation d'exactitude automatique ont déjà commencé au travers de ce qui a été nommée la "méthode point-radius" pour le

géoréférencement et le calcul de incertitude associée aux données (Wieczorek et al. 2004)
- Le traitement de deux localités nommées, (exemple. “10 km W of Toowoomba towards Dalby”) produit un résultat nul.

Un autre programme d’analyse, RapidMap Geocoder (NMNH 1993) avait été développé en 1993 par le Muséum National d’Histoire Naturelle américain et le musée Bernice P. Bishop à Hawaï, uniquement pour des localités hawaïennes. Cependant, il n’avait pas été considéré comme une réussite et son développement avait été arrêté. Certaines informations utiles sur la méthodologie d’analyse utilisée sont malgré tout disponibles sur l’Internet à : http://users.ca.astound.net/specht/rm/tr_place.htm.

GeoLoc-CRIA

GeoLoc est un programme simple pour le Web pour trouver des localités au Brésil, une distance et une direction connues à partir d’une localité répertoriée dans un gazetier. Il a été développé à CRIA (Marino *et al.* in prep.). GeoLoc fonctionne d’une façon semblable au programme eGaz (voir ci-dessous) et peut être trouvé à http://splink.cria.org.br/tools/ (CRIA 2004a). Le prototype inclut un certain nombre de dictionnaires spécialisés et fournit à l’utilisateur la possibilité de sélectionner celui à utiliser lorsque plus d’un est disponible pour une même zone. Il fournit également une marge d’erreur calculée.

Un exemple peut être vu dans la figure 5, où la latitude et la longitude d’une localité à 25 km NE de Campinas dans l’état de São Paulo, Brésil, est recherchée. D’abord, on trouve la localité pour Campinas en utilisant un des nombreux dictionnaires spécialisés ou les enregistrements de *species*Link (enregistrements obtenus à travers l’organisation d’une recherche décentralisée d’un ensemble de bases données concernant principalement l’état de São Paulo). Ensuite, en ajoutant “25 km” et “NE” (encerclé) et en cliquant sur le bon “Campinas” (flèche), les résultats apparaitront sur une carte (figure 6). Le géocode donné est -46.9244, -22.7455 avec une marge d’erreur de 9.754 km (encerclé). Ces informations (latitude, longitude, erreur) sont déjà stockées dans l’outil de copier/coller Microsoft et peuvent être collé dans n’importe quel fichier Microsoft Windows compatible tel que Word, Excel, et Access. La carte montre également l’emplacement de “Campinas” d’après trois sources -celle en rouge étant celle choisie en accord avec le point “25 km NE de Campinas”. On peut zoomer et travailler sur la carte, et de nombreuses couches environnementales peuvent être activées ou désactivées.

Le programme peut aussi mener à une feuille de calcul EXCEL de localités et produire un tableau HTML des résultats pour les recherches à venir, ou une feuille de calcul EXCEL. Le problème majeur de ce programme est qu’il n’est disponible que pour les localités brésiliennes. Les algorithmes sont en train d’être incorporés dans le plus vaste projet Biogeomancer.

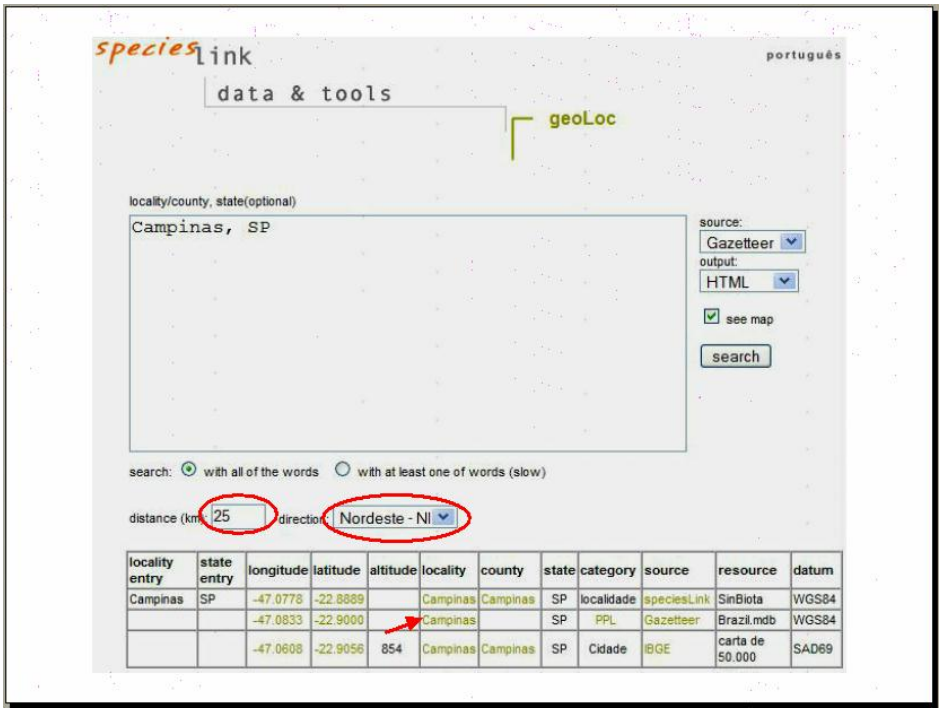


Fig. 5. *Utilisation du module « geoLoc » du programme CRIA pour trouver le géocode d'une localité à 25 km NE de Campinas, SP.*

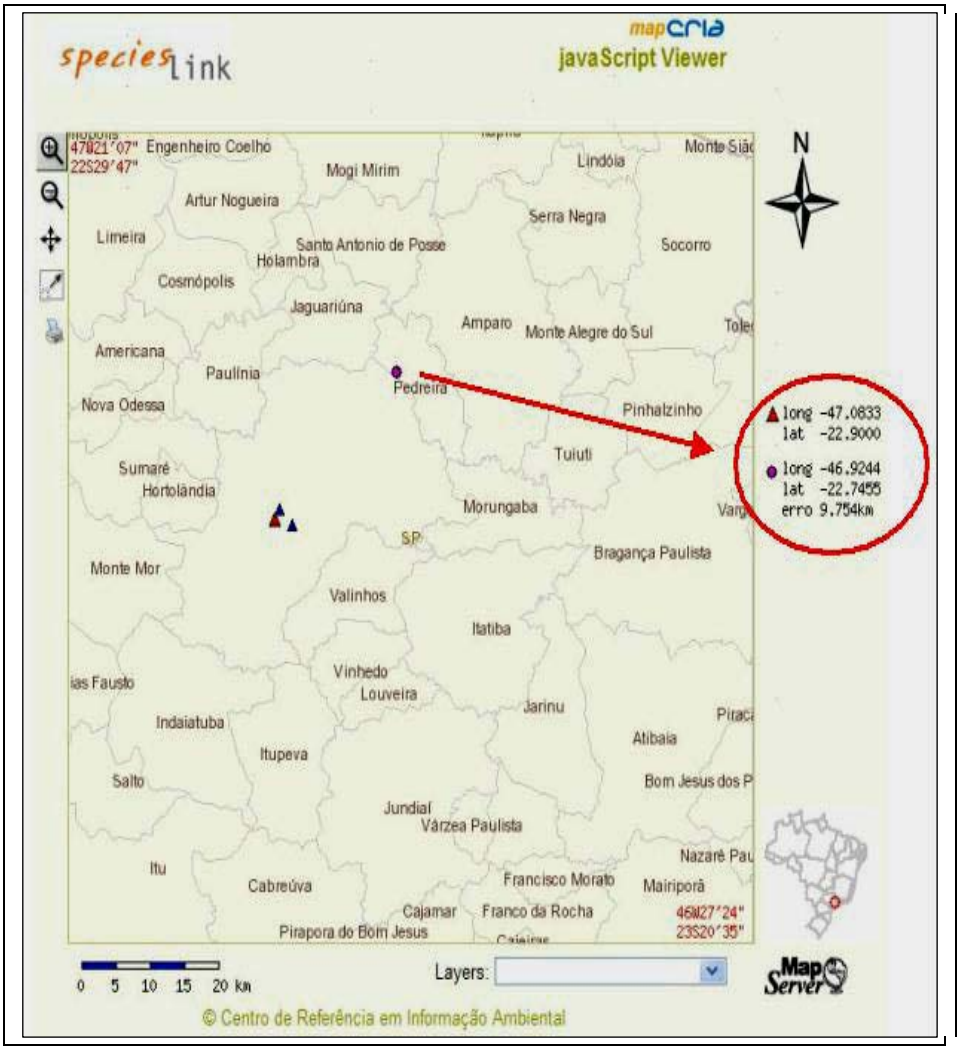


Fig. 6. *Résultats de la sélection ci-dessus montrant la localité de "Campinas" (d'après les différentes sources) et le point à 25 km NE de Campinas, avec des informations de géocode et d'erreurs associées (dans le cercle).*

GEOLocate

GEOLocate (Rios et Bart *n.dat*.) est un programme de géoréférencement développé par le Muséum d'Histoire Naturelle de l'Université de Tulane. Il a été conçu pour faciliter la tâche d'attribution de coordonnées géographiques à une donnée de localité associée à une collection d'histoire naturelle. Les buts principaux de GEOLocate sont de :

- développer un algorithme pour convertir des données d'histoire naturelle textuelles en longitude et latitude pour l'Amérique du Nord ;
- Fournir une interface pour la visualisation et les ajustements ultérieurs des coordonnées générées ;
- Fournir une solution simple pour que les utilisateurs importent et géoréférencent leurs données ;
- Fournir une option de mise à jour automatique.

L'algorithme standardise d'abord le rayon de la localité en termes communs et décompose les distances, directions et identifiants géographiques clés tels que les lieux ayant un nom. Cette information est utilisée ensuite en étant complétée par des gazetteer (incluant le nom des lieux, la longueur des rivières, l'exploitation des terres et les endroits traversés par des routes ou des rivières) pour déterminer les coordonnées géographiques. Le programme permet également à l'utilisateur de déplacer des localités vers l'étendue d'eau la plus proche.

•Le programme est disponible à l'Université de Tulane, et une démonstration en ligne est disponible à http://www.museum.tulane.edu/geolocate/demo.aspx (figure 7).

Fig. 7. *Un exemple de l'interface de GEOLocate qui utilise une version de démo en ligne pour identifier les coordonnés géographiques de Cambridge, Ontario.*

Le programme ne fonction seulement que pour l'Amérique du Nord (Mexique, USA et Canada) mais les concepteurs sont en train de travailler à son extension afin d'y inclure le monde entier. D'autres développements incluront la compatibilité avec DiGIR, un support multilingue, et des techniques de validation avancées (N.Rios pers. com. 2004).

eGaz

eGaz (Shattuck 1997) est un programme développé à la Collection Nationale d'Insectes de la CSIRO, *Australian National Insect Collection*, pour assister les herbiers et les musées dans l'identification et l'ajout de géocodes à leurs enregistrements de spécimens. Après le développement, il avait été incorporé dans le logiciel de saisie de données et de gestion des spécimens BioLink (Shattuck and Fitzsimmons 2000) après le développement de celui-ci. eGaz est disponible en tant que partie du progiciel Biolink (voir ci-dessous).

eGaz élimine le besoin de cartes papiers et de règles pour déterminer la latitude et la longitude des villes, villages, lacs ou autres lieux nommés. eGaz peut aussi calculer la latitude et la longitude pour des sites placés à distance et direction connues à partir d'un lieu nommé. Le programme permet l'inclusion facile de dictionnaires spécialisés de n'importe quelle région, et des gazetteer pour la plupart du monde sont disponibles en téléchargement depuis le site du CSIRO (http://www.biolink.csiro.au/gazfiles.html).

eGaz est un produit Microsoft Windows fournissant deux fenêtres: une fenêtre de gazetteer et une fenêtre de cartes (figure 8). Il permet à l'utilisateur de chercher une localité sous la forme "lieu nommé", "distance" et "direction" afin d'obtenir un géocode et de le transférer à un fichier.

L'exemple présenté dans la figure 8 traite de l'obtention d'une latitude et d'une longitude d'un endroit décrit comme "80 km SSW de Toowoomba", Queensland, Australie. La première étape est de charger le gazetier adéquat et d'y sélectionner "Toowoomba" (**A**). Il y a un grand nombre d'options, mais j'ai sélectionné la ville de Toowoomba (étiquetée POPL pour zone peuplée - *« Populated Place »*). La localité de Toowoomba apparait en rouge sur la carte (**B**). La distance "80" est entrée dans le champ de Distance et le menu déroulant est utilisé pour sélectionner "km" et "SSW" (**C**). La localité sélectionnée apparaît sous forme de point bleu sur la carte (**D**). La localité, avec la latitude et la longitude, apparaissent également en bas de la fenêtre du gazetier (**E**). Simplement en cliquant sur cette zone et en sélectionnant "copier" cette information peut être copiée et collée dans n'importe quel fichier Microsoft Windows compatible (Word, Excel, Access). La latitude et la longitude (à une résolution d'une minute d'arc) apparaissent également (**F**), et ceci peut être copié de façon similaire. Alternativement, en allant dans le menu d'édition et en sélectionnant "Copier Lat/Long", le géocode peut être copié avec une précision d'une seconde d'arc (1/3600 de degré).

Il est également possible d'aller sur la carte et de zoomer sur un point. D'autres couches telles que des réseaux routiers (au format de fichier ESRI Shapefile) peuvent être chargées pour permettre un positionnement plus exact des points. Par exemple, en le déplaçant peut être vers la route la plus proche si l'on collecte des données depuis une voiture. L'outil de sélection peut être utilisé en cliquant sur le point pour obtenir le géocode avec une résolution d'une seconde d'arc. À nouveau, simplement cliquer avec la souris ou utiliser Edition/Copier/Coller, permet à l'information d'être copiée dans le fichier approprié.

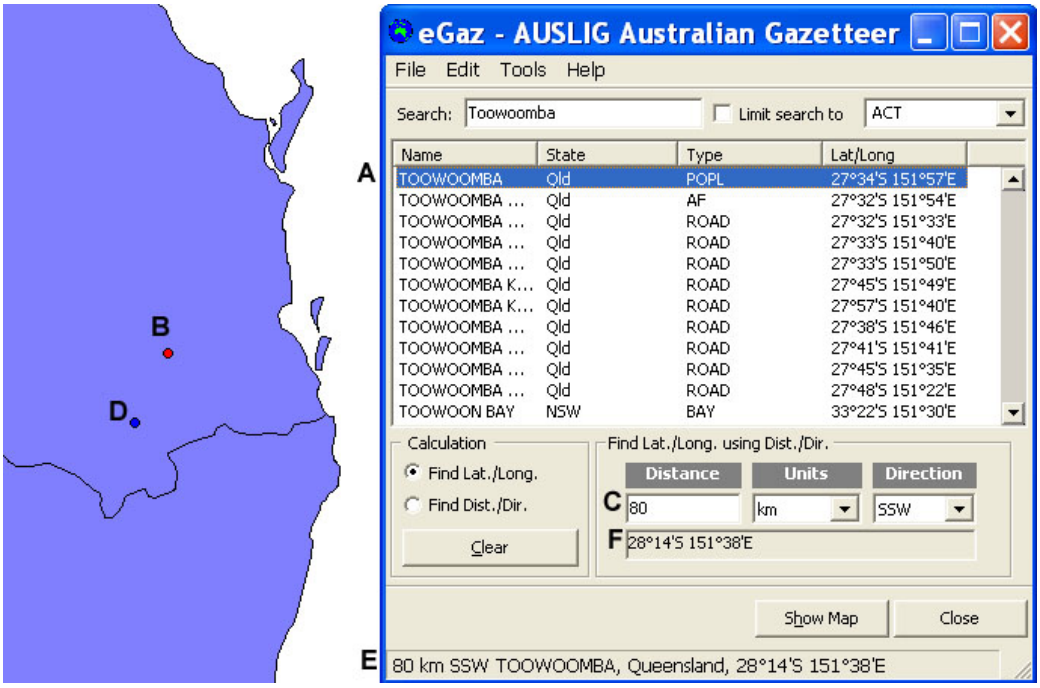


Fig. 8. *Échantillon de résultat à partir d'eGaz, montrant la détermination de la latitude et de la longitude pour une position à 80 km SSW de Toowoomba, Queensland, Australie. A. Information sur Toowoomba à partir du lexique. B. Localisation sur carte de Toowoomba. C. Champs d'encodage montrant les localités à 80 km SSW. D. Localisation sur carte de 80 km SSW de Toowoomba. E. Détails sur la localité. F. Latitude et longitude de la nouvelle localité.*

Diva-GIS

Diva-GIS est un programme GIS gratuit développé pour les musées et les herbiers[4]. Il inclut notamment un algorithme qui assiste dans l'assignation de coordonnées aux spécimens lorsqu'elles sont manquantes. Un prétraitement est nécessaire pour organiser les données dans un format accepté par le programme, mais un certain nombre de bases de données commencent déjà à structurer leurs données de cette manière. Le fichier de saisie demande que les données de localisation textuelles soient décomposées dans un certain nombre de champs spécialisés tels que "Lieu nommé 1 *(« Named Place 1 »)*", "Distance 1", "Direction 1" et "Lieu nommé 2 *(« Named Place2 »)*", "Distance 2", "Direction 2". Par exemple, le rapport de localité :

"*poussant dans une localité appelée Ulta, 25.2 km E de Chilla*"

serait analysé ainsi :

Lieu Nommé 1:	Ulta
Distance 1:	

[4] Note du traducteur (FT) : de part sa polyvalence Diva-Gis est également un logiciel fort utilisé par des biologistes actifs dans d'autres domaines que la botanique et la conservation d'herbiers

Direction 1:	
Lieu Nommé 2:	Chilla
Distance 2:	25.2km
Direction 2:	E

et

"14 km ESE de Sucre sur la route de Zudanez" serait analysée ainsi:

Lieu Nommé 1:	Sucre
Distance 1:	14 km
Direction 1:	ESE
Lieu Nommé 2:	Zudanez
Distance 2:	
Direction 2:	

Cependant, un seul groupe de champs "Lieu Nommé", "Distance" et "Direction" sera en mesure de fournir le géocodage pour de nombreux enregistrements, et cela constitue l'intégralité des informations que la plupart des institutions obtiendront. Les auteurs de Diva-GIS (Hijmans *et al.* 2005) recommandent d'arrondir les distances au nombre entier inférieur pour contrebalancer les inexactitudes de données et pour prendre en compte les cas où 25 km au Nord d'un lieu signifie 25 km par route et non pas à vol d'oiseau. Je recommanderai le contraire, et rapporterai les chiffres les plus précis possibles, et placerai une précision du chiffre dans un champ de "précision" en mètres.

Quand un fichier de saisie a été sélectionné, le nom et le chemin du fichier de résultat désigné, et les noms de champs appropriés sélectionnés à partir d'un menu déroulant, l'algorithme est lancé et produit des résultats (figure 9). L'algorithme utilise un Gazetteer approprié pour assigner les coordonnées.

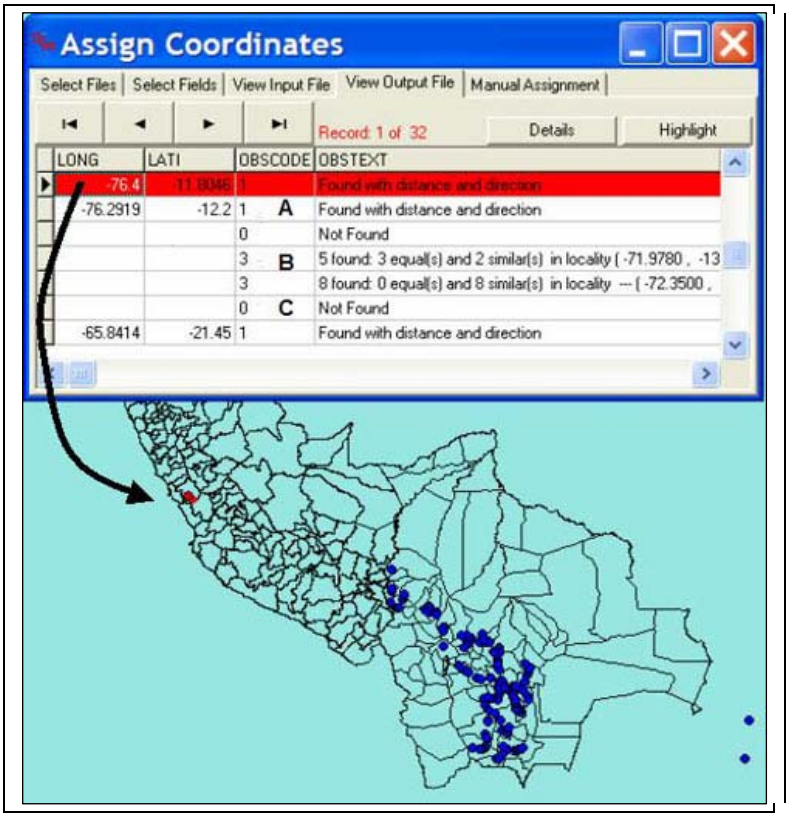


Fig. 9. *Résultats avec Diva-GIS* en utilisant des géocodes attribués automatiquement. ***A.*** *Géocodes non-ambigus trouvés et attribués par le programme.* ***B.*** *Géocodes ambigus identifiés.* ***C.*** *Géocode approprié non trouvé.*

Comme montré dans cet exemple (figure 9), le programme a trouvé des résultats non ambigus dans le ou les gazetier(s) pour un certain nombre d'enregistrements utilisant le champs "lieu connu" dans le fichier de saisie et leur a attribué un géocode calculé de façon inappropriée (**A**). Une fois que le fichier de résultat a été chargé et qu'un fichier Shapefile a été créé, chacun de ces rapports peuvent être mis en valeur pour produire un point aux couleurs facilement repérables sur la carte. Dans un certain nombre de cas, le programme a trouvé plusieurs résultats possibles dans le ou les gazetier(s) pour le "Lieu nommé" et a mentionné cette ambiguïté de manière appropriée (**B**). Mais dans d'autres cas (**C**), le programme a été incapable de trouver des résultats dans le gazetier.

Dans le cas d'enregistrements où un certain nombre de valeurs concordantes possibles est trouvé, il est possible de passer à l'étape suivant en double-cliquant sur l'une des lignes (**B**) ce qui produit un nouveau fichier de résultats (figure 10).

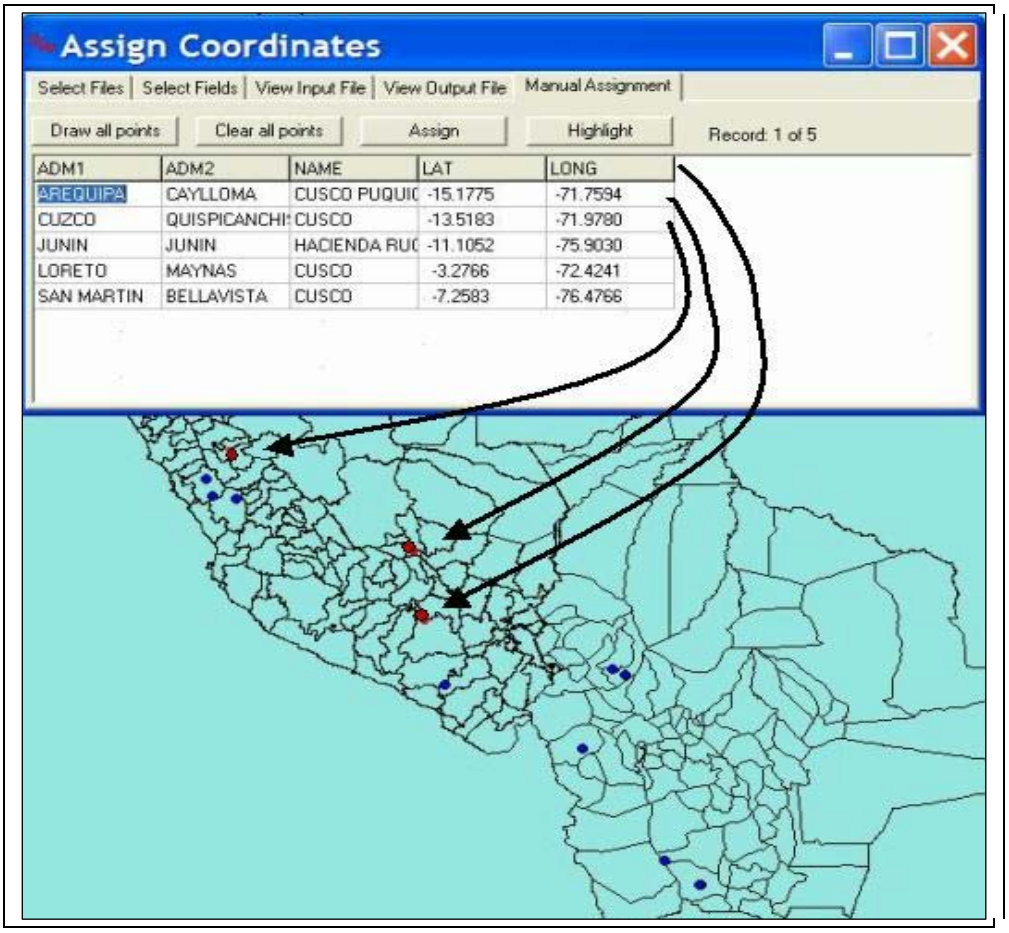


Fig. 10. *Résultats de Diva-GIS montrant les géocodes possibles pour des données où l'utilisation d'un gazetier a produit un certain nombre d'alternatives crédibles.*

Dans le cas du rapport montré dans la figure 11, le programme a identifié cinq localisations alternatives possibles à partir du ou des gazetier(s) et les a présenté sur le GIS pour laisser le choix aux utilisateurs. Lorsque l'un d'eux est choisi, il suffit juste de cliquer sur le bouton "attribuer" *(« Assign »)* pour que cela soit ajouté au fichier de résultats. Alternativement, on peut décider d'une autre localisation en même temps et utiliser " l'attribution manuelle" *(« Manual Assignement »)* pour ajouter un géocode ou modifier un de ceux déjà assignés.

Vérification des géocodes et validation

Il y a quatre façons principales de procéder pour vérifier et valider les géocodes sur les spécimens entrés en base de données. Il s'agit d'une utilisation de bases de données ayant pour but de vérifier les incohérences internes, l'utilisation de systèmes d'information géographiques, l'utilisation d'espaces environnementaux pour vérifier les valeurs extrêmes et l'utilisation de données statistiques pour vérifier les incohérences dans les espaces géographiques ou environnementaux.

Utiliser les bases de données

Contrôles internes

La plupart des bases de données d'espèces incluent un certain nombre d'informations redondantes.

Par exemple, l'état dans lequel la collecte avait été effectuée et un champ d'information textuelle de localisation. Certaines bases de données incluent un champ de "lieu connu le plus proche" et ceci peut également dupliquer les informations au sein du champ de localité. Des contrôles peuvent être menés pour vérifier que la ville citée ou le lieu connu le plus proche sont localisés dans le bon état ou district, ou même pays s'il est mentionné dans un autre champ.

Vérifier les informations d'une base de données à partir d'enregistrements similaires est également possible. Par exemple, vérifier toutes les localités par rapport aux latitudes et longitudes mentionnées. Il peut y avoir une base de données contenant 5 activités de récolte d'une même localisation -"10 km N de Campinas, SP" par exemple. Ont-elles doutes la même latitude et longitude ou est-ce que l'une ou plusieurs diffèrent-elles du reste de façon significative? Voir également la discussion sur les Règles d'Association Ordinales , ci-dessous.

Nettoyage de données (speciesLink)

Le module de nettoyage de donnée CRIA du Système de Distribution d'Information de speciesLink (CRIA 2002) inclut un certain nombre de protocoles pour identifier les erreurs possibles afin d'aider les gestionnaires de collections à traiter leur données. À l'heure actuelle, ceci n'est disponible qu'en portugais mais une version anglaise est en projet. Une partie de cet outil a pour but d'identifier les erreurs de noms. Les protocoles incluent :

- Des listes de noms (famille, *family*, genre, *genus*, espèces, *species*, sous-espèces, *subspecies*) accompagnées du nombre d'occurrences dans la base de données en question. Un rapide coup d'œil sur un exemple (figure 11) montre un certain nombre de problèmes évidents. La première ligne montre qu'il y a 101 occurrences dans la base de données dont des enregistrements non identifiés au delà de la famille. La deuxième ligne montre une occurrence avec un nom de famille "4606euphorbiaceae" et la troisième ligne montre 5 enregistrements uniquement identifiés au niveau de la de la famille pour Acanthaceae.

family	genus	species	subspecies	ocor_col
[]	[]	[]	[]	101
[4606euphorbiaceae]	sp [Julocroton]	[humilis]	[var. subpannosus]	1
[Acanthaceae]	[]	[]	[]	5

Fig. 11. *Extrait du module de Nettoyage de données de CRIA montrant quelques erreurs possibles.*

- L'examen des erreurs possibles dans les noms de genre. C'est le cas lorsque le nom de famille est le même, que le nom de l'espèce est le même, que le nom de genre est similaire (identifiés grâce à des algorithmes de type soundex) mais que l'orthographe du nom générique est différent. Ces données produites en sortie montrent aussi le nombre d'occurrences pour chacun des noms dans la base de données étudiée, et le nombre total d'occurrences dans toutes les bases de données auxquelles on a accès grâce à speciesLink. L'exemple (figure 12) montre deux orthographes différentes pour le genre *Hieronyma* (avec aussi deux orthographes différentes pour *alchornioides*, mais celles ci sont identifiées sous un protocole différent) ainsi que le nombre d'occurrences pour chacune. On peut cliquer sur " sp " et cela vous permet de chercher une série de base de données à la fois internes et externes à l'organisation. L'outil inclut des ressources

telles que l'Index International du Nom des Plantes, *International Plant Name Index* (IPNI), species 2000, etc… qui peuvent aider l'utilisateur à déterminer ce qui pourrait être la bonne orthographe.

family	genus	species	subspecies	ocor_col	ocor_total
[Euphorbiaceae]	sp [Hyeronima]	[alchorrneoides]	[]	3	3
[Euphorbiaceae]	sp [Hieronyma]	[alchorrneoides]	[]	1	1
[Euphorbiaceae]	sp [Hyeronima]	[alchornioides]	[]	9	9
[Euphorbiaceae]	sp [Hieronyma]	[alchornioides]	[]	17	17

Fig. 12. *Extrait du module de Nettoyage de données de CRIA montrant quelques erreurs possibles.*

- L'examen des erreurs possibles en matière de nom ou d'épithètes d'espèces. De même que pour le nom de genre, cet outil recherche des noms où le genre est le même, où le soundex de l'épithète de l'espèce est le même mais où il y a une différence dans l'orthographe de l'épithète de l'espèce. À nouveau, le résultat montre le nombre total d'occurrences dans la base de données étudiée, dans les bases de données accessibles et le statut des noms dans species2000[5]. L'exemple (figure 13) montre un nombre de noms d'espèce avec des alternatives. Le nombre d'occurrences de chaque nom, en accord avec le statut dans species2000 lorsqu'il est disponible, peut donner une indication permettant de savoir quelle orthographe peut être erronée.

genus	species	subspecies	ocor_col	ocor_total	status_sp2000
sp [Acacia]	[polyphylla]	[]	83	217	accepted name
sp [Acacia]	[polyphyllla]	[]	1	1	
sp [Banisteriopsis]	[argyrophylla]	[]	25	164	
sp [Banisteriopsis]	[argirophylla]	[]	2	2	
sp [Bauhinia]	[cuyabensis]	[]	19	19	provisionally accepted name
sp [Bauhinia]	[cuiabensis]	[]	1	2	
sp [Bignonia]	[unguis-cati]	[]	1	5	unambiguous synonym
sp [Bignonia]	[unguiscati]	[]	2	2	

Fig. 13. *Extrait du module de Nettoyage de données de CRIA montrant quelques erreurs possibles.*

- L'examen des différences et les erreurs possibles dans le nom des auteurs. La figure 14 montre le nombre de possibilités pour un seul nom d'espèce. À nouveau, cliquer sur "sp" permet de fouiller d'autres bases de données pour aider à déterminer quelle peut être l'alternative à utiliser.

genus	species	subspecies	author	ocor_col	ocor_total
sp [Actinostemon]	[concolor]	[]	[Müll.Arg.]	0	1
sp [Actinostemon]	[concolor]	[]	[(Spreng.) M.Arg.]	0	13
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Muell.Arg.]	0	17
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Mull. Arg.]	1	2
sp [Actinostemon]	[concolor]	[]	[(spreng.) Müll. Arg.]	0	2
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Müll. Arg.]	0	52
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Müll.Arg.]	88	139
sp [Actinostemon]	[concolor]	[]	[(Spreng) Müll. Arg.]	0	6
sp [Actinostemon]	[concolor]	[]	[(spr.) Muell. Arg.]	0	1
sp [Actinostemon]	[concolor]	[]	[(Spr.) Muell.Arg.]	0	2

[5] http://www.species2000.org

Fig. 14. *Extrait du module de Nettoyage de données de CRIA montrant quelques erreurs possibles.*

- L'examen des différences dans les noms de famille et de sous-espèces fonctionne de la même manière.

D'autres protocoles sont utilisés pour identifier des erreurs géographiques possibles dans les jeux de données, et ceci est traité dans *Données Spatiales* ci-dessous. CRIA ne détient pas les données, et ne leur fait aucun changements mais fournit un service aux détenteurs de données, pour les aider à identifier les erreurs possibles dans leur bases de données. Il incombe ensuite à ces détenteurs de décider quels sont les formats corrects et quelles enregistrements devraient être corrigées et lesquels ne devraient pas l'être.

Bases de données externes

Des erreurs dans différents aspects des données occurrences des espèces peuvent être identifiées par la mise en relation avec des bases de données externes. De telles bases de données peuvent inclure des modèles d'élévation numérique, *Digital Elevation Models*, des bases de données de topographie spatiale, des lexiques et des itinéraires de collecteurs.

La plupart des bases de données sophistiquées peuvent être utilisées pour vérifier l'exactitude des champs d'altitudes en comparant l'altitude citée avec celle d'une base de données de modèle d'élévation numérique, *Digital Elevation Models* (DEM). Il est important que le DEM utilisé soit à une échelle appropriée, et à cause de l'exactitude variable de la plupart des données de spécimens, des erreurs (dans le contenu des données ou leur interprétation) peuvent se produire s'il n'est pas utilisé de façon critique. Une telle technique a été utilisée avec succès avec ERIN (*Environmental Resources Information Network,* Réseau d'Information en Ressources Environnementales) en Australie pendant plus de dix ans (Chapman, non publié). Le processus utilise un traitement par lots par le biais d'une base de données ORACLE® et peut vérifier (ou attribuer) les rapports d'altitudes à un taux de plus de 3000 par minute.

Plus récemment, des bases de données spatiales sophistiquées ont été développées, telles que le Moteur de base de données spatiale de l'ESRI, *Spatial Database Engine*, (ArcSDE®) (ESRI 2003) et PostGIS[6] qui permettent des recherches pour des bases de données plus complexes en utilisant les géocodes eux-mêmes. Ce type de logiciel est cependant très coûteux et seulement très peu de musées et d'herbiers sont susceptibles d'avoir les moyens ou le besoin de les utiliser et c'est pourquoi ces méthodes ne sont pas étudiées plus en détail dans ce dossier.

Des gazetiers existent pratiquement dans le monde entier sous une forme ou une autre, et sont souvent disponibles sous la forme de base de données téléchargeable [7]. Ils peuvent être utilisés pour vérifier les champs appropriés au sein de la base de données du spécimen dans le but de vérifier l'exactitude des informations. Un soin particulier doit être employé lors de l'utilisation de beaucoup de ces bases de données puisqu'elles contiennent souvent elles-mêmes des erreurs (voir figure 15 par exemple), et il est important que le bon gazetteer d'une zone particulière soit utilisé et à la bonne échelle. De plus, beaucoup de noms de lieux peuvent être ambigus (ex: il existe des

[6] Note du traducteur (FT) : Postgis est un module gratuit et open-source (libre de droit) pour le traitement des données
géographiques à utiliser en combinaison avec le système de base de données relationnelle lui-aussi gratuit PostgreSQL.

[7] Note du traducteur (FT) : Le site de DIVA-GIS permet d'accéder gratuitement aux gazetteer du NIMA (*U.S. National Imagery and Mapping Agency*) : http://www.diva-gis.org/Data

centaines de “Sandy Creeks” en Austalie) (Chapman et Busby 1994), ou incluent des noms de lieux historiques qui n’apparaissent pas dans les gazetteer. Il y a aussi le problème lié au sens du nom des lieux (Wieczorek 2001a). Un des aspects du nouveau projet BioGeomancer (voir les commentaires dans d’autres sections) est l’intégration de gazetteer avec bases de données biologiques utilisant des technologies de services Web Services . On espère également améliorer ces gazetteer grâce à la participation du public et particulièrement de commencer à inclure des localisations liées à des collections historiques.

Une méthode rarement employée mais qui a un grand potentiel, est la vérification croisée avec les bases de données des localités des collecteurs. Jusqu’à ce jour, très peu de bases de données de ce genre existent (mais la base de données de Harvard cité précédemment serait un bon point de départ général en matière de botanique [8]), et d’autres sont peu à peu développées. Peterson *et al.* (2003) ont récemment suggéré une méthode statistique originale utilisant les oiseaux du Mexique en exemple. Ils ont trié l’accès aux collections d’un collecteur particulier sur des critères de temps et ont déterminé un radius maximum de mouvements possibles pour chaque jour (ou groupe de jours). Utilisant une approche basée sur une formule dans EXCEL, ils ont identifiés les erreurs possibles en matière de spécimens tombant en dehors du champ calculé. Des méthodes similaires pourraient être utilisées dans les bases de données elles-même-voir discussion dans la partie « Règles d’Associations Ordinales » (« *Ordinal Association Rules »)*. De telles méthodes ne fonctionneront cependant que si les collections entrées en base de données sont suffisamment grandes pour créer un tel itinéraire.

Contrôle GIS

Les Systèmes d’Information Géographique, *Geographic Information Systems* (GIS, en français SIG) sont des outils très utiles qui sont devenus bien plus faciles d’utilisation ces derniers temps. Les formes de GIS s’échelonnent des systèmes de fonctionnalité perfectionné et chers des produits gratuits ou disponibles en magasin avec une fonctionnalité plus limitée. Beaucoup de ces SIG gratuits sont cependant suffisamment puissants pour fournir la plupart des fonctionnalités requises pour un herbier ou un musée et peuvent facilement être adaptés afin de fournir des protocoles de vérification et de nettoyage de données.

	Points	Lignes	Polygones
Points	▪Est un voisin de ▪est attribué (« allocated »)	▪Est près de ▪repose sur	▪Est un centroïde de ▪est à l’intérieur de
Lignes		▪croise ▪joint	▪entrecroise ▪est une frontière de
Polygones			▪Se superpose avec ▪est adjacent à

Tableau 3 *: Relation entre les classes d’objets (Gatrell 1991)*

Le SIG peut aussi être utilisé pour vérifier la cohérence logique au sein de la base de données. La redondance dans l’encodage topologique peut être utilisé pour détecter les défauts dans la structure des données telles que des données manquantes et des polygones non étiquetés (Chrisman 1991). Le SIG permet les interrelations de couches spatiales pour détecter les erreurs et ceci, associé à la

[8] http://www.huh.harvard.edu/databases/cms/download.html

visualisation, est un atout majeur.

L'utilisation d'un simple outil SIG pour placer un point (enregistrement de spécimens) contre des polygones (régions, états, pays, territoires) peut aider à détecter des données qui ne coïncident pas entre-elles (géographiques ou altitudinales). C'est un test commun utilisé dans les systèmes SIG connu sous le nom de méthode "point-polygone" (*« point-in-polygon »)*. Elle est utilisée pour les SIG pour s'assurer que les bouées marines n'apparaissent pas sur les terres, que les rivières n'apparaissent pas en dehors de leur lit, etc. Un des tests les plus important qu'un SIG peut mener sur les données primaires d'espèce est de vérifier que les enregistrements supposés se rapporter aux terres se trouvent réellement sur les terres, que ceux supposés être dans l'océan s'y trouve également. Ce simple contrôle effectué après le chargement d'un grand jeu de données dans un SIG rend le mauvais positionnement d'un grand nombre d'enregistrements évident à constater. Des contrôles pour les enregistrements mal placés au moyen d'un SIG peuvent varier de la simple inspection visuelle à des contrôles plus automatisés. L'inspection visuelle peut être très utile pour déterminer si les données tombent dans le bon pays, par exemple. Si vous avez une base de données d'enregistrements brésiliens, en utilisant un SIG vous pouvez rapidement identifier les données mal placées parce qu'elles sont placées en dehors du Brésil. Par exemple, la figure 15 montre des enregistrements d'un gazetteer de noms de lieux brésiliens à accès public comportant des erreurs évidentes. Des erreurs dans les enregistrements de spécimens peuvent similairement être identifiées en utilisant cette méthodologie.

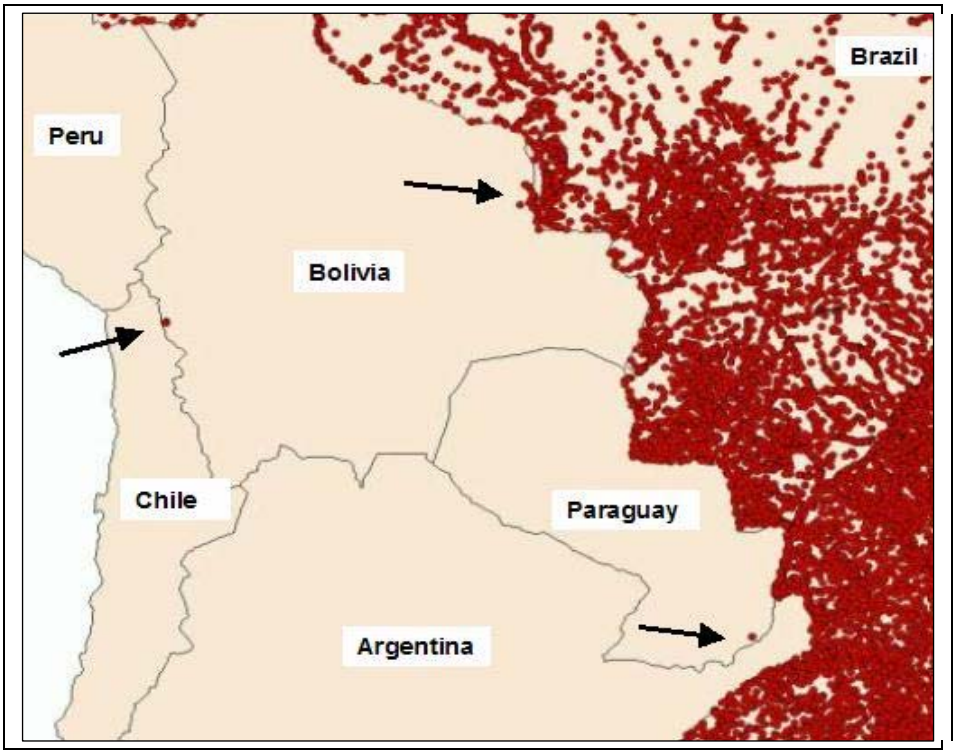


Fig. 15*. Enregistrements d'un gazetier de nom de lieux brésiliens montrant un ensemble erreurs, dont l'une est évidente puisqu'elle s'étend à la frontière chilo-bolivienne, et une autre dans le sud du Paraguay.*

Un certain nombre d'outils, par exemple Diva-GIS (Hijmans *et al*. 2005) et l'outil de nettoyage de données CRIA (CRIA 2005) ont des protocoles qui assistent dans l'identification des erreurs.

Le GIS peut aussi être utilisé pour vérifier que les enregistrements tombent à l'extérieur d'un type de végétation particulier, un type de sol ou de géologie, etc. Certaines espèces sont extrêmement spécifiques à un certain type géologique- calcaire, grès etc. Si vous avez les frontières de ces couches, tout enregistrement tombant à l'extérieur devrait être regardé comme une possible erreur et coché pour une vérification plus poussée (Chapman et al. 2001). La figure 16 cartographie une espèce ne se manifestant que dans des milieux hautement minéralisé de sols de Serpentenite, et deux résultats (marqués 'a' et 'b') apparaissent comme de probables erreurs. Après vérification, le résultat 'a' n'a que la localisation 'Goomeri' - la ville la plus proche de l'affleurement de Serpentenite, et a été géocodée avec la latitude et la longitude de la ville. Le résultat 'b' est assez près de l'affleurement et est sûrement mal placé de par la précision du géocode donné.

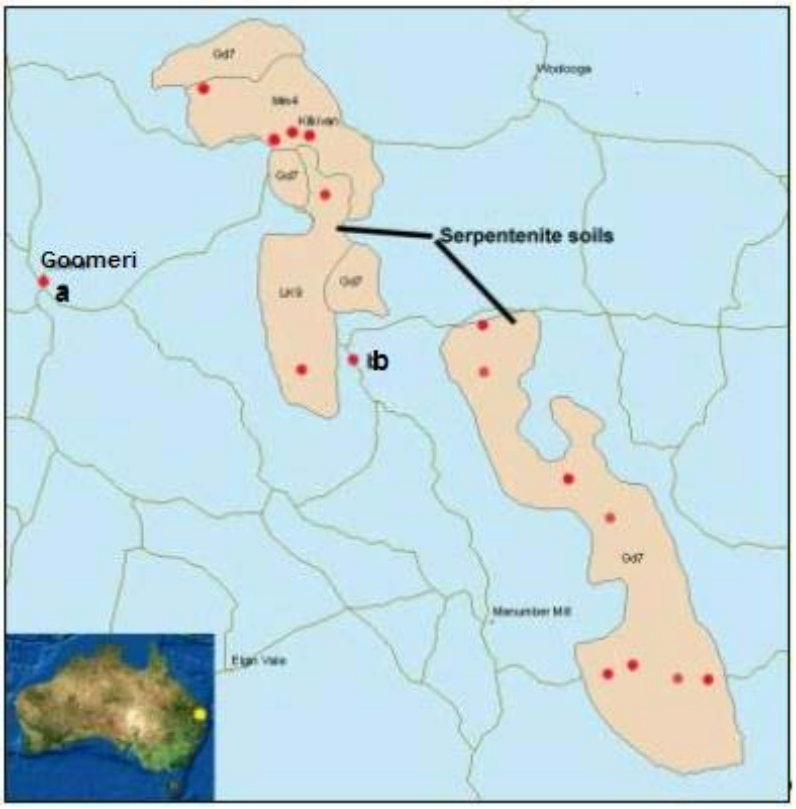


Fig. 16. *Enregistrements d'espèces (en rouge) qui ne sont trouvés que dans les milieux hautement minéralisé de sols de Serpentenite. Les résultats 'a' et 'b' comportent sûrement des erreurs de géocodage.*

L'identification des itinéraires des collecteurs (Chapman 1988, Peterson *et al.* 2003) permet de vérifier les erreurs possibles, si par exemple, la date de la collection ne correspond pas au tracé suivi en particulier par c e collecteur. Ceci pourrait être particulièrement utile pour les collecteurs des 18e et 19e siècles, avant que les collecteurs soient capables de couvrir de vastes distances dans une journée par hélicoptère, avion ou véhicules à moteur. Dans l'exemple de la figure 17, les collections entre le 22 et le 25 février et de la première moitié de mars devraient être dans la zone Pentland-Lolworth en Australie (entourée), mais si elles sont en dehors de cette zone, elles sont susceptibles d'inclure des erreurs dans la date de la collection, ou dans le géocode (Chapman 1988). A nouveau, utiliser un SIG pour cartographier à la fois un itinéraire et les enregistrements d'espèces peut être très utile. Un autre exemple est l'utilisation d'un SIG mobile au Népal pour tracer les routes des

collecteurs le long des rivières (Lampe et Reide 2002).

D'autres emplois du SIG incluent par exemple la définition de la zone-tampon *(« buffering »)* délimitant des localisations probables, par exemple, des courants pour les poissons ou de plantes aquatiques, la côte pour les espèces du littoral, les zones d'altitude pour les espèces alpines ou d'autres connues pour avoir des zones d'altitudes alpines distinctes. De cette manière, toute information en dehors de ces zones-tampon devrait nécessiter d'être vérifiée. Il faut y consacrer beaucoup de soin, comme avec les poissons par exemple, il pourrait signifier que les données figurant en dehors de la zone tampon ne sont pas des erreurs du tout, mais que ces espèces pourraient apparaître des courants trop petits pour être rapportés sur une carte. Ces tests peuvent généralement seulement faire ressortir les données suspectes, et c'est ensuite aux contrôles individuels de déterminer ce qui peuvent être de vraies erreurs et celles qui peuvent être des valeurs existant réellement en dehors de la norme.

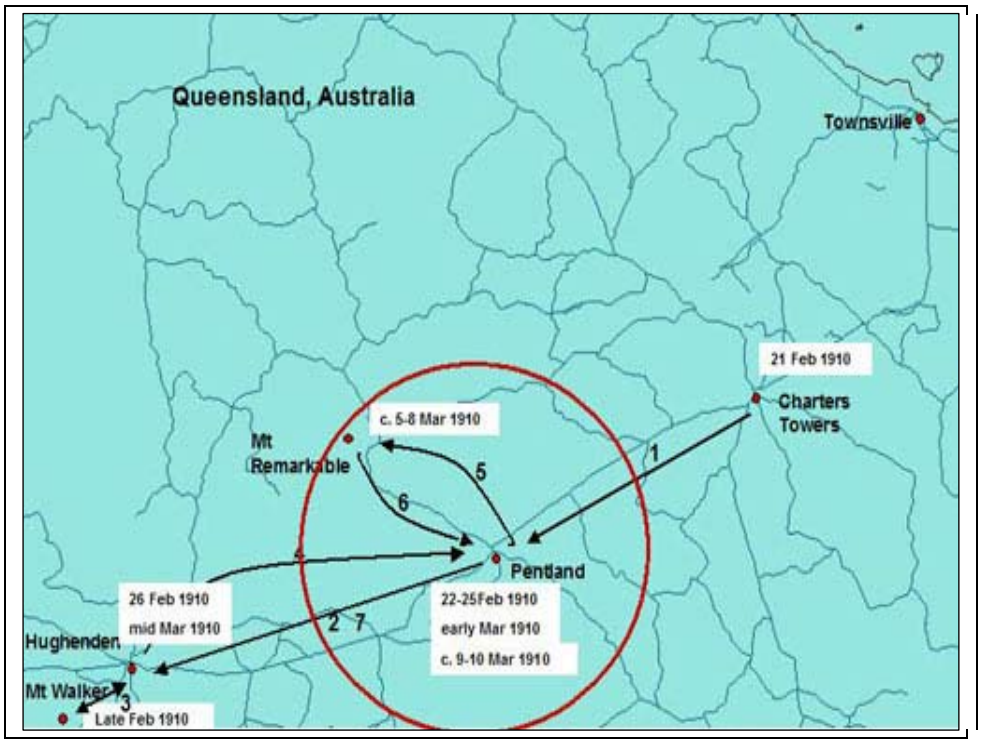


Fig. 17.*Localités de collecte de Karl Domin à Queensland, Australie en 1910 (Chapman 1988). Il voyagea en train de Thionville à Hughenden, en s'arrêtant à Charters Towers et Pentland. Il est ensuite retourné à Pentland, Mount Remarkable, et la région de Lolworth à cheval où il y passa environ dix jours avant de retourner à Hughenden en train. Les dates sont approximatives.*

Valeurs aberrantes dans les espaces géographiques et environnementaux

Il existe un certain nombre de méthode permettant de détecter les valeurs aberrantes dans les données celles–ci, et nous les décrivons ci-dessous. Les données d'histoire naturelle sont très variées et ne se conforment généralement pas à des distributions statistiques standards. Dès lors, comme le suggèrent Maletic et Marcus (2000), plus d'une méthode est souvent nécessaire pour trouver la plupart des valeurs aberrantes.

Détection de valeurs géographiques aberrantes

Un programme du CRIA au Brésil ('spOutiler') permet à l'utilisateur de taper ou de copier/coller des enregistrements dans une boîte sur l'Internet, de les lier vers des fichiers, ou de soumettre un fichier XML d'enregistrement de spécimens et de recevoir des informations sur des valeurs géographiques aberrantes. Les enregistrements sont soumis sous la forme : "id, latitude, longitude, altitude" et le programme retourne les informations sur les erreurs possibles, à la fois sous forme textuelle et de cartes (Marino *et al*. en préparation). Il permet également à l'utilisateur d'identifier les jeux de données soit comme terrestre (on-shore) ou comme maritimes (off-shore) et une fois encore le programme retournera une liste de données non concordantes. Il s'agit d'un programme exceptionnel qui s'avérera très utiles aux biologistes. Il est aussi possible aux utilisateurs de soumettre un document en ligne et de le retrouver annoté d'informations sur des erreurs possibles. Une version en ligne peut être trouvée à http://splink.cria.org.br/tools/ (CRIA 2004b).

Dans la figure 14, une liste de localités a permis de produire quatre enregistrements ayant des erreurs possibles: trois avec une erreur possibles de latitude, une avec une erreur possible d'altitude. Ces points sont ensuite affichés sur la carte associée avec les erreurs possibles identifiées en rouge.

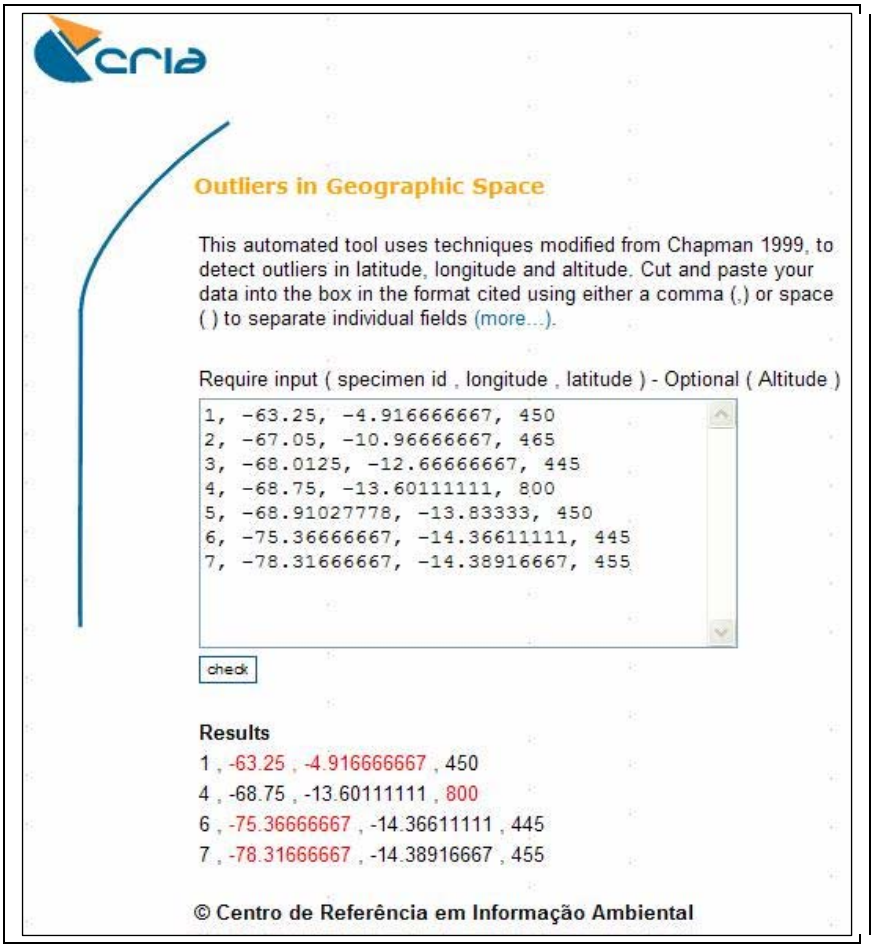


Fig. 18. *Affichage des prototypes de valeurs extrêmes dans le « Geographic Space system » du CRIA identifiant les enregistrements de n° 1, 4, 6 et 7 comme ayant des erreurs possibles*

dans le géocodage.

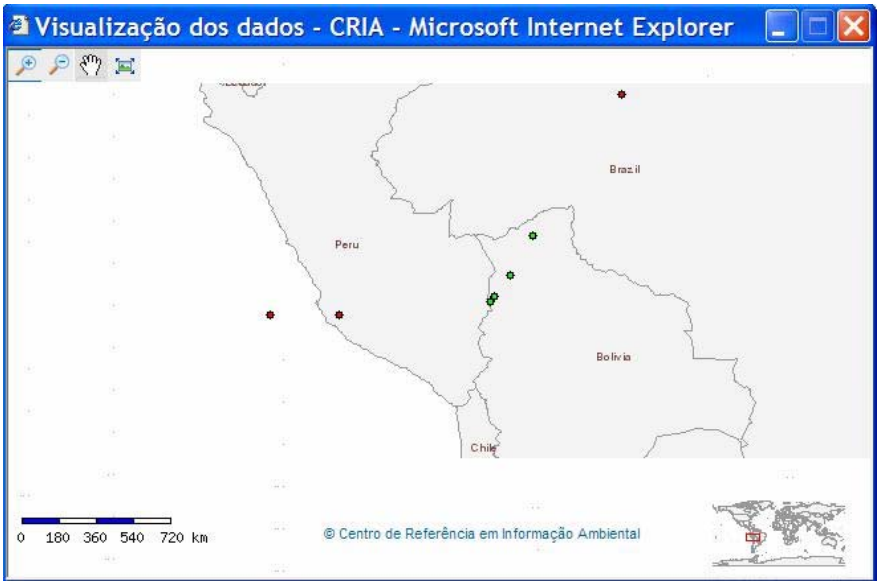


Fig. 19. *Résultat sous forme de carte affichant les enregistrements identifiés comme suspects (en rouge) à partir de la figure 14.*

Des programmes auxquels le public peut accéder utilisent cette méthode:

- **spOutlier**-CRIA (CRIA 2004b, Marino *et al*. in prep).
- **Data Cleaning-CRIA** (CRIA 2005).
- **Diva-GIS** (Hijmans *et al.* 2005)

Courbes cumulatives des fréquences

Les premières versions du programme BIOCLIM (Nix 1986, Busby 1991) étaient utilisées pour détecter les valeurs extrêmes possible en excluant les enregistrements tombant en dehors du 90ème centile de n'importe quel élément du profil climatique pour le taxon, ou en utilisant des courbes cumulatives de fréquence (Busby 1991, Lindemeyer *et al.* 1991) où le pourcentage donné peut varier. Même si ces techniques sont toujours en utilisation et sont faciles d'emploi (Houlder *et al.* 2000, Hijmans *et al.* 2005), elles ne sont pas à utiliser pour les taxons n'incluant pas d'erreurs importantes *(« genuine errors »)*, ou comportant beaucoup d'erreurs. Elles sont aussi suspectes pour les échantillons de petites tailles (Chapman et Busby 1994, Chapman 1999).

Une modification du logiciel Diva-GIS (Hijmans *et al.* 2005) a mené à l'inclusion de la méthode statistique *Reverse Jackknifing* (Chapman 1999) discutée plus bas, et les enregistrements identifiés grâce à cette méthode ont été reliés aux courbes cumulatives de fréquence en soulignant chaque paramètres .

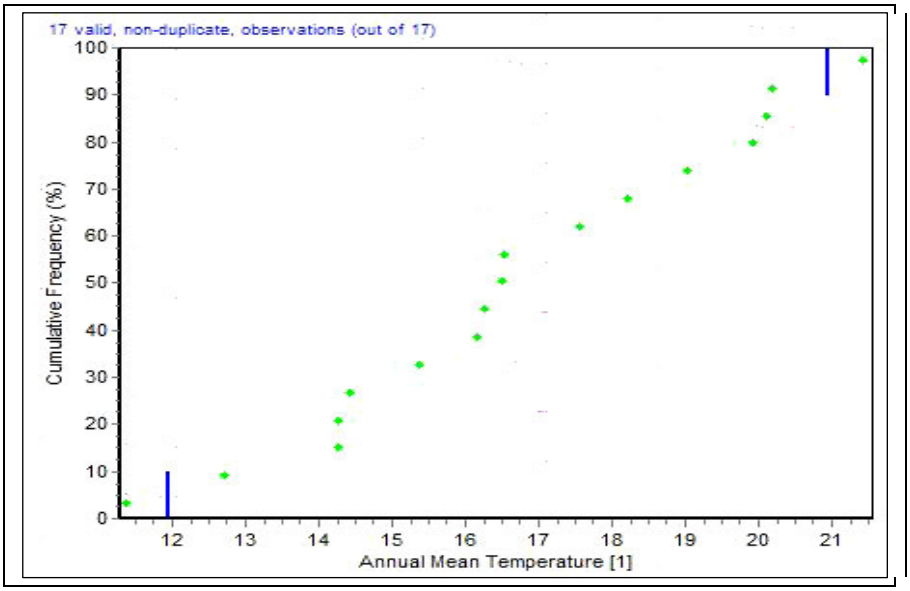


Fig. 20. *Courbe cumulative de fréquence utilisée pour détecter des valeurs extrêmes dans les espaces climatiques en utilisant les températures annuelles moyennes. La ligne bleue représente 97,5%, le point en bas à gauche (ou même les deux points en bas à gauche), peut être vu comme une valeur extrême possible nécessitant une vérification de géocode.*

Programmes accessibles au public utilisant cette méthode:

- **Diva-GIS** (Hijmans *et al.* 2005)
- **ANUCLIM** (Houlder *et al.* 2000).

Analyse en composantes principales (ACP)

En utilisant la dispersion *(« scatter »)* des points dans l'analyse en composantes principales d'une couche climatique en comparaison à une autre, on peut aider l'identification des valeurs extrêmes possibles et donc des erreurs de géocodes possible. C'est une méthode de validation de données assez puissante mais, à moins que le processus ne soit automatisé d'une façon ou d'une autre pour l'identification de multiples enregistrements anormaux, la méthode peut être assez fastidieuse puisqu'il faudrait passer en revue toutes les combinaisons de composants climatiques possibles utilisées.

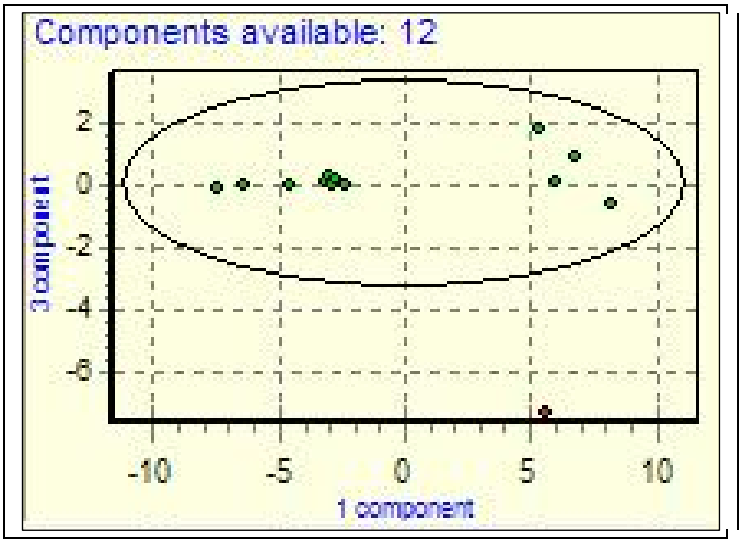


Fig. 21. *Analyse de Composants Principaux, montrant un point (rouge) identifié comme étant une valeur extrême et donc comme étant une erreur possible (FloraMap, Jones et Gladkov 2001).*

Programmes accessibles au public utilisant cette méthode:

- **FloraMap** (Jones et Gladkov 2001)
- ***PATN*** *vers. 3.01 (Belbin 2004)*

Partitionnement de données (« Cluster *Analysis »)*

L'identification de valeurs extrêmes en utilisant des distances euclidiennes ou d'autres unités de mesure de distance peut parfois identifier des valeurs extrêmes qui n'auraient pas été identifiées par des méthodes au niveau du champ (Johnson et Wichern 1998, Maletic et Marcus 2000). Le partitionnement de données peut être utilisée pour aider à identifier les groupes multiples de population semblables (en utilisant les espaces climatiques ou d'autres critères), et peut donc également être utilisé pour identifier des groupements isolés qui sont soit des cas uniques *(« unicates »)* soit des petits groupes séparés d'autres groupes par une distance significative. À nouveau, c'est une méthode assez utile et apparemment assez robuste, mais qui dépend beaucoup de la méthode de clustering utilisée et qui peut être complexe à calculer informatiquement.

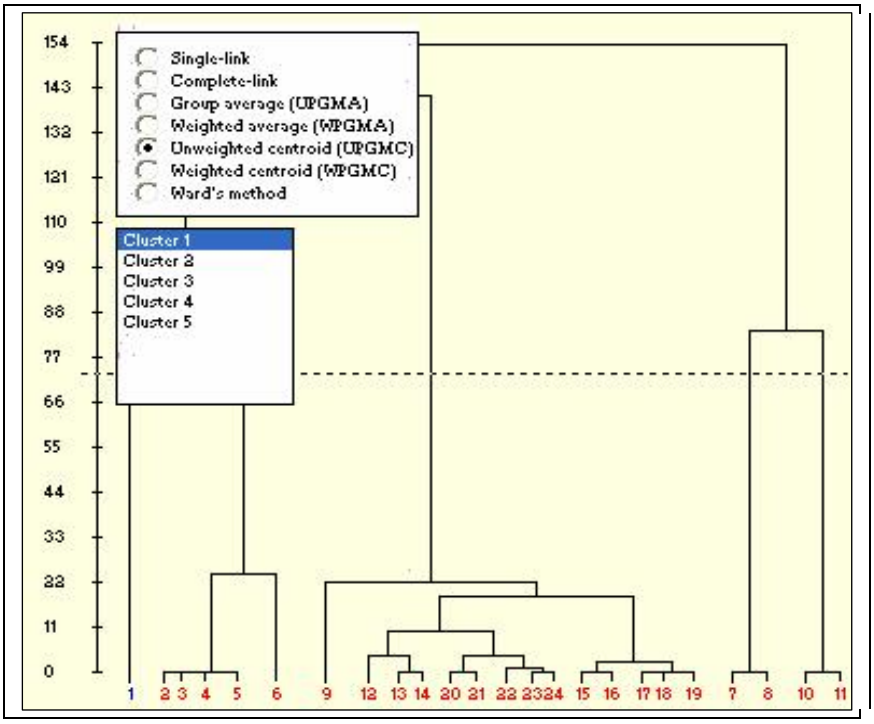


Fig. 22. *Partitionnement des données montrant un groupe unique («unicate cluster ») (#1 – en bleu) qui peut être vu comme une valeur extrême (FloraMap, Jones et Gladkov 2001).*

Programmes à accès public utilisant cette méthode :

- **FloraMap** (Jones et Gladkov 2001)
- **PATN** Vers. 3.01 (Belbin 2004)

Enveloppe Climatique

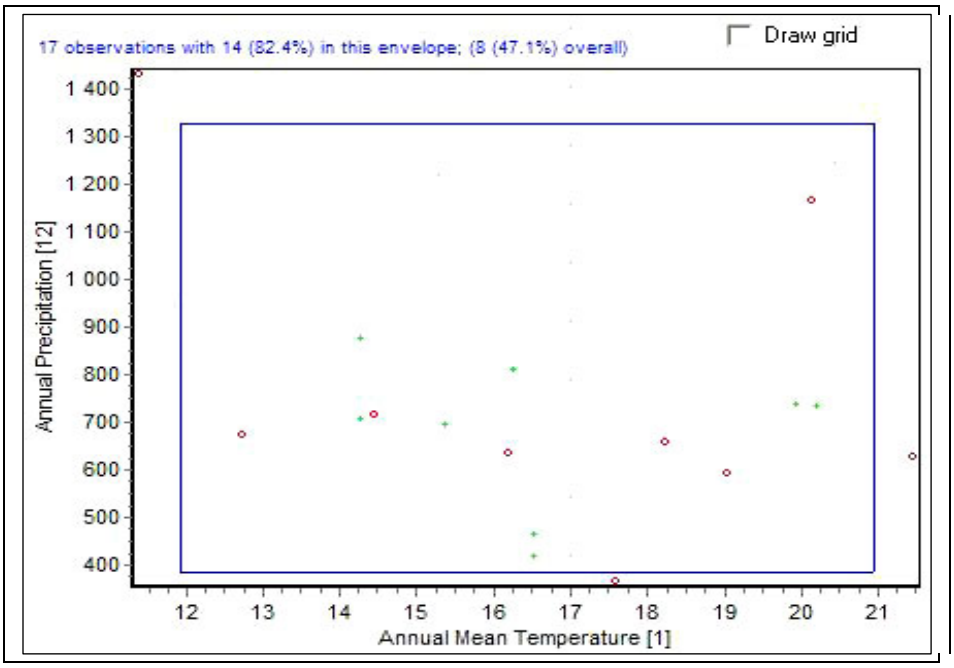


Fig. 23. *Enveloppe Climatique pour BIOCLIM en utilisant une enveloppe au 97,5ème centile des températures et précipitations annuelles moyennes. Les résultats marqués en rouge sont ceux tombant en dehors des 64 enveloppes possibles.*

La méthode d'enveloppe climatique est une extension de la méthodologie des courbes cumulatives de fréquence citée ci-dessus, mais regroupe chacune des couches climatiques dans une boite multi-dimensionnelle ou enveloppe pouvant permettre d'examiner deux dimensions à la fois, de façon similaire à l'analyse en composantes principales. Des valeurs extrêmes dans les courbes cumulatives de fréquence formant la totalité des couches climatiques peuvent être identifiées de cette manière.

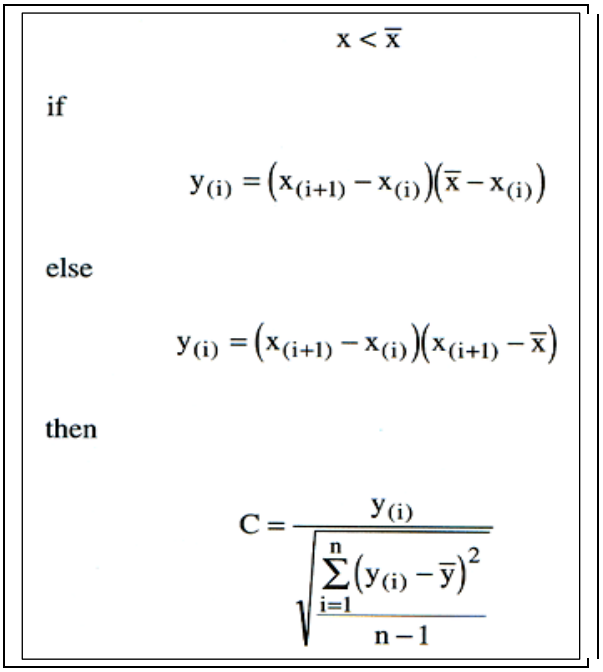
Programmes accessible au public utilisant cette méthode:

- ***Diva-GIS*** *(Hijmans et al. 2005)*

« Jackknife » inverse

Cette technique utilise une version modifiée de l'algorithme de *« Jackknife » inverse* pour extraire les valeurs aberrantes à la fin d'une série de points dans n'importe quel nombre de profils climatiques. En 1992, cette méthode était développée à ERIN en Australie pour automatiquement détecter les valeurs aberrantes d'espaces climatiques (Chapman 1992, 1999, Chapman et Busby

1994) et donc identifier les enregistrements suspects parmi les milliers d'espèces modelées à l'époque. Cette méthode s'est avérée extrêmement fiable dans l'identification automatique des enregistrements, avec une forte proportion (environ 90%) des enregistrements identifiés comme suspects s'avérant être de vraies erreurs.



$$x < \bar{x}$$

if

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(\bar{x} - x_{(i)}\right)$$

else

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(x_{(i+1)} - \bar{x}\right)$$

then

$$C = \frac{y_{(i)}}{\sqrt{\frac{\sum_{i=1}^{n}\left(y_{(i)} - \bar{y}\right)^2}{n-1}}}$$

•**Fig. 24**. *Formule pour déterminer la Valeur Critique (C) dans un algorithme de détection des valeurs extrêmes où C= Valeur Critique (Chapman 1999). Cette formule a été utilisée en Australie depuis 1992 pour détecter les valeurs extrêmes dans les milieux environnementaux (climatiques). Cette formule a récemment été modifiée (2005) en divisant la valeur de C par la variable 'x' et a été incorporée dans la version 5.0 de Diva-GIS (Hijmans et al. 2005). Ceci a amélioré sa fiabilité dans l'utilisation avec de paramètres ayant des ordres de grandeur importants comme les précipitations ou l'altitude.*

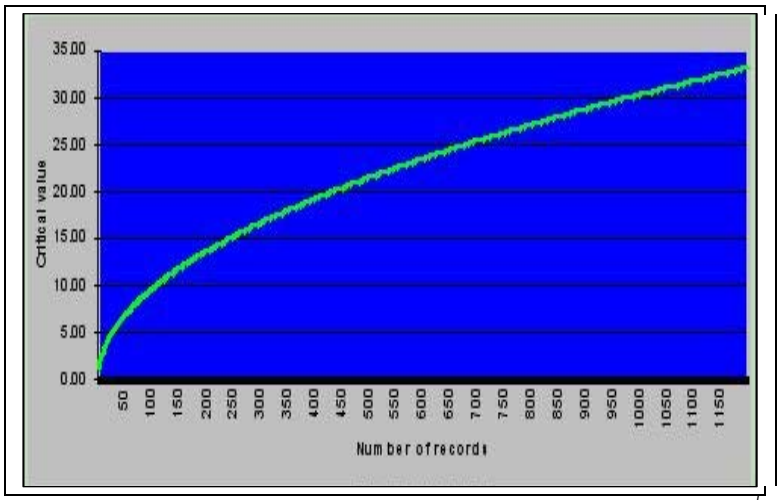


Fig. 25. *Courbe de valeurs de seuil (Threshold Value Curve), (T=0.95(√n)+0.2 ou 'n' est le nombre d'enregistrements). Les valeurs au dessus de la courbe sont regardées comme suspects, les valeurs en dessous comme "valides" (Chapman 1999).*

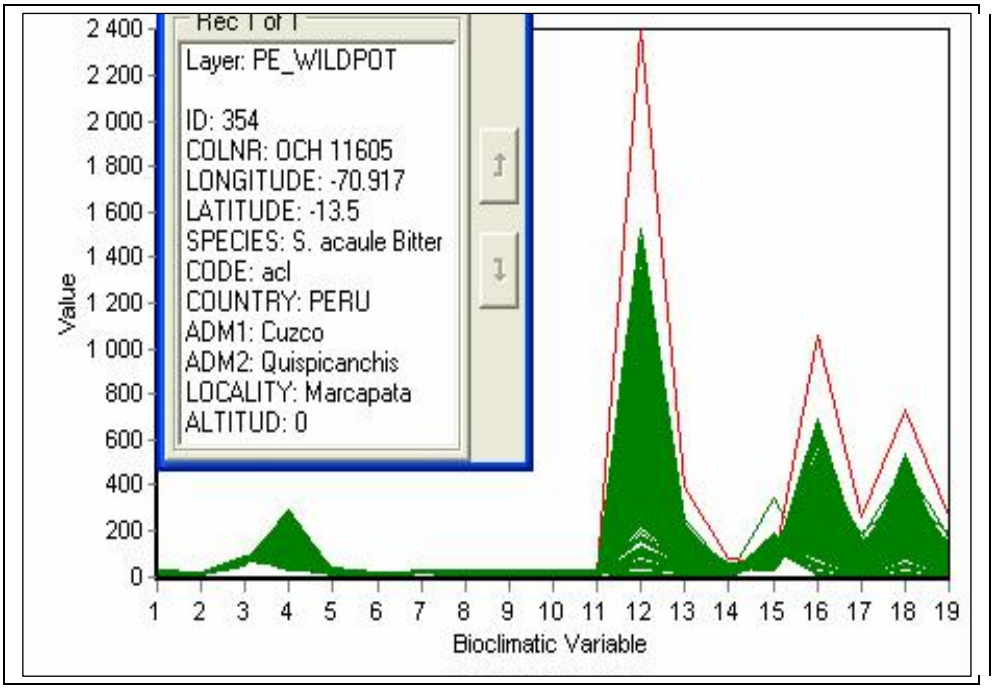


Fig. 26. *Algorithme de détection de valeurs extrêmes de Diva-GIS utilisant la méthode du « Reverse Jackknife ». Le programme a identifié une valeur extrême possible (en utilisant l'option sélectionnée pour montrer seulement les enregistrements qui semblaient douteux selon au moins 6 des 19 critères possibles).*

Programmes accessibles au public utilisant cette méthode:

- **Diva-GIS** (Hijmans *et al.* 2005)
- Également en programmation dans le nouveau kit d'outils de BioGeomancer et qui devrait être disponible mi 2006

Valeurs Extrêmes de Paramètre

Les valeurs extrêmes de paramètre sont une méthode similaire à celle d'enveloppe climatique qui identifie les enregistrements aux extrêmes de chaque courbe de fréquence cumulative et les mets en paquet dans un fichier de résultats journalisé. De cette manière, il est possible d'identifier des enregistrements particuliers qui sont extrêmes dans plus d'un paramètre climatique.

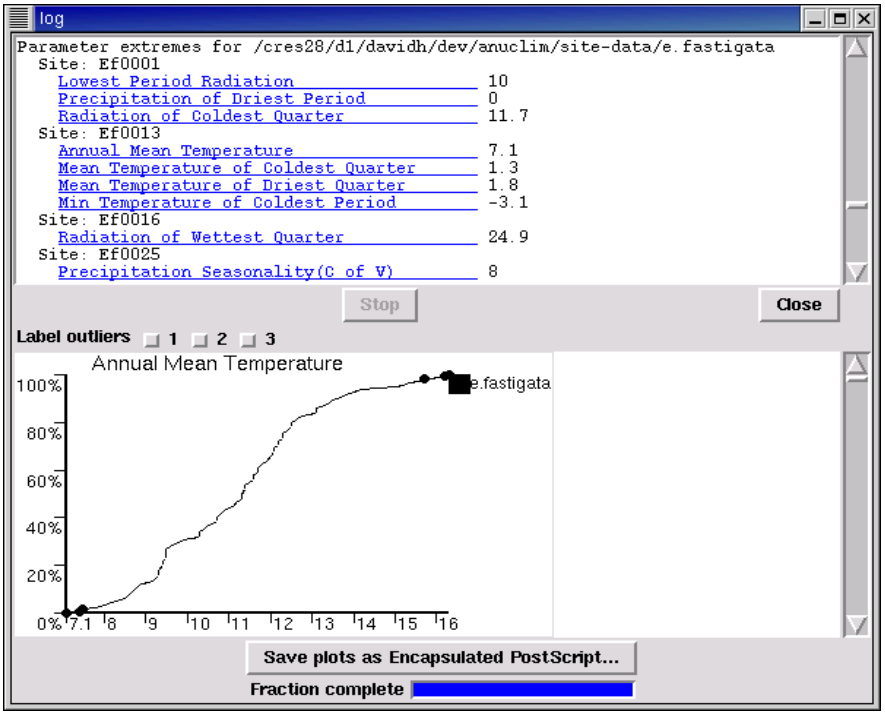


Fig 27. *Fichier journalisé issu de la version 5.1 de ANUCLIM (Houlder et al. 2000) montrant les valeurs extrêmes des paramètres pour l'Eucalyptus fastigata (en haut) et une courbe d'accumulation de l'espèce associée (en bas).*

Programmes accessibles au public utilisant cette méthode:

- **ANUCLIM** (Houlder *et al.* 2000).

Autres Méthodes

Beaucoup des méthodologies listées ci-dessous sont très simples et disponibles dans de nombreux paquets d'outils statistiques de référence. Certaines ne semblent pas avoir été utilisées dans la détection des erreurs dans les données biologiques, mais des exemples avec des types de données similaires indiquent qu'il vaudrait la peine de l'essayer. Un certain nombre de ces méthodes ainsi que d'autres sont élaborées chez Legendre et Legendre (1998). Un certain nombre d'autres méthodes de détection de valeurs extrêmes qu'il serait bon d'essayer peuvent être trouvées dans les publications de Barnett et Lewis (1994).

Ecarts types par rapport à la Moyenne

La plus prometteuse de ces autres méthodes serait peut être de rechercher un certain nombre de dérivations standards à partir d'une valeur moyenne d'après le théorème de Chebyshev (Barnett and Lewis 1994). Maletic et Marcus (2000) on testé un certain nombre de d'écarts par rapport à la moyenne en utilisant 5000 enregistrements navals personnels, avec 78 champs du même type (date) et trouvé qu'utiliser 5 fois l'écart type générait les meilleurs résultats. Il serait nécessaire de mener des tests sur un certain nombre de jeux de données sur les collections, et particulièrement de le faire

pour un nombre beaucoup plus faible de données que Maletic et Marcus. Des tests préliminaires avec des chiffres bas que j'ai menés en utilisant des données d'altitude n'ont jusqu'ici pas paru prometteurs.

Ecarts types par rapport à la Médiane

Un autre groupe de tests statistiques non-paramétriques utilise la relation avec la médiane plutôt que la moyenne. Deux méthodes possibles sont le test «U» de Mann-Whitney et le test de Kuskall-Wallis qui s'intéressent aux hypothèses alternatives où deux (Mann-Whitney), trois ou plus (Kuskall-Wallis) populations diffèrent seulement en fonction de la médiane (Barnett and Lewis 1994, Lowry 2005). Je n'ai pas vu de tels exemples appliqués à la détection de valeurs extrêmes dans les données de sous-espèces, mais des tests pourraient valoir le coup d'être menés.

Utilisation de Distributions Modélisées

Les modèles de distribution dérivés de la modélisation de distribution des espèces tels que ceux produits par l'utilisation de GARP (Stockwell et Peters 1999, Pereira 2002) ou Lifemapper (Université du Kansas 2003b), pourraient être utilisé pour identifier des nouveaux enregistrements tombant en dehors de la distribution prédite. Cette méthode, bien que facile à utiliser, est limitée par la qualité de la distribution prédite. Si tous les enregistrements d'une espèce n'ont pas été utilisés pour développer le modèle, alors il peut comporter des déficiences. De plus, utiliser simplement les frontières externes de la distribution ne prend pas en compte la nature isolée des bons modèles qui identifient les niches adéquates au sein de la totalité de la distribution géographique.

Analyse de motifs *(« Pattern Analysis »)*

L'analyse de motifs peut être utilisée pour identifier les enregistrements ne se conformant pas à des motifs déjà existants dans la base de données. Une variété de méthodes peut être utilisée pour l'analyse des motifs, dont l'association, le partage, la classification, le regroupement, l'ordination et l'utilisation de réseaux tels que des arbres couvrants de poids minimal (Belbin 2004). Certaines de ces méthodes ont été abordées plus en détail précédemment. Un motif peut généralement être défini comme un groupe de rapports ayant des caractéristiques similaires (Maletic et Marcus 2000), mais le choix du "bon motif de référence", s'il existe, peut influencer le résultat (Weiher et Keddy 1999).

Programmes accessibles au public utilisant cette méthode:

- **PATN** (Belbin 2004)

Règles d'association ordinale

Les règles d'association tentent de trouver la relation ordinale qui a tendance à reporter un grand pourcentage de données (Marcus *et al.* 2001). Elles peuvent être utilisées pour les données catégoriques et quantitatives. En d'autres mots, elles recherchent des comportements types tels que si la plupart du temps A<B, alors si dans un cas A>B il s'agit sûrement d'une erreur. Avec les données quantitatives, les règles peuvent être utilisées en accord avec d'autres méthodes statistiques qui utilisent la moyenne, la médiane, l'écart-type et les tranches de centiles pour la détection des valeurs extrêmes. Avec cette méthode, plus on a de données, meilleurs seront les résultats. Dans beaucoup de cas, elle peut être utilisée sur une base de données entière et non pas seulement sur les données d'une seule espèce. Par exemple, si une espèce A dans une Végétation de type B la plupart du temps, un enregistrement montrant qu'elle se manifeste dans une Végétation de type C pourrait être une erreur. Également, toutes les données collectées par une même personne ne

devraient pas s'étendre sur les 15 premières années de vie du collecteur ni dépasser 100 année de vie ou bien dépasser la date de décès. De telles lois peuvent également être utilisées en conjonction du périmètre d'activité d'un collecteur (voir ci-dessus). Par exemple, si une collection date d'avant 1900, alors deux collections rassemblées par les mêmes personnes le même jour ne devraient pas être séparées de plus de x kilomètres.

Programmes en accès public utilisant cette méthode:

- **PATN** (Belbin 2004).

Données descriptives

Les contrôles d'erreurs dans les données descriptives sont plus compliqués à couvrir ici de par la nature très variée de ce qui peut être inclus dans de telles bases de données. La nature structurée de ces bases de données lorsqu'elles existent autorise cependant l'établissement de plus de règles .

Conception des bases de données

La clé pour maintenir une bonne qualité de données dans les bases de données descriptives est de suivre de bonnes procédures de conception, et lorsque c'est possible de concevoir les bases de données à partir de standards tels que DELTA (Dallwitz *et al*. 1993) ou le nouveau standard SDD (*Structure of Descriptive Data*, structure de données descriptives : http://160.45.63.11/Projects/TDWG-SDD/) qui est en train d'être développé par le Groupe de Travail de Bases de Données Taxonomiques, *Taxonomic Databases Working Group* (TDWG).

Contrôles d'édition

De par la nature structurée des bases de données descriptives, elles laissent place à l'utilisation de contrôles d'édition. Par exemple, la plupart des champs de données descriptives ont des contraintes variées inhérentes, et ont souvent un jeu de caractères ou de symboles bien développés à partir desquels les entrées sont choisies. Des erreurs peuvent toujours se produire cependant, particulièrement avec des valeurs continues dont les unités peuvent être confondues (ex: millimètres et centimètres). Les unités utilisées devraient être enregistrées et préférablement dans un champ séparé comme recommandé dans le standard SDD. De plus, la standardisation des unités au sein d'une base de données devrait être menée à tous les endroits possibles. Par exemple, décider d'utiliser les « mm » en permanence, ou les « cm », etc., plutôt que de mélanger les deux ce qui peut mener à des erreurs, particulièrement lorsque que la saisie de données est effectuée par plusieurs personnes. Des tests peuvent être menés sur ces champs pour chercher les extrêmes (exemple: en utilisant les courbes cumulatives de fréquence décrites dans *Données Spatiales*, ci-dessus), pour regarder les valeurs extrêmes en utilisant les écarts-types de la médiane ou de la moyenne, etc… Souvent, en mettant les résultats sous forme graphique, il est possible d'identifier les rapports porteurs d'erreurs possibles. Certains autres types d'erreurs permettant d'identifier des occurrences d'erreurs incluent (English 1999):

- ***Valeurs de données manquantes***
  Recherche des champs vides où des valeurs devraient être affichées. Là où il y a besoin de faire appel à une valeur manquante ou « nulle » dans un champ, conserver les raisons de son utilisation dans un champ séparé est une bonne pratique : par exemple "pas pertinent, pas mesuré ou inconnu"...
- ***Valeurs de données incorrectes***
  Cela implique la recherche des erreurs typographique, les interversions systématiques de touches de claviers tapées par erreur, de données entrées au mauvais endroit (ex: caractères alphanumériques entrés dans un champ numérique), et les valeurs de données entrées de force dans un champ qui réclame toujours une valeur mais pour lequel l'opérateur de données ne connaît pas la valeur et donc ajoute une valeur fictive. Des valeurs arbitraires sont parfois ajoutées dans des champs pour "tromper" les méthodes statistiques là où les champs vides ou les valeurs égales à zéro ne sont pas autorisées. Ceci réclame un soin tout particulier.
- ***Valeurs de données non-atomiques***

Recherche des champs où plus d'un fait du monde réel est saisi.

- ***«Champs Schizophrènes »***
  Recherche des champs utilisés pour une fonction pour laquelle ils n'avaient pas forcément été conçus.
- ***Occurrences doublées***
  Recherche des valeurs qui pourraient se référer au même objet du monde réel. Ceci peut se produire assez souvent lorsqu'on combine deux bases de données utilisant des terminologies différentes.
- ***Valeurs de données contradictoires***
  Cela se produit quand deux bases de données reliées entre elles n'utilisent pas forcément les mêmes listes de valeur, et lorsque combinées elles montrent des incohérences. C'est ici que l'utilisation de standards de transferts tels que le SDD mentionné ci-dessus rentre en jeu.

Documentation des erreurs

Comme mentionné dans le document associé sur les Principes de *Qualité des Données, «Principles of Data Quality »* (Chapman 2005a), la documentation et la vérification des données est essentielles pour maintenir la qualité des données et éviter de dupliquer les vérifications d'erreurs. Sans une bonne documentation, les utilisateurs ne peuvent pas déterminer si les données sont aptes à l'utilisation.

Il est assez peu utile pour tout le monde que les contrôles de qualité de données soient entrepris, et les corrections sont apportées, s'ils ne sont pas complètement documentés (Chapman 2005a). Un système de traçabilité historique et d'audit des corrections de données est nécessaire puisque qu'il est toujours possibles que les erreurs perçues ne soient en fait pas des erreurs et que des changements apportés amènent de nouvelles erreurs. Sans trace des ces changements, il ne serait pas possible de d'annuler ces "corrections". Ceci est particulièrement important lorsque ces changements sont effectués par une autre personne que celle à l'origine de la production des données (Chapman 2005a).

Il existe différentes façons de développer des fichiers d'historique ou d'audits des modifications (par exemple, enregistrer les changements apportés à la base de données sur la durée tout en enregistrant quels contrôles de qualités de données ont été menés et quand). Ces systèmes de traçages sont importants pour que les erreurs puissent être retrouvées et que les gestionnaires de données ne mènent pas de contrôles déjà effectués, et aussi pour que les altérations et ajouts aux données soient documentés pour des raisons légales ou autres (par exemple, informer utilisateurs qui auraient peut être utilisé les données en faisant connaître quels changements ont pu être apportés depuis la dernière fois qu'ils y avaient accédé). Une façon de créer ces journaux de bords et audit est de constituer n d'une base de données temporelle par l'ajout d'une série d'horodatages *(« timestamps »)*. Ainsi, un horodatage transactionnel et événementiel durant laquelle les données devraient être stockées dans une base de données (Wikipedia[9]). Une autre méthode est de faire des exportations XML périodiques des enregistrements qui ont changé, ou de portions de données où des changements ont été effectués.

Comme mentionné chez Chapman (2005a):

> *"les contrôles de qualité des données entrepris sur des données par un utilisateur peuvent identifier un certain nombre d'enregistrements suspects. Ces enregistrements peuvent alors être vérifiés et s'avérer être des données correctes comportant des cas réels de valeurs extrêmes. Si cette information n'est pas documentée dans l'enregistrement, quelqu'un d'autre risque ultérieurement d'entreprendre d'autres tests qui identifieraient à nouveau les mêmes enregistrements comme suspects."*

Tout comme mentionné dans le document associé sur les *Principes De Qualité des Données* (*Principles of Data Quality,* Chapman 2005a):

> *Une façon d'être sûr que les erreurs sont complètement documentées est de l'inclure dans les premières étapes planifiant la conception et la construction de la base de données. Des champs additionnels de données de qualité/fiabilité de données peuvent alors être incorporés.*

[9] http://en.wikipedia.org/wiki/Temporal_database

Des champs tels que la fiabilité des géocodes la, source de l'information pour le géocode et l'altitude, la mention de la personne ayant ajouté l'information.«Est-ce que le nouveau géocode a été ajouté par un collecteur utilisant un GPS, ou un opérateur de saisie de données à une date ultérieure utilisant une carte à une échelle particulière ? », « Est-ce que l'altitude a été générée automatiquement par DEM, et si oui, quelle était la source du GEM, sa date, son échelle?» etc…. Toutes ces informations seront utiles pour déterminer plus tard si l'information est importante pour un usage particulier, et l'utilisateur de la donnée peut ainsi en décider.

De plus, des champs dédiés à la validation des données -le "qui, quand, comment et quoi" des contrôles de validation menés devraient être ajoutés à la base de données pour tracer et auditer la validation, les vérifications d'erreurs et le nettoyage de données menés sur la base de données. Idéalement, cela devrait également être ajouté au niveau de l'enregistrement comme mentionné précédemment.

Visualisation de l'Erreur

Il reste encore beaucoup de travail à effectuer afin de développer de bonnes méthodes de visualisation des erreurs en matière de données primaires d'espèce. Les deux exigences pour la visualisation sont:

- Visualisation pour la vérification et le nettoyage des erreurs
- Visualisation pour la présentation

Le deuxième, visualisation pour la présentation, a été couvert dans le document sur les *Principes de Qualité des Données* (Chapman 2005a).

Le SIG est l'outil le plus commun pour visualiser les erreurs spatiales. Faire simplement apparaître sur une carte les données primaires d'espèce en les superposant à des couches topographiques peut assister la détection des erreurs. Les systèmes de SIG vont des simples systèmes *on-line* utilisés surtout pour la mise en carte *en-ligne* et la présentation d'informations jusqu'aux systèmes autonomes qui varient du plus simple au plus hautement sophistiqué.

De nombreuses institutions utilisent déjà un SIG pour la production de cartes, et ceux-ci sont facilement adaptables à une utilisation dans le cadre de la vérification des erreurs. D'autres institutions, cependant, n'utilisent pas de SIG régulièrement et considèrent que l'acquisition d'un GIS ne correspondrait pas à leur besoins, mais il existe des programmes SIG gratuits qu'il est facile d'apprendre et simples à utiliser; qui répondraient adéquatement à la plupart des exigences des petites institutions. Au moins un de ceux là – Diva-GIS (Hijmans *et al.* 2005) – a été spécialement créé pour être utilisée par des petits musées ou herbiers et inclut différentes méthodes de détection des erreurs décrites dans ce document, tout comme des algorithmes de visualisation et de modelage.

Pour les données non spatiales, la meilleure façon de visualiser les erreurs est par l'utilisation de feuilles de calculs dans des logiciels tableurs et de graphiques. Un simple graphique de valeurs identifiera rapidement les enregistrements ne correspondant pas à un comportement type. Les graphiques simples sont faciles à mettre en place et à remplir à partir de bases de données et représentent ainsi une méthode standard de vérification des erreurs.

Il y a une tendance grandissante dans la communauté spatiale à utiliser des techniques telles que l'analyse de Monte Carlo *(« Monte Carlo Analysis »)* pour produire des estimations sur l'étendue et l'importance des erreurs (Flowerdew 1991). Cette analyse se prête à une bonne visualisation, et il s'agit d'une bonne façon de transmettre les erreurs aux utilisateurs. Même si certains logiciels courants qui incluent les méthodes de Monte Carlo sont devenus assez chers (ex: Canoco 4.5 pour Windows[10] et S-Plus[11]), il existe des alternatives gratuites, par exemple le gadget PopTool pour Microsoft Excel (Hood 2005).

Visualiser l'exactitude

Comme mentionné dans la partie *Géoréférencement* ci-dessus, les enregistrement sous forme de points de données primaires d'espèce ne sont pas vraiment des points, mais ont une forme d'erreur qui leur est associée. En faisant apparaître le point avec la précision associée(« l'empreinte »), une

[10] http://www.microcomputerpower.com/
[11] http://www.microcomputerpower.com/

bonne compréhension de ce que la collection signifie et sa relation avec le monde réel peuvent aussi être visualisés.

C'est une zone de recherche qui a un besoin urgent d'être approfondie à l'égard des données primaires d'espèce- le développement de techniques pour visualiser l'incertitude et montrer des empreintes d'exactitude. Au lieu de représenter enregistrement de collection comme un point de latitude et de longitude il faut inclure la précision qui lui est associée et donc présenter la localisation comme son empreinte au moyen d'un cercle, d'une ellipse, d'un polygone ou même d'une grille. Les techniques GIS, comme la délimitation de zones-tampon, représentent un bon outil pour développer des empreintes le long de rivières ou de routes. Le programme « Biogeomancer » traite certains de ces aspects, mais il est peu probable qu'il mène à un système complètement opérationnel en temps disponible[12].

[12] Note du traducteur (FT) : à l'heure de la traduction française de ce document (2010) le développement de BioGeomancer a continué et l'équipe de développement travalel toujours à son amélioration (notamment en ce qui concerne le traitement des données par lots (batchs)

Outils Cités

1. Logiciels

ANUCLIM

Description: Un progiciel de modelage bioclimatique contenant un ensemble de programmes dont la version la plus récente de BIOCLIM. Le programme inclut un certain nombre de méthodes d'identification des erreurs dans les données-sources de spécimens.
Version: 5.1 (2004).
Détenteur: Centre de Ressources et d'Etudes Environnementales, *Centre for Resource and Environmental Studies* (CRES), Université Nationale Australienne, Canberra, Australie.
Coût: $AUD1000.
Auteur: Houlder et al. 2000.
Téléchargement: http://cres.anu.edu.au/outputs/software.php

BioLink

Description: Un progiciel conçu pour gérer les informations basées sur les taxons telles que la nomenclature, la distribution, la classification, l'écologie, la morphologie, les illustrations, les multimédia et la littérature.
Version: 2.1 (2005).
Détenteur: Collection d'Insectes Nationale Australienne, *Australian National Insect Collection*, CSIRO, Canberra, Australie.
Coût: Gratuit.
Auteur: Shattuck et Fitzsimmons 2000.
Téléchargement: http://www.biolink.csiro.au/.

BIOTA

Description: Un système de gestion de données de biodiversité et de collection. Son interface graphique facile à utiliser exploite le pouvoir d'une base de données entièrement relationnelle.
Version: 2.03 (2004).
Détenteur: Robert K. Colwell, Connecticut, USA.
Coût: Version de Démonstration: gratuite: Version complète: $US200-600.
Auteur: Collwell 2002.
Téléchargement: http://viceroy.eeb.uconn.edu/Biota2Pages/biota2_download.html

Biótica

Description: Conçu pour gérer les départements de conservation, les données nomenclaturales, géographiques et écologiques et pour assister dans la capture des données et la mise à jour.
Version: 4.0 (2003).
Détenteur: CONABIO, Mexico City, Mexique.
Coût: $US290.
Auteur: Conabio 2002.
Téléchargement :
http://www.conabio.gob.mx/informacion/biotica_ingles/doctos/distribu_v4.0.html.

BRAHMS

Description: Un logiciel de base de données pour la recherche botanique et la gestion des collections. Il fournit une aide pour la gestion des noms, l'assainissement des collections et les recherches taxonomiques.
Version: 5.58 (2005).
Détenteur: Université d'Oxford, Oxford, UK.
Coût: Gratuit.
Auteur: Université d'Oxford 2004.
Téléchargement: http://storage.plants.ox.ac.uk/brahms/defaultNS.html.

Desktop GARP

Description: Un progiciel pour prédire et analyser la distribution des espèces sauvages (conception de modèles prédictifs de distribution).
Version: 1.1.3 (2004)
Détenteur: Université du Kansas, Lawrence, Kansas, USA et Centro de Referência em Informação Ambiental (CRIA), Campinas, Brésil.
Coût: Gratuit.
Auteur: Pereira 2002.
Téléchargement: http://www.lifemapper.org/desktopgarp/Default.asp?Item=2&Lang=1.

Diva-GIS

Description: Un système d'information géographique développé pour l'analyse des données de biodiversité. Il inclut plusieurs outils de modélisation simples et plusieurs algorithmes de vérification de qualité des données.
Version: 7.2.2(2010).
Détenteur: R.J. Hijmans *et al.,* Université de Californie, Berkeley.
Coût: Gratuit
Auteur: Hijmans *et al.* 2005
Téléchargement: http://www.diva-gis.org

eGaz

Description: Un programme développé pour assister les musées et les herbiers dans l'identification et l'ajout de géocodes à leurs données de spécimens.
Détenteur: Australian National Insect Collection, CSIRO, Canberra, Australia.
Coût: Gratuit
Auteur: Shattuck 1997.
Téléchargement: http://www.biolink.csiro.au/egaz.html

FloraMap

Description: Un outil logiciel pour prédire la répartition des plantes et d'autres organismes à l'état sauvage.
Version: 1.02 (2003).
Détenteur: Centro Internacional de Agricultura Tropical (CIAT), Columbie.
Coût: $US100.
Auteur: Jones et Gladkov 2001.
Téléchargement: http://www.floramap-ciat.org/ing/floramap101.htm.

GeoLocate

Description: Un programme de géoréférencement pour faciliter la tâche d'attribution de coordonnées géographiques aux données de localités associées à des collections d'histoire naturelle.

Version: 2.0 (2003).
Détenteur: Museum Tulane d'Histoire Naturelle, Belle Chasse, LA, USA.
Coût: Gratuit.
Auteur: Rios et Bart *n.dat*.
Commande: http://www.museum.tulane.edu/geolocate/order.aspx.

PATN

Description: Un progiciel complet et polyvalent pour extraire et afficher les comportements–types *(« patterns »)* dans les données multivariées.
Version: 3.01 (2004).
Custodian: Blatant Fabrications Pty Ltd (Lee Belbin)
Coût: $US299.
Auteur: Belbin 2004.
Téléchargement: http://www.patn.com.au/.

PopTools

Description: Module polyvalent pour versions PC de Microsoft Excel qui facilite l'analyse des matrices de modèlse de populations et les processus simulatoires et stochastiques.
Version: 2.6.6 (2005).
Détenteur: Greg Hood, Albany, W.A., Australie.
Coût: Gratuit
Auteur: Hood 2005
Téléchargement: http://www.cse.csiro.au/poptools/.

Specify

Description: Un système de gestion de collection pour les musées d'histoire naturelle et les herbiers.
Version: 4.6 (2004).
Détenteur: Biodiversity Research Center, Université du Kansas, Lawrence, Kansas, USA.
Coût: Gratuit
Auteur: Université du Kansas 2003a
Téléchargement: http://www.specifysoftware.org/Specify/specify/download.

XPER2 [13]

Description: Logiciel assistant l'identification, la description d'observations morpho-anatomiques et l'attribution de noms scientifiques, permet l'échange de clés de déterminations taxonomiques interactives.
Version: 2.1 (2010).
Détenteur: Université Pierre et Marie Curie (Paris 6, France)
Coût: Gratuit
Auteur: Université Pierre et Marie Curie 2010
Téléchargement: http://www.specifysoftware.org/Specify/specify/download.

[13] Logiciel mentionné par le traducteur (FT 2010)

2. Ressources en ligne

BioGeoMancer

Description: Un service de géoréférencement pour les collecteurs, nettoyeurs et utilisateurs de spécimens d'histoire naturelle.
Détenteur: Museum d'Histoire Naturelle Peabody, Connecticut, USA.
Auteur: Peabody Museum *n.dat.*
Lieu: http://www.biogeomancer.org
Notes: Le projet BioGeomancer a récemment (2005) été étendu pour devenir une collaboration internationale entre des musées d'histoire naturelle avec pour but l'amélioration des outils de géoréférencement et de vérification de qualité des données. Ces outils sont disponibles depuis mi 2006 sous forme de programmes autonomes et de services en ligne.

Data Cleaning (CRIA)

Description: Un outil de vérification de données et d'identification d'erreurs développé par le CRIA pour aider les assainisseurs de données disponibles via la système d'information distribuées de speciesLink pour identifier les possibles erreurs dans leur bases de données. Ces erreurs incluent les erreurs de nomenclature et géographiques.
Détenteur: Centro de Referência em Informação Ambiental (CRIA), Campinas, Brésil.
Lieu: http://splink.cria.org.br/dc
Notes: Certains des algorithmes développés dans cet outils (particulièrement les outils géographiques) sont en voie d'être incorporés dans le kit Biogeomancer dans le cadre du projet de collaboration internationale censé se terminer mi 2006.

geoLoc

Description: Un outil d'assistance pour le géoréférencement des données des collections biologiques.
Détenteur: Centro de Referência em Informação Ambiental (CRIA), Campinas, Brésil.
Lieu: http://splink.cria.org.br/geoloc?&setlang=en

Georeferencing Calculator

Description: Une applette java pour aider dans le géoréférencement des localités descriptives telles que celles trouvées dans les collections de musées d'histoire naturelle.
Détenteur: Université de Californie, Berkeley, CA, USA.
Lieu: http://manisnet.org/manis/gc.html

Lifemapper

Description: Logiciel utilisant l'Internet pour récolter les enregistrements de plantes et d'animaux à partir des musées d'histoire naturelle et qui utilise un algorithme de modélisation pour prédire la répartition.
Détenteur: Biodiversity Research Center, Université du Kansas, Lawrence, Kansas, USA.
Lieu: http://www.lifemapper.org/

spOutlier

Description: Un outil automatisé pour détecter les valeurs aberrantes dans les latitudes, longitudes et altitudes et pour identifier les enregistrements « errants » à l'intérieur ou à

l'extérieur des côtes dans les données d'histoire naturelle.

Détenteur: Centro de Referência em Informação Ambiental (CRIA), Campinas, Brésil.

Lieu: http://splink.cria.org.br/outlier?&setlang=en

3. Normes et Directives

DELTA

Description: Le format DELTA (Description du Langage pour la Taxonomie, *DEscription Language for Taxonomy*)est une méthode flexible d'encodage de descriptions taxonomiques par traitement informatique.

Norme: Adoptée par TDWG comme une norme pour l'échange de données.

Auteur: Dallwitz *et al.* 1993.

Lieu: http://delta-intkey.com/

HISPID

Description: Normes d'information d'Herbiers et Protocoles pour l'Interchangement des données.

Détenteur: Commité des Responsables des Herbiers Australiens, *Committee of Heads of Australian Herbaria*. Adoptée comme une norme TDWG.

Auteur: Conn 1996, 2000.

Lieu: http://plantnet.rbgsyd.nsw.gov.au/Hispid4/

MaNIS Georeferencing Guidelines

Description: Contient des informations sur l'attribution de coordonnées géographiques et les distances d'erreurs maximums pour ces coordonnées.

Détenteur: Université de Californie, Berkeley, CA, USA.

Lieu: http://manisnet.org/GeorefGuide.html

Manual de Procedimentos para Georreferenciar

Description: Manuel développé par CONABIO au Mexique comme directives pour le géoréférencement des collections d'histoire naturelle. En espagnol avec un résumé anglais en préparation.

Auteur: CONABIO 2005.

Lieu: pas de version électronique disponible pour le moment.

MaPSTeDI Georeferencing Guidelines

Description: Guide au processus de géoréférencement de spécimens du projet MaPSTeDI.

Détenteur: Université du Colorado Regents, Denver, CO, USA.

Lieu: http://mapstedi.colorado.edu/geocoding.html

Plant Names in Botanical Databases

Description: Le but de ces normes est de spécifier comment les noms de plantes scientifiques pourraient être organisés dans les bases de données botaniques.

Détenteur: Groupe de Travail de Bases de Données Taxonomiques, *Taxonomic Databases Working Group* (TDWG).

Lieu: http://www.tdwg.org/plants.html

SDD

Description: Le sous-groupe SDD du TDWG a été établi pour développer un standard XML

international pour capturer et gérer les données descriptives d'organisations.

Détenteur: Taxonomic Databases Working Group (TDWG)..

Lieu: http://www.tdwg.org/standards/116/

TDWG Standards

Description: Le TDWG a développé des normes à utiliser pour les données de biodiversité depuis de nombreuses années. Ces normes ont été et sont toujours développées pour une variété de problèmes liés au stockage, à la documentation et à la répartition des données relatives aux espèces ainsi qu'aux occurences d'espèces.

Détenteur: Taxonomic Databases Working Group (TDWG).

Lieu: http://www.tdwg.org/standards/
http://www.tdwg.org/activities/

Conclusion

Errores ad sua principia referre, est refellere
Rapporter des erreurs à leur origine revient à les réfuter

(Ref. 3 Co. Inst. 15)

L'âge de l'information a signifié que les institutions de collections naturelles sont devenus part intégrante du processus de prise de décision environnementale et les acteurs politiques insistent de plus en plus sur la pertinence et la valeur en retour des moyens qu'ils placent dans ces institutions. Il est donc dans le meilleur intérêt des institutions de collections naturelles qu'ils produisent un produit de qualité se elles entendent rester perçues comme ayant une forte valeur ajoutée auprès des bailleurs de fonds.

Les meilleures pratiques pour la mise en bases de données des informations de musées, d'herbiers et d'institutions conservant des données observationnelles et de terrain impliquent de rendre les données les plus précises possibles et l'utilisation des techniques et méthodologies les plus appropriées pour assurer que les données sont aussi bonnes qu'elles peuvent l'être. Pour s'assurer que c'est bien le cas, il est essentiel que les erreurs d'entrées de données soient réduites au minimum, et que les nettoyages de données et validations en cours d'exécution soient intégrés dans des protocoles quotidiens de gestion des informations et des données.

Il n'y a pas de bonne ou de mauvaise qualité de données (Chapman 2005a). Les données sont des données, et leur utilisation détermine leur qualité. Cependant, les fournisseurs de données se doivent d'assurer autant que possible que les données ne contiennent aucune erreur. A lui seul aucun test ne sera jamais suffisant pour identifier toutes les erreurs d'un jeu de données, et il est donc important d'utiliser un ensemble de méthodes répondant au mieux aux exigences de l'organisme qui les utilise et aux données qu'elle détient. De plus, la collaboration entre les institutions, les fournisseurs de données, les scientifiques et les informaticiens de même qu'entre les utilisateurs de données est nécessaire pour améliorer la qualité des données, pas uniquement au sein des institutions de collection individuelles mais également au travers de la totalité des collections où cette collaboration entre en jeu.

Une bonne documentation est sans doute la technique la plus importante de la gestion des données. Quels que soient les tests qui ont été menés sur les données, ils devraient être documentés de manière complète. Ce n'est que de cette façon que les utilisateurs des données peuvent vraiment être informés sur leurs natures et leur exactitude probable.

En cette période d'augmentation des données et d'échange d'information, la notoriété des collections et des institutions risque de dépendre de la qualité et de la disponibilité des informations qu'elle détient (Redman 1996, Dalcin 2004), plus que de la qualité des scientifiques comme c'était le cas par le passé. C'est un fait, et les deux ne peuvent plus être séparés. Une bonne gestion de données et d'informations va de pair avec une bonne activité scientifique et ensemble elles devraient permettre de bonnes données et de bonnes informations.

Remerciements

Texte

Références bibliographiques

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

ABRS. 2004. *Australian Faunal Directory*. Canberra: Australian Biological Resources Study. http://www.deh.gov.au/biodiversity/abrs/online-resources/abif/fauna/afd/index.html [Accessed 12 Apr. 2005].

ANBG. 2003. *Australian Plant Name Index*. Canberra: Australian National Botanic Gardens. http://www.anbg.gov.au/apni/index.html [Accessed 12 Apr. 2005].

Barnett, V. and Lewis, T. 1994. *Outliers in Statistical Data*. Chichester, UK: Wiley and Sons.

Beaman, R.S. 2002. Automated georeferencing web services for natural history collections **in** *Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002* http://www.cria.org.br/eventos/tdbi/flora/reed [Accessed 12 Apr. 2005]

Beaman, R., Wieczorek, J. and Blum, S. 2004. Determining Space from Place for Natural History Collections in a Distributed Library Environment. *D-Lib Magazine* Vol. 10(5). http://www.dlib.org/dlib/may04/beaman/05beaman.html [Accessed 12 Apr. 2005].

Belbin, L. 2004. *PATN vers. 3.01*. Blatant Fabrications http://www.patn.com.au [Accessed 12 Apr. 2005].

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 13 Apr. 2005].

BioCASE. 2003. *Biological Collection Access Service for Europe*. http://www.biocase.org [Accessed 12 Apr. 2005].

Birds Australia. 2004. *Birds Australia Rarities Committee (BARC).* http://users.bigpond.net.au/palliser/barc/barc-home.html [Accessed 12 Apr. 2005].

Bisby, F.A. 1994. *Plant Names in Botanical databases.* TDWG Standard. http://www.tdwg.org/plants.html [Accessed 12 Apr. 2005].

Bisby, F.A., Zarucchi, J.L., Schrire, B.L., Roskov, Y.R., Heald, J. and White, R.J. (eds). 2002. *ILDIS World Database of Legumes* ver. 6.05. http://www.ildis.org/ [Accessed 12 Apr. 2005].

Blakers, M., Davies, S.J.J.F. and Reilly, P.N. 1984. *The Atlas of Australian Birds*. Melbourne: Melbourne University Press.

Blum, S. 2001. *Georeferencing Natural History Collection Localities at the California Academy of Sciences*. http://www.calacademy.org/research/informatics/GeoRef/index.html [Accessed 12 Apr. 2005].

Brickell, C.D., Baum, B.R., Hetterscheid, W.L.A., Leslie, A.C., McNeill, J., Trehane, P., Vrugtman, F. and Wiersema, J.H. (eds) 2004. *International Code for Cultivated Plants* ed. 7. Edinburgh, U.K.: ISHS. http://www.actahort.org/books/647/ [Accessed 11 Apr. 2005].

Brummitt, R.K. and Powell, C.E. (eds). 1992 *Authors of Plant Names.* Kew: Royal Botanic Gardens, Kew. http://www.ipni.org/index.html [Accessed 12 Apr. 2005]

Burbidge, A.A. 1991. Cost Constraints on Surveys for Nature Conservation **in** Margules, C.R. and Austin, M.P. (eds). *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*. Canberra: CSIRO.

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

Busby, J.R. 1991. BIOCLIM – a bioclimatic analysis and prediction system. Pp. 4-68 **in** Margules, C.R. and Austin, M.P. (eds) *Nature Conservation: Cost Effective Biological Surveys and data Analysis*. Melbourne: CSIRO.

Chapman, A.D. 1988. Karl Domin in Australia **in** *Botanical History Symposium. Development of Systematic Botany*

in Australasia. Ormond College, University of Melbourne. May 25-27, 1988. Melbourne: Australian Systematic Botany Society, Inc.

Chapman, A.D. 1991. Australian Plant Name Index pp. 1-3053. *Australian Flora and Fauna Series* Nos 12-15. Canberra: AGPS.

Chapman, A.D. 1992. Quality Control and Validation of Environmental Resource Data **in** *Data Quality and Standards: Proceedings of a Seminar Organised by the Commonwealth Land Information Forum, Canberra, 5 December 1991*. Canberra: Commonwealth land Information Forum.

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. *et al*. 2002. *Guidelines on Biological Nomenclature*. Canberra: Environment Australia. http://www.deh.gov.au/erin/documentation/nomenclature.html [Accessed 12 Apr. 2005].

Chapman, A.D. 2004. Guidelines on Biological Nomenclature. Brazil edition. Appendix J 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 11 pp. http://splink.cria.org.br/docs/appendix_j.pdf [Accessed 12 Apr. 2005].

Chapman, A.D. 2005a. *Principles of Data Quality*. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/data_quality [Accessed 1 Aug. 2005].

Chapman, A.D. 2005b. *Uses of Primary Species-Occurrence Data*. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/uses_of_data [Accessed 1 Aug. 2005].

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. and Milne, D.J. 1998. *The Impact of Global Warming on the Distribution of Selected Australian Plant and Animal Species in relation to Soils and Vegetation*. Canberra: Environment Australia

Chapman, A.D., Bennett, S., Bossard, K., Rosling, T., Tranter, J. and Kaye, P. 2001. Environment Protection and Biodiversity Conservation Act, 1999 – Information System. *Proceedings of the 17th Annual Meeting of the Taxonomic Databases Working Group, Sydney, Australia 9-11 November 2001*. Powerpoint: http://www.tdwg.org/2001meet/ArthurChapman_files/frame.htm [Accessed 12 Apr. 2005].

CHAH 2002. *AVH - Australian's Virtual Herbarium*. Australia: Council of Heads of Australian Herbaria. http://www.chah.gov.au/avh/avh.html [Accessed 12 Apr. 2005].

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.

Christidis, L. & Boles, W.E. 1994. *Taxonomy and Species of Birds of Australia and its Territories*. Royal Australasian Ornithologists Union, Melbourne. 112 pp.

Clarke, K.C. 2002. *Getting Started with Geographic Information Systems*, 4th edn. Upper Saddle River, NJ, USA: Prentice Hall. 352 pp.

Colwell, R.K. 2002. *Biota: The Biodiversity Database Manager*. Connecticut, USA: University of Connecticut http://viceroy.eeb.uconn.edu/Biota [Accessed 12 Apr. 2005].

CONABIO. 2002. *The Biótica Information system*. Mexico City: Comisión national para el conocimiento y uso de la biodiversidad. http://www.conabio.gob.mx/informacion/biotica_ingles/doctos/acerca_biotica.html [Accessed 12 Apr. 2005

CONABIO. 2005. *Manual de Procedimentos para Georreferenciar.* Mexico: Comisión para el Conocimiento y Uso de la Biodiversidad México (CONABIO).

Conn, B.J. (ed.) 1996. *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data.* Version 3 (Draft 1.4). Sydney: Royal Botanic Gardens.

http://www.bgbm.org/TDWG/acc/hispid30draft.doc [Accessed 10 Apr 2005]

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 Jul. 2003].

Croft, J.R. (ed.) 1989. *HISPID – Herbarium Information Standards and Protocols for Interchange of Data*. Canberra: Australian National Botanic Gardens.

CRIA. 2002. *speciesLink*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/ [accessed 12 Apr. 2005]

CRIA. 2004a. *GeoLoc-CRIA*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/tools/ [Accessed 12 Apr. 2005].

CRIA. 2004b. *spOutlier-CRIA*. Centro de Referência em Informação Ambiental. http://splink.cria.org.br/tools/ [accessed 1 Mar. 2005]

CRIA (2005), *speciesLink. Dados e ferramentos. Data Cleaning*. Centro de Referência em Informação Ambiental. http://splink.cria.org.br/dc/ [Accessed 12 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. 2005].

Dallwitz, M.J., Paine, T.A. and Zurcher, E.J. (1993). *User's guide to the DELTA System: a general system for processing taxonomic descriptions*. 4th edn. http://delta-intkey.com/ [Accessed 12 Apr. 2005].

DEH. 2005a. *Threatened Species*. Canberra: Department of Environment and Heritage. http://www.deh.gov.au/biodiversity/threatened/species/index.html [Accessed 12 Apr. 2005].

DEH. 2005b. *Species Profile and Threats Database*. Canberra : Department of Environment and Heritage. http://www.deh.gov.au/cgi-bin/sprat/public/sprat.pl [Accessed 7 Apr. 2005].

Dorr, LJ. 1997. *Plant Collectors in Madagascar and the Comoro Islands*. Kew, UK: Royal Botanic Gardens, Kew.

English, L.P. 1999. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc., New York.

ESRI. 2003. *ArcSDE: The GIS Gateway to Relational Databases*. http://www.esri.com/software/arcgis/arcinfo/arcsde/overview.html [Accessed 12 Apr. 2005

Farr, E. and Zijlstra, G. (eds). *n.dat*. *Index Nominum Genericorum (Plantarum).* On-line version. http://ravenel.si.edu/botany/ing/ [Accessed 21 Jul. 2004].

Flowerdew, R., 1991. Spatial Data Integration. pp. 375-387 **in**: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

Froese, R. and Bisby, F.A. (eds). 2004. *Catalogue of Life 2004*. Los Baños, Philippines: Species 2000. http://www.sp2000.org/AnnualChecklist.html [Accessed 7 Apr. 2005].

Froese, R. and Pauly, D. 2004. *Fishbase.* Ver. 05/2004. The Philippines: World Fish Center. http://www.fishbase.org/ [Accessed 10 Apr. 2005].

Fundación Biodiversidad 2005. *Proyeto Anthos – Sistema de información sobre los plantas de España*. http://www.programanthos.org/ [Accessed 8 Apr. 2005].

Gatrell, A.C. 1991. Concepts of Space and Geographical Data. pp. 119-134 **in:** Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

GBIF. 2003a. *What is GBIF?* http://www.gbif.org/GBIF_org/what_is_gbif [Accessed 12 Apr. 2005].

GBIF. 2003b. *GBIF Work Program 2004*. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/GBIF_org/wp/wp2004/GB7_20WP2004-v1.0-approved.pdf [Accessed 12 Apr. 2005].

GBIF. 2004. *Data Portal*. Copenhagen: Global Biodiversity Information Facility.

http://www.gbif.net/portal/index.jsp. [Accessed 12 Apr. 2005].

Geographic Names Board. 2003. *Guidelines for Naming of Roads*. Sydney: Geographic Names Board of New South Wales. http://www.gnb.nsw.gov.au/newsroom/road_naming_guideline.pdf [Accessed 21 Jul. 2004].

Greuter, W. *et al.* 1984-1989. *Med-Checklist: a critical inventory of vascular plants of the circum-mediterranean countries*. 4 Vols. Botanical Garden and Botanical Museum Berlin-Dahlem.

Hepper, F.N. and Neate, F. 1971. *Plant Collectors in West Africa.* Utrecht, The Netherlands: Oosthoeks' Uitgeversmaatschappij.

Hijmans, R.J., Guarino, L., Bussink, C., Mathur, P., Cruz, M., Barrentes, I. and Rojas, E. 2005 *DIVA-GIS Version 5. A geographic information system for the analysis of biodiversity data*. http://www.diva-gis.org [Accessed 30 Jul. 2004].

Hobern, D. and Saarenmaa, H. 2005. *GBIF Data Portal Strategy.* Draft Version 0.14. Copenhagen, Denmark: Global Biodiversity Information Facility. http://circa.gbif.net/Public/irc/gbif/dadi/library?l=/architecture/portal_strategy_1/ [Accessed 7 Apr. 2005].

Hood, G.M. 2005. *PopTools version 2.6.6*. Canberra: CSIRO Sustainable Ecosystems. http://www.cse.csiro.au/poptools [Accessed 13 Apr. 2005].

Houlder, D. Hutchinson, M.J., Nix, H.A. and McMahaon, J. 2000. *ANUCLIM 5.1 Users Guide*. Canberra: Cres, ANU. http://cres.anu.edu.au/outputs/anuclim.php [Accessed 12 Apr. 2005].

IAPT. 1997. *Names in Current Use for Extant Plant Genera ver. 1.0*. on-line version. International Association for Plant Taxonomy. http://www.bgbm.org/iapt/ncu/genera/Default.htm [Accessed 12 Apr. 2005].

ICSM. 2001. *Guidelines for the Consistent Use of Place Names*. Intergovernmental Committee on Survey and Mapping: Committee for Geographic Names in Australia. http://www.icsm.gov.au/icsm/cgna/consistent_pnames.pdf. [Accessed 12 Apr. 2005].

Index Herbariorum. (1954-1988) *Index Herbariorum Part 2: Collectors*. Various compilers. Utrecht/Antwerp, The Hague/Boston

Part 2(1): Collectors A-D (1954). *Regnum Vegetabile* vol. 2 (A-D),

Part 2(2): Collectors E-H (1957). *Regnum Vegetabile* vol. 9 (E-H),

Part 2(3): Collectors I-L (1972). *Regnum Vegetabile* vol. 86 (I-L),

Part 2(4): Collectors M (1976). *Regnum Vegetabile* vol. 93 (M),

Part 2(5): Collectors N-R (1983). *Regnum Vegetabile* vol. 189 (N-R),

Part 2(6): Collectors S (1986). *Regnum Vegetabile* vol. 114 (S),

Part 2(7): Collectors T-Z (1988). *Regnum Vegetabile* vol. 117 (T-Z).

IPNI. 1999. *International Plant Names Index*. http://www.ipni.org/index.html [accessed 12 Apr. 2005].

IOPI. 2003. *Global Plant Checklist*. International Organization for Plant Information (IOPI). http://www.bgbm.fu-berlin.de/IOPI/GPC/ [Accessed 12 Apr. 2005].

Johnson, R.A. and Wichern, D.W. 1998. *Applied Multivariate Statistical Analysis*. 4th edn. New York, NY: Prentice Hall.

Jones P.G. and Gladkov, A. 2001. *Floramap Version 1.01*. Cali, Colombia: CIAT. http://www.floramap-ciat.org/ing/floramap101.htm [Accessed 12 Apr. 2005].

Koch, I. 2003. *Coletores de plantas brasileiras*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 12 Apr. 2005].

Lampe, K.-H. and Riede, K. 2002. *Mapping the collectors: the georeferencing bottleneck*. Poster given to TDWG meeting, Indaiatuba, Brazil.

http://www.cria.org.br/eventos/tdbi/bis/Poster-200dpi.html [accessed 12 Apr. 2005].

Legendre P. and Legendre L. (1998): Numerical Ecology. *Developments in Environmental Modeling* 20, Second English Edition, Elsevier, Amsterdam, 853p. http://www.bio.umontreal.ca/legendre/numecol.html [Accesssed 12 Apr. 2005].

Lindemeyer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F. and Tanton, M.T. 1991. The Conservation of Leadbeater's Possum, *Gymnobelidus leadbeateri* (McCoy): A Case Study of the Use of Bioclimatic Modelling. *J. Biogeog.* **18:** 371-383.

Lowry, R. 2005. Concepts and Applications of Inferential Statistics. *VassarStats: Web Site for Statistical Computation*. http://faculty.vassar.edu/lowry/webtext.html [Accessed 8 Apr. 2005]

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 12 Apr. 2005].

MaNIS. 2001. *The Mammal Networked Information System*. http://manisnet.org/manis [Accessed 12 Apr. 2005].

Marcus, A., Maletic, J.I. and Lin, K.-I. 2001. Ordinal Association Rules for Error Identification in Data Sets pp. 589-591 in *Proceedings of the 10th ACM Conference on Information and Knowledge Management (ACM CIKM 2001). Atlanta, GA*. http://www.cs.wayne.edu/~amarcus/papers/cikm01.pdf [Accessed 12 Apr. 2005].

Margules, C.R. and Redhead, T.D. 1995. *BioRap. Guidelines for using the BioRap Methodology and Tools*. Canberra: CSIRO. 70pp.

Marino, A., Pavarin, F., de Souza, S. and Chapman, A.D. in prep. *Simple on line tools for geocoding and validating biological data*. To be submitted.

Neldner, V.J., Crossley, D.C. and Cofinas, M. 1995. Using Geographic Information Systems (GIS) to Determine the Adequacy of Sampling in Vegetation Surveys. *Biological Conservation* 73: 1-17.

Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes in Longmore, R.C. (ed). Atlas of Australian elapid snakes. *Australian Flora and Fauna Series* No. **7:** 4-15. Canberra: Australian Government Publishing Service.

NMNH. 1993. *RapidMap. Geocoding locality descriptions associated with herbarium specimens*. U.S. National Musuem of Natural History and Bernice P. Bishop Museum, Honolulu. http://users.ca.astound.net/specht/rm/ [Accessed 12 Apr. 2005].

Peabody Museum. *n.dat*. *BioGeoMancer*. http://www.biogeomancer.org [Accessed 12 Apr. 2005].

Peterson, A.T., Navarro-Siguenza, A.G. and Benitez-Diaz, H. 1998. The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis* **140:** 288-294.

Peterson, A.T., Stockwell, D.R.B. and Kluza, D.A. 2002. Distributional Prediction Based on Ecological Niche Modelling of Primary Occurrence Data pp. 617-623 **in** Scott, M.J. *et al.* eds. *Predicting Species Occurrences. Issues of Accuracy and Scale.* Washington: Island Press.

Peterson, A.T., Navarro-Siguenza, A.G. and Pereira, R.S. 2003. Detecting errors in biodiversity data based on collectors' itineraries. *Bulletin of the British Ornithologists' Club* 124: 143-151. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/PNP_BBOC_2004.pdf. [Accessed 12 Apr. 2005].

Pfeiffer, U., Poersch, T. and Fuhr, N. 1996. Retrieval effectiveness of proper name search methods. *Information Processing and Management* **32(6):** 667-679.

Platnick, N.I. 2004. *The World Spider Catalog*. New York: The American Museum of Natural History. http://research.amnh.org/entomology/spiders/catalog81-87/INTRO3.html [Accessed 12 Apr. 2005].

Podolsky, R. 1996. *Software Tools for the Management and Visualization of Biodiversity Data*. NY, USA: United Nations Development Project. http://www3.undp.org/biod/bio.html [Accessed 13 Apr. 2005].

Pollock, J.J. and Zamora, A. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of ACM* **27(4):** 358-368.

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.

Rios, N.E. and Bart, H.L. Jr. *n.dat. GEOLocate. Georeferencing Software. User's Manual*. Belle Chasse, LA, USA: Tulane Museum of Natural History. http://www.museum.tulane.edu/geolocate/support/manual_ver2_0.pdf [Accessed 12 Apr. 2005].

Roughton, K.G. and Tyckoson, D.A. 1985. Brousing with sound: Sound-based codes and automated authority control. *Information Technology and Libraries* **4(2):**130-136.

Ruggiero, M. (ed.) 2001. *Integrated Taxonomic Information System*. http://www.itis.usda.gov/ [Accessed

12 Apr. 2005].

Pereira, R.S. 2002. *Desktop Garp*. Lawrence, Kansas: University of Kansas Center for Research. http://www.lifemapper.org/desktopgarp/Default.asp?Item=2&Lang=1 [Accessed 13 Apr. 2005].

Shattuck, S.O. 1997. eGaz, The Electronic Gazetteer. *ANIC News* **11:** 9 http://www.ento.csiro.au/biolink/egaz.html [Accessed 12 Apr. 2005].

Shattuck, S.O. and Fitzsimmons, N. 2000. *BioLink, The Biodiversity Information Management System*. Melbourne, Australia: CSIRO Publishing. http://www.ento.csiro.au/biolink/software.html [Accessed 12 Apr. 2005].

Steenis-Kruseman, M.J. van 1950. Malaysian Plant Collectors and Collections. *Flora Malesiana* Vol. 1. Leiden, The Netherlands.

Stockwell, D. and Peters, D. 1999. "The GARP modelling system: problems and solutions to automated spatial prediction." *International Journal of Geographical Information Science* **13**(2): 143-158.

University of Colorado Regents. 2003a. *mapstedi. Geocoding*. Denver: University of Colorado MaPSTeDI project. http://mapstedi.colorado.edu/geocoding.html [Accessed 12 Apr. 2005].

University of Colorado Regents. 2003b. *GeoMuse*. Denver: University of Colorado MaPSTeDI project. http://www.geomuse.org/mapstedi/client/start.jsp [Accessed 12 Apr. 2005].

University of Kansas. 2003a. *Specify.* Biological Collections Management. Lawrence, Kansas: University of Kansas http://www.specifysoftware.org/Specify/ [Accessed 12 Apr. 2005].

University of Kansas. 2003b. *LifeMapper.* Lawrence, Kansas: University of Kansas – Informatics Biodiversity Research Center. http://www.lifemapper.org/ [Accessed 12 Apr. 2005

University of Oxford. 2004. *BRAHMS. Botanical Research and Herbarium Management System*. Oxford, UK: University of Oxford http://storage.plants.ox.ac.uk/brahms/defaultNS.html [Accessed 27 Jul 2004].

Weber, W.A. 1995. Vernacular Names: Why Oh Why?. *Botanical Electrical News* No. 109. http://www.ou.edu/cas/botany-micro/ben/ben109.html [Accessed 7 Apr. 2005].

Weiher, E. and Keddy, P. (eds). 1999. *Ecological Assembly Rules: Perspectives, Advances, Retreats.* Cambridge, UK: Cambridge University Press. 418 pp.

Wieczorek, J. 2001a. *MaNIS: Georeferencing Guidelines*. Berkeley: University of California, Berkeley - MaNIS http://manisnet.org/manis/GeorefGuide.html [Accessed 12 Apr. 2005].

Wieczorek, J. 2001b. *MaNIS: Georeferencing Calculator*. Berkeley: University of California, Berkeley - MaNIS http://manisnet.org/manis/gc.html [Accessed 12 Apr. 2005].

Wieczorek, J. and Beaman, R.S. 2002. Georeferencing: Collaboration and Automation in *Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002* http://www.cria.org.br/eventos/tdbi/bis/georeferencing [Accessed 12 Apr. 2005].

Wieczorek, J., Guo, Q. and Hijmans, R.J. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18(8): 745-767.

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

Williams, P.H., Marguiles, C.R. and Hilbert, D.W. 2002. Data requirements and data. sources for biodiversity priority area selection. *J. Biosc.* **27(4):** 327-338.

Index

ANUCLIM 56,62,68
BioCase 25
BIOCLIM 55
BioGeoMancer 4,12,36,70
BioLink 12,42,68
Biota 12
Biótica 12
BRAHMS 12,69
calculateur de géoréférencement 4,71
certitude d'identification 13
conception de la base de données 13
conception des bases de données 16
DELTA 64,72
détection des erreurs 3
Diva-GIS 43,44,45,46,51,55,56,59,61,67,69
données de taxonomie et de nomenclature 13
données descriptives 64,72
eGaz 4,12,41,42,69
erreurs de nomenclature et de taxonomie 5
erreurs spatiales 5
exactitude spatiale 32
FloraMap 57,58
fusionner les bases de données 6
GARP 63,69
géocode 31,32,33
geoLoc 4,12,38,71
GEOLocate 4,12,40,69
géoréférencement 32
géoréférencement
  directives de MaNIS 72
  directives de MaPSTeDI 72
herbier virtuel australien 25,34
HISPID 12,14,28
Lifemapper 63
nettoyage de données 2
nettoyage de données avec les outils du CRIA 71
nettoyage des données 6,11
nettoyage des noms 55
noms de plantes dans les bases de données botaniques 72
noms scientifiques 15
occurrences doublées 65
orthographe des noms 15
PATN 57,58,63,64,70
Phonix 19,22
portail GBIF 25,34

prévenir les erreurs....4
RapidMap Geocoder....38
ressources en ligne....70
saisie des données....15
SDD....64,72
Skeleton-Key....19,22
Soundex....18,22
speciesLink....25,28,34,38
Specify....12,70
spOutlier....55
systèmes d'information géographique....50
TDWG....64,73
valeurs aberrantes....71
valeurs de données contradictoires....65
valeurs de données incohérentes....19
valeurs de données incorrectes....18,64
valeurs de données manquantes....18
valeurs de données non-atomiques....19,65
vérification des erreurs....15,18,46
visualisation pour la présentation....66