PRINCÍPIOS DE QUALIDADE DE DADOS

Arthur D. Chapman[1]

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

GBIF
www.gbif.org

[1] Australian Biodiversity Information Services
PO Box 7491, Toowoomba South, Qld, Australia
email: info@gbif.org

Citação sugerida:
Chapman, A. D. (2015). *Princípios de Qualidade de Dados*. Versão 1.0 pt em Português lançada em abril 2015 e traduzida para pelo Nó Português do GBIF (www.gbif.pt) e pelo representante brasileiro do GBIF, SiBBr (Sistema de Informação sobre a Biodiversidade Brasileira, www.sibbr.gov.br). Versão original em Inglês lançada em jul 2005. Copenhagen: Global Biodiversity Information Facility. 81 pp. ISBN: 87-92020-58-5. Disponível on-line em http://www.gbif.org/orc/?doc_id=5990.

ISBN/DoI: 87-92020-58-5 (10 figuras), 978-87-92020-58-1 (13 figuras).
EAN: 9788792020581.
URI persistente: http://www.gbif.org/orc/?doc_id=5990.

Língua: Portugesa.

Disclaimer:
Este trabalho foi encomendado para Arthur Chapman em 2004 pela Secretaria GBIF para destacar a importância da qualidade dos dados no que se refere aos dados de ocorrência primários. O texto foi revisto em 2015 para a precisão e expandida através de anexos a até b. Nosso entendimento dessas questões e as ferramentas disponíveis para facilitar a verificação de erros está evoluindo rapidamente. Portanto, esperamos que haverá futuras versões deste documento e gostaria de receber input via info@gbif.org.

Controle de versões do documento

Versão	Descrição	Data de lançamento	Autor(es)
1.0 pt	Primeira versão pública	abril 2015	Traduzido pelo GBIF.PT e SiBBr. Editado por AGT.

Crédito arte da capa: GBIF Secretariat, 2015. Imagem: *Amata phegea* (Linnaeus 1758) pelo Per de Place Bjørn, 2005.

ISBN 978-87-92020-58-1
90000 >
9 788792 020581
9 788792 020581 >

Conteúdo

Introduçao à versão em português

A questão da qualidade de dados e da sua ‘aptidão para o uso’ (fitness for use) é fundamental para a missão e a estratégia futura do GBIF. Trata-se de uma prioridade alta, amplamente reconhecida tanto pelos participantes da rede mundial como pelos usuários dos seus serviços de dados sobre a biodiversidade: é preciso uma melhoria significativa da consistência e da qualidade dos dados livremente disponíveis através da colaboração da comunidade do GBIF. Esta colaboração já conseguiu em possibilitar centenas de estudos utilizando dados mediados pelo GBIF, para aumentar o entendimento essencial sobre a vida da planeta e informar as politicas e as decisões para enfrentar o crise da perda de biodiversidade. Porém, a sustentabilidade do GBIF depende de um esforço constante para responder cada vez mais efetivamente aos requerimentos dos stakeholders na pesquisa, entre os tomadores de decisões e na sociedade geral. Neste sentido, os princípios de qualidade de dados articulados com tanta claridade por Artur Chapman em 2004 ficam tão relevantes hoje como eram quando este manual foi escrito. Permanece o documento mais consultado entre os recursos disponibilizados a GBIF.org/resources. Portanto, com o crescimento impressionante do envolvimento no compartilho e na publicação de dados sobre a biodiversidade entre a comunidade lusófona, é o momento certo para publicar esta edição do manual na língua português. Agradecemos sinceramente os esforços voluntários dos nossos colegas Inês Paulino do Nó Português do GBIF e Pedro Guimarães do representante brasileiro do GBIF, SiBBr (Sistema de Informação sobre a Biodiversidade Brasileira) na tradução aqui apresentada. Esperamos que seja uma ferramenta útil para auxiliar publicadores de dados nos países lusófonos em contribuir ainda mais para o conhecimento mundial sobre a biodiversidade.

Tim Hirsch
Vice-diretor
Secretariado do GBIF
Março de 2015

Introduçao

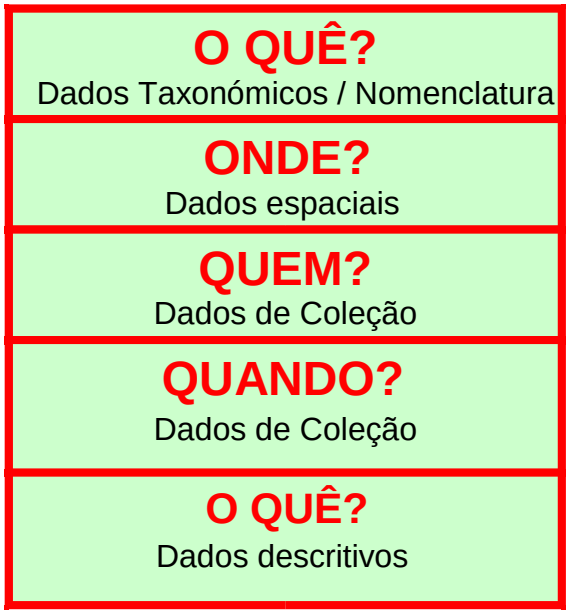

Fig. 1. Cinco perguntas destacando áreas relevantes para a qualidade dos dados de biodiversidade

Os Princípios de qualidade de dados tornaram-se o ponto central em diferentes atividades desde negócios (SEC 2002), medicina (Gad e Taulbee 1996), GIS (Zhang e Goodchild 2002) deteção remota[1] (Lunetta e Lyon 2004) e muitas outras, mas só agora se tornaram universalmente aceites por museus e pela comunidade taxonómica. O rápido aumento da disponibilização e troca de dados taxonómicos e de ocorrência de espécies tornou estes princípios importantes e de se ter em conta devido à crescente exigência dos utilizadores, que cada vez mais pedem informação com maior qualidade e detalhe. Muitas vezes os dados dos museus são vistos, pela comunidade externa, como inaceitáveis na tomada de decisões em termos de conservação do ambiente, mas será isto o resultado da qualidade dos dados, ou da sua documentação? Estes dados são de uma importância elevada. Devido à colheita ao longo do tempo, os dados são base insubstituível sobre a biodiversidade, no período de tempo em que os Humanos tiveram um grande impacto sobre esta (Chapman e Busby 1994). São um recurso essencial nos esforços para a conservação do ambiente, pois fornecem registos[2] completos de ocorrências de espécies em áreas que sofreram alterações de habitat devido à desflorestação, agricultura, urbanização, alterações climáticas ou que foram modificadas de outras formas (Chapman 1999).

Estas são algumas das ideias que tentarei explicar abaixo, mostrando uma série de princípios de qualidade de dados que devem ser tomados como fundamentais na atividade de museus e herbários à medida que partilharem os seus dados para a comunidade em geral.

[1] No Brasil, sensoriamento remoto.
[2] No Brasil, registro.

A qualidade de dados e erros nos dados são assuntos muitas vezes negligenciados em base de dados ambientais, sistemas de modelação[3], SIG, sistemas de suporte de decisão, etc. Muitas das vezes os dados são usados sem critério e sem ter em consideração os erros associados, o que levará a resultados errados, informação mal interpretada, decisões ambientais imprudentes e aumento de custos.

> "*Dados de espécimes de plantas e animais presentes em museus e herbários representam um recurso vasto, dando não só informação presente, a localização destas entidades, mas também informações históricas de há centenas de anos*" (Chapman e Busby 1994).

Existem muitos princípios de qualidade de dados que se aplicam quando se lida com dados de espécies, especialmente com os aspetos espaciais desses dados. Estes princípios estão envolvidos em todos as fases da gestão dos dados. A perda de qualidade em qualquer destas fases implica uma redução da sua aplicabilidade e uso. Estas fases são:

- Recolha[4] e registo de dados no momento da colheita,
- Manipulação de dados antes da digitalização (preparação de etiquetas, copia dos dados para um registo, etc.),
- Identificação da coleção (espécime, observação) e dos seus registos,
- Digitalização dos dados,
- Documentação dos dados (recolha e registo dos metadados),
- Armazenamento e arquivo de dados,
- Apresentação e disseminação de dados (publicação em papel e por via eletrónica, base de dados acessíveis através da Web, etc.),
- Utilização de dados (análise e manipulação).

Todas estas fases têm influência na qualidade final dos dados ou na sua "aptidão para o uso" e aplicam-se a todos os aspetos dos dados na parte taxonómica e nomenclatural dos dados, o "O Quê?", na parte espacial o "Onde?" e na outra informação, como no "Quem?" e no "Quando?" (Berendsohn 1997).

Antes de se discutir detalhadamente sobre qualidade de dados e a sua aplicação em dados de ocorrência de espécies, há que definir e descrever determinados conceitos. Estes são o próprio conceito de qualidade de dados, exatidão e a precisão, que muitas vezes se confundem, e que se entende por dados primários de espécie e dados de ocorrência de espécies.

"Não subestime a elegância simples da melhoria da qualidade. Não são necessárias aptidões especiais para além do trabalho em equipa5, formação6 e disciplina. Qualquer um pode ser um contribuidor efetivo" (Redman, 2001).

[3] No Brasil, modelagem.
[4] No Brasil, coleta.
[5] No Brasil, equipe.
[6] No Brasil, treinamento.

1. Definições

1.1. Dados de ocorrência de espécies

Os dados de ocorrência de espécies incluem, neste texto, os dados presentes na etiqueta de espécimes ou de lotes depositados em museus e herbários, dados de observação ou de estudos ambientais. Em geral, estes dados são o que designamos por dados pontuais através de linhas (dados referentes a transetos de estudos ambientais, colheitas ao longo de um rio), polígonos (observações através de uma área definida, como um parque natural) ou dados em grelha (observações ou registos de avaliações ao longo de uma grelha regular) estejam também incluídos. Em geral estamos a falar de dados georreferenciados, ou seja, registos com referências geográficas que os associam a um lugar em particular no espaço, com coordenadas (Latitude, Longitude, UTM) ou não (tendo descrições da localidade, altitude, profundidade) e de tempo (data, hora do dia). Em geral, os dados estão também associados a um nome taxonómico, mas colheitas não identificadas podem igualmente ser incluídas. A designação do termo "dados de ocorrência de espécies" é usada ocasionalmente como "dados primários de espécies".

1.2. Dados primários de espécies

"Dados primários de espécies" é um conceito usado para descrever os dados elementares da colheita ou dados sem qualquer tipo de atributos espaciais. Incluí os dados taxonómicos e nomenclaturiais sem atributos espaciais, tais como nomes, taxa e conceitos taxonómicos sem referências geográficas associadas.

1.3. Exatidão e Precisão

Exatidão e *Precisão* são conceitos regularmente confundidos, e as diferenças não são geralmente entendidas. As diferenças são melhor explicadas com um exemplo (Figura1).

Exatidão refere-se à aproximação dos valores medidos, observações ou estimativas ao valor real ou verdadeiro (ou ao valor que é aceite como verdadeiro – por exemplo, as coordenadas de um ponto de referência) como mostra a figura 1.

Precisão (ou resolução) pode ser divida em dois tipos principais. A p*recisão estatística*, que é a proximidade das observações repetidas entre si. Não tem a ver com a sua relação com o valor real, e podem ter alta precisão, mas baixa exatidão como mostra a figura 1a. A p*recisão numérica* é o número de dígitos significativos com os quais uma observação é registada[7], situação tornada bastante mais evidente com o uso de computadores. Por exemplo, uma base de dados pode produzir um registo decimal de latitude/longitude até à 10ª casa decimal (ou seja,.01 mm) enquanto que na realidade o registo tem uma resolução não maior do que 10-100 (3-4 casas decimais). Esta situação leva a uma falsa impressão quer da resolução, quer da exatidão.

Estes termos – exatidão e precisão- também podem ser aplicados a dados não-espaciais e a dados espaciais. Por exemplo, a coleção pode ter uma identificação até ao nível da subespécie (i.e. tem uma elevada precisão), mas pode ser o taxon errado (i.e. tem baixa

[7] No Brasil, registrada

exatidão), ou pode ser identificada apenas até ao nível da família (i.e. elevada exatidão, mas baixa precisão).

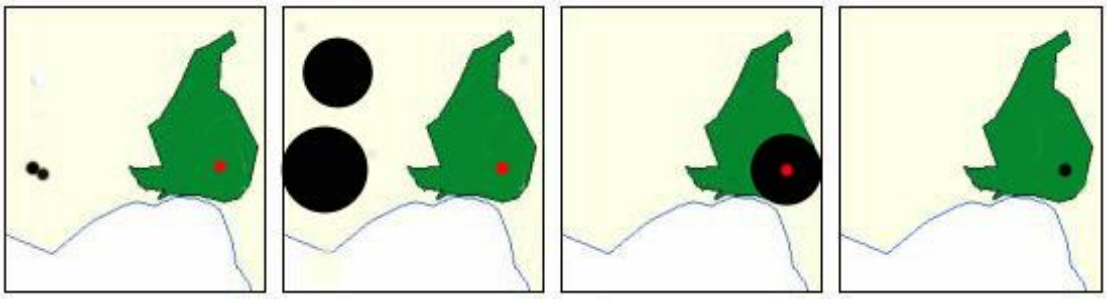

Fig. 2. *Mostra a diferença entre exatidão e precisão num contexto espacial. Os pontos vermelhos mostra a verdadeira localização, os pontos pretos, representam as localizações relatadas pelo coletor: (a) alta precisão, baixa exatidão, (b) baixa precisão, baixa exatidão mostrando erros aleatórios, (c) baixa precisão, elevada exatidão, e (d) elevada precisão e elevada exatidão.*

1.4. Qualidade

A qualidade quando aplicada aos dados tem várias definições, mas no mundo geográfico existe uma definição amplamente aceite a da "aptidão para o uso" (Chrisman 1983) ou "uso potencial". Esta é a definição adotada pela maioria dos padrões de transferência de dados espaciais modernos (ANZLIC 1996a, USGS 2004). É também utilizada na economia e no mundo dos negócios. Alguns autores (English 1999, por exemplo) acreditam que a definição "aptidão para o uso" é um pouco restritiva e discutem uma que envolva a aptidão dos dados para o uso potencial ou futuro.

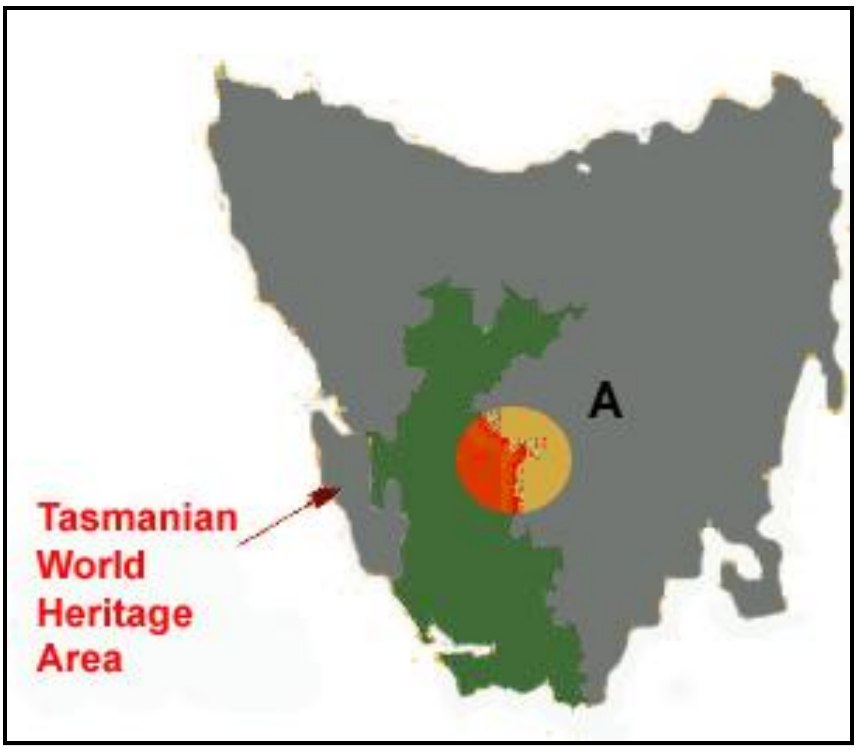

Fig. 3. *Mapa da Tasmânia, Austrália, que mostra o registo (A) recolhido com exatidão de 0.5º (cerca de 50 Km), como mostra o círculo. A área potencial da colheita (determinado usando o valor de exatidão) sobrepõe-se à Área de Património Mundial da Tasmânia (Tasmanian World Heritage Area).*

Um exemplo do uso do conceito de "aptidão para o uso" pode ser visto na figura 2. Uma colheita de uma espécie em particular (marcada por "A") tem de exatidão 0.5º de Latitude (cerca de 50 km). Se alguém estiver a preparar uma lista de espécies da Tasmânia e quiser saber onde esta espécie ocorre, o registo é capaz de responder à questão, ou seja, a coleção tem "aptidão para o uso" e pode ser considerada como tendo elevada qualidade para o propósito. Por outro lado, se alguém quiser saber se a espécie ocorre ou não na Área de Património Mundial da Tasmânia, com estas informações não é possível afirmar tendo em consideração apenas as informações do registo, pode ocorrer ou não. Estes dados não têm aptidão para este uso e têm baixa qualidade. Os valores de latitude e longitude numa base de dados podem ser muito precisos e aparentar terem elevada exatidão e isto pode ser mal interpretado pelo utilizador do registo se este não possuir um valor de exatidão associado.

Casos semelhantes ocorrem com componentes não espaciais dos dados quando um erro de identificação, por exemplo, pode tornar os dados pouco úteis e não aptos para o uso. Se alguém está a estudar a distribuição de uma espécie (ou a sua fisiologia ou ecologia, etc.) e tem associado um nome errado ao espécime ou observação pode levar a interpretações e resultados errados.

A qualidade de dados é multidimensional e envolve a gestão, modelação, análise, controlo de qualidade e seguro, armazenamento e apresentação dos dados. De forma independente, a qualidade de dados, como referido por Chrisman (1991) e Strong *et al.* (1997), está relacionada com o uso e não pode ser avaliada independentemente do utilizador. Numa

base de dados, os dados não têm verdadeira qualidade ou valor (Dalcin 2004); apenas têm valor *potencial* que só é realizado quando alguém os usa para fazer algo útil. A informação de qualidade está relacionada com sua capacidade para satisfazer os seus utilizadores e as suas necessidades (English 1999).

Redman (2001), sugeriu que para os dados estarem aptos a serem usados, estes devem estar e ser acessíveis, precisos, oportunos, completos, consistentes com outras fontes, relevantes, abrangentes, detalhados a um nível aceitável, serem fáceis de ler e de interpretar.

Um aspeto que um curador de dados tem de considerar é o que é necessário fazer para aumentar a usabilidade da sua base de dados e assim chegar a um público mais amplo (ou seja, aumentar o seu uso ou relevância potencial) e, assim, torná-los mais aptos para uma ampla gama de propósitos. Haverá um compromisso, entre o aumento da usabilidade e esforço necessário para adicionar essa funcionalidade e usabilidade extra. Isso, pode exigir a automatização dos campos de dados, acrescentando informações de georeferenciação[8], etc.

"Os dados são de alta qualidade se forem adequados para o uso para os quais foram produzidos, em operações, tomada de decisão e planeamento" (Juran, 1964).

1.5. Garantia de Qualidade/Controlo de qualidade

A diferença entre garantia de qualidade e controlo de qualidade nem sempre é clara. Taulbee (1996) fez a distinção entre os dois conceitos e apontou que não pode existir um sem o outro se os objetivos de qualidade forem para ser seguidos.

Esta autora define:

- *Controlo de qualidade* é uma avaliação da qualidade baseada em normas internas, processos e procedimentos estabelecidos para controlar e monitorizar a qualidade; e
- *Garantia de qualidade* é uma avaliação importante baseada em normas externas ao processo e é a revisão das atividades e processos de controlo de qualidade para assegurar que o produto final segue as normas de qualidade predeterminadas.

Numa abordagem orientada para a área dos negócios, Redman (2001) define *Garantia de Qualidade* como

> "*aquelas atividades que são projetadas para produzir produtos de informação livres de defeitos que atendem às necessidades mais importantes dos clientes, ao menor custo possível*".

Como aplicar estes conceitos na prática não é claro, e na maioria dos casos os conceitos parecem ser sinónimos[9] para descrever a prática geral na gestão da qualidade de dados.

1.6. Incerteza

A incerteza pode ser pensada como uma "medida de lacunas no conhecimento ou informação sobre uma quantidade desconhecida cujo valor real pode ser estabelecido se

[8] No Brasil, georreferenciamento.
[9] No Brasil, sinônimos.

existir um dispositivo de medição perfeito" (Cullen and Frey 1999). A incerteza é uma propriedade ligada à compreensão dos dados pelo observador, estando mais ligada a ele do que aos dados em si mesmos. Existe sempre incerteza nos dados; a dificuldade está em recolher, compreender e visualizar essa incerteza, para que os outros possam igualmente entendê-la. Incerteza é o conceito chave para a compreensão e avaliação de risco.

1.7. Erro

O erro engloba a imprecisão e inexatidão dos dados. Existem diversos fatores que contribuem para o erro.

> "*A opinião normal acerca dos erros e incertezas é de que são maus. Isto não é necessariamente verdadeiro, pois pode ser útil para se perceber como os erros e a incerteza aparecem, como é que podem ser geridos*[10] *e possivelmente reduzidos... Uma boa compreensão dos erros e da sua propagação levam a um controlo de qualidade ativo*" (Burrough e McDonnell 1998).

Os erros são geralmente vistos como aleatórios ou sistemáticos. *Erros aleatórios* referem-se a desvios da verdade de um modo aleatório. *Erros sistemáticos* ou enviesamentos surgem devido a um desvio uniforme dos valores e às vezes são descritos como tendo "exatidão relativa" no mundo cartográfico (Chrisman 1991). Ao determinar a "aptidão para uso", os erros sistemáticos podem ser aceitáveis para algumas aplicações e impróprio para outras. Um exemplo pode ser o uso de um datum[10] geodésico diferente - em que, se utilizados em toda a análise, podem não causar problemas de maior. Os problemas surgirão, no entanto, quando forem utilizados dados de diferentes fontes e com diferentes enviesamentos por exemplo, fontes de dados que usam diferentes datums geodésicos, ou onde identificações podem ter sido efetuadas utilizando uma versão anterior de um código de nomenclatura.

"*Como os erros são inevitáveis, devem ser reconhecidos como uma dimensão fundamental dos dados*" (Chrisman 1991). Só quando um erro está incluído na representação dos dados é possível responder a questões sobre as limitações dos dados, e mesmo sobre as limitações do conhecimento atual. Os erros conhecidos nas três dimensões do espaço, atributo e tempo precisam ser medidos, calculados, gravados e documentados.

1.8. Validação e Limpeza

A validação é um processo usado para determinar se os dados são inexatos, incompletos ou não razoáveis. O processo pode incluir controlo dos formatos, da integridade, de razoabilidade e de limite, revisão dos dados para identificar anómalos (geográficos, estatísticos, temporais ou ambientais), ou outros erros, e avaliação dos dados por especialistas na área (p.e. especialistas taxonómicos). Estes processos têm como resultado a sinalização, documentação e consequente controlo de registos suspeitos. O controlo da validação pode envolver também verificação de conformidade com os padrões aplicáveis, regras e convenções. Uma fase chave na validação e limpeza dos dados é identificar a

[10] Diferentes datums geográficos podem levar a trocas sistemáticas na posição atual (de coordenadas lat/long) até cerca de 400m em algumas zonas da Terra.

origem dos erros e focar-se na prevenção desses erros para que não voltem a ocorrer (Redman 2001).

A limpeza de dados (*Data cleaning*) refere-se ao processo de "reparar" os erros que foram identificados nos dados durante o processo de validação. Este conceito é sinónimo de "*data cleansing*", ainda que alguns utilizem este termo para abranger tanto a validação como a limpeza de dados. É importante no processo de limpeza que estes não sejam inadvertidamente perdidos, e as alterações na informação sejam realizadas com muito cuidado. Normalmente, é melhor manter ambas as versões, a antiga (dados originais) e a nova (dados corrigidos) lado a lado na base de dados de modo que, se os erros forem feitos no processo de limpeza, a informação original possa ser recuperada.

Foram produzidas uma série de ferramentas e diretrizes nos últimos anos para ajudar no processo de validação e limpeza de dados de espécies. Esta parte vai ser abordada no documento "*Principles and Methods of Data Cleaning*". O processo manual de limpeza de dados é trabalhoso, demorado e ele próprio está sujeito a erros (Maletic e Marcus 2000).

Os passos gerais para a limpeza de dados (conforme Maletic e Marcus 2000) é:

- Definir e determinar tipos de erros
- Procurar e identificar ocorrências de erros
- Corrigir erros
- Documentar as ocorrências dos erros e os diferentes tipos
- Modificar os procedimentos de entrada de dados para reduzir erros futuros.

1.9. Veracidade na Etiquetagem

A veracidade da etiquetagem é geralmente entendida como sendo a documentação da qualidade dos bens e dos produtos para venda ou tornados acessíveis a terceiros. Para dados de ocorrência de espécies, a veracidade na etiquetagem é geralmente composta por metadados, desde que estes documentem completamente os aspetos de qualidade, procedimentos e métodos de controlo de qualidade e/ou parâmetros estatísticos de qualidade relevantes para os dados. A veracidade na etiquetagem é a principal função que conduz à certificação e acreditação, nos casos em que esta é apropriada. A maioria dos museus e herbários já o fazem no que diz respeito à informação sobre especialistas e a data em que a identificação foi realizada (informações de determinação), mas isso raramente é estendido a outras informações no registo ou nos dados de observação e sem *voucher*.

1.10. Utilizadores

Quem são os utilizadores[11]? Os utilizadores de dados envolvem pessoas de todas as fases da cadeia de informação (Figura 3). No caso de dados primários de espécies, inclui utilizadores da própria instituição que produziu os dados primários como taxonomistas, gestores, investigadores, técnicos, coletores, assim como utilizadores externos e a jusante como políticos e decisores, cientistas, agricultores, florestais e horticultores, gestores ambientais, ONG's (ambientais e de produção), médicos, farmacêuticos, profissionais de

[11] No Brasil, usuários.

indústria, jardins botânicos e zoológicos, público em geral (incluindo jardineiros amadores) e utilizadores comunitários. Os dados de ocorrência de espécies têm imensos utilizadores e envolvem praticamente toda a comunidade, de uma forma ou de outra.

Os dados primários de espécies nem sempre foram recolhidos, tendo em conta a futura utilização pela comunidade. Tradicionalmente, os dados, especialmente de museus e herbários, tinham como principal objetivo fornecer informação para investigação taxonómica ou biogeográfica. Este foi um processo essencial, mas no mundo de hoje os provedores de financiamento para estas instituições, muitas vezes as agências governamentais, estão à procura de um maior retorno financeiro, e portanto, que os dados tenham maior valor através da sua disponibilidade para usos adicionais. Em particular, os governantes olham para o uso de dados procurando usá-los para uma melhor tomada de decisão, gestão ambiental ou planeamento em conservação (Chapman e Busby 1994), e os curadores destes dados não podem dar-se ao luxo de ignorar as necessidades dos seus utilizadores. Com um bom mecanismo de *feedback*, os utilizadores podem dar a sua opinião sobre a qualidade dos dados, o que pode ser uma importante ligação na cadeia de qualidade de dados como discutido abaixo.

Determinar as necessidades dos utilizadores é um trabalho difícil[12] *e árduo. Mas não há outra solução senão fazê-lo e a recompensa é ótima.*

[12] No Brasil, difício.

2. Princípios de Qualidade de Dados

> "*A experiência tem mostrado que o tratamento de dados a longo prazo e a sua gestão dentro de uma estrutura coordenada produz uma economia considerável de valor persistente*" (NLWRA 2003).

Os princípios de qualidade de dados necessitam de ser aplicados em todas as fases do processo de gestão de dados (colheita, digitalização, armazenamento, análise, apresentação e uso). Existem duas chaves para o melhoramento da qualidade dos dados – a prevenção e a correção. A prevenção de erros está diretamente ligada tanto com a recolha dos dados como com a inserção desses dados numa base de dados. Se bem que podem e devem ser realizados esforços consideráveis na prevenção do erro, o facto é que estes continuarão a existir em grandes conjuntos de dados (Maletic e Marcus 2000) e a validação e correção de dados não podem ser ignoradas.

A prevenção dos erros é de longe superior à deteção de erros, já que a deteção é muito dispendiosa e não pode garantir 100% de sucesso (Dalcin 2004). A deteção de erros, no entanto, tem um papel importante a desempenhar quando se trata de coleções históricas (Chapman e Busby 1994, English 1999, Dalcin 2004) que é o caso de muitos dados primários de espécie ou de ocorrência de espécies aqui considerados.

Comece por definir uma visão dos dados, desenvolver uma política de dados e implementar uma estratégia de dados – não por desenvolver atividades de "limpeza de dados" não planeadas[13], não coordenadas e não sistémicas.

2.1. A Visão

É importante para as organizações terem uma visão no que se refere a uma boa qualidade de dados. Isto aplica-se a organizações que queiram disponibilizar os dados a outros. Uma boa visão da qualidade de dados normalmente aumenta a visão global das organizações (Redman 2001) e aumenta os procedimentos operacionais da organização. No desenvolvimento desta visão, os gestores devem focar-se em conseguir um enquadramento da gestão integrada no qual a liderança, pessoas, *hardware*, aplicações (*software*), controlo de qualidade e os dados são trazidos em conjunto com ferramentas adequadas, linhas orientadoras e procedimentos padrão para manter os dados e torná-los em produtos de informação de qualidade (NLWRA 2003).

Uma visão de qualidade de dados:

- Força uma organização a pensar sobre os seus dados a longo prazo e sobre a necessidade de informação e a sua relação com o sucesso da organização a longo prazo,
- motiva ações na direção correta, e.g. no sentido da qualidade,
- fornece uma base sólida para a tomada de decisão tanto dentro como fora da organização,

[13] No Brasil, planejadas.

- formaliza o reconhecimento dos dados e informação como sendo ativos fundamentais da organização,
- maximiza a utilização dos dados e informação da organização, evita a duplicação, facilita parcerias e melhora a equidade de acesso,
- e maximiza a integração e interoperabilidade.

2.2. Política

Tal como uma visão, uma organização necessita de ter uma política para implementar essa visão. O desenvolvimento de uma política de qualidade de dados sólida:

- Força a organização a pensar de modo abrangente acerca da qualidade e a reexaminar as suas práticas diárias,
- Formaliza o processo de gestão dos dados,
- Ajuda a organização a ser mais clara acerca do seus objetivos a respeito de
    - redução de custos,
    - melhorar a qualidade de dados,
    - melhorar o serviço e relação com o cliente e
    - melhorar o processo de tomada de decisões,
- Proporciona aos utilizadores confidencialidade e estabilidade no acesso e uso dos dados provenientes dessa instituição,
- Melhora a relação e comunicação com os clientes da organização (tanto os provedores de dados como os seus utilizadores),
- Melhora a posição da organização perante a comunidade, e
- Melhora as possibilidades de ter melhor financiamento à medida que os objetivos de boas práticas são alcançados.

2.3. A Estratégia

Devido à vasta quantidade de dados mantidos pelas instituições, há necessidade de desenvolver uma estratégia para a captura e verificação dos dados (ver também sob *Priorização,* abaixo). Uma boa estratégia a seguir (tanto para a inserção de dados como para o controlo da qualidade) é definir objetivos de curto, médio e longo-prazo. Por exemplo (segundo Chapman e Busby 1994):

- **Curto-prazo.** Dados que podem ser montados e verificados durante um período de 6 a 12 meses (geralmente inclui dados que já estão numa base de dados e novos dados que exigem uma menor verificação de qualidade).
- **Intermédio.** Dados que podem ser inseridos numa base de dados durante um período de 18 meses, com um pequeno investimento em recursos e dados cuja avaliação pode ser realizada usando métodos internos simples de qualidade.
- **Longo-prazo.** Dados que podem ser inseridos e/ou verificados num espaço de tempo mais alongado usando acordos colaborativos, métodos de verificação mais sofisticados, etc. Pode envolver trabalhar com a coleção de forma sistemática selecionando:
    - Grupos taxonómicos que tenham sido recentemente revistos ou que estejam no processo de estudo dentro da instituição.

- Coleções importantes (tipos, coleções especiais de referência, etc.)
- Grupos chave (famílias importantes, taxa com significância nacional, taxa ameaçados, taxa de importância ecológica/ambiental).
- Taxa de regiões geográficas chaves (e.g. de países em desenvolvimento com o objetivo de partilha de dados com os países de origem, áreas geográficas de importância para a instituição).
- Taxa que fazem parte de acordos de colaboração com outras instituições (por exemplo, um acordo para a base de dados do mesmo taxa em várias instituições).
- Movendo-se de forma sistemática, desde o início ao fim da coleção.
- As aquisições recentes, de em detrimento de coleções de registadas anteriormente.

Alguns dos princípios de uma boa gestão de dados que devem ser incluídos na estratégia são (de acordo com NLWRA 2003):

- Não reinventar os mecanismos de gestão da informação.
- Procura de eficiência na recolha de dados e procedimentos de controlo de qualidade.
- Partilhar dados, informação e ferramentas sempre que possível.
- Usar padrões existentes ou desenvolver padrões novos e robustos em conjunto com outros.
- Promover o desenvolvimento de redes e parcerias.
- Apresentar um caso de negócios para a coleção e gestão de dados.
- Reduzir a duplicação na colheita de dados e controlo da qualidade de dados.
- Olhar para além do uso imediato dos dados e examinar as exigências dos utilizadores.
- Assegurar a implementação de uma boa documentação e metadados

2.4. A prevenção é melhor que a cura

O custo de introduzir uma coleção numa base de dados pode ser substancial (Armstrong, 1992), mas é apenas uma fração do custo de verificar e corrigir os dados numa data posterior. É melhor prevenir erros do que corrigi-los mais tarde (Redman 2001) sendo de longe a opção mais barata. Fazer correções retrospetivamente pode significar que os dados com erros já tenham sido usados numa série de análises antes de terem sido corrigidos, causando custos *A posteriori* das decisões tomadas com dados pobres ou de realização de novas análises.

A prevenção de erros nada faz aos erros já existentes na base de dados, no entanto, a validação e limpeza de dados continua a ter um papel importante no processo da qualidade destes dados. O processo de limpeza é importante na identificação da causa dos erros que já estão incorporados na base de dados e deve levar a procedimentos que garantam que estes erros não sejam repetidos. A limpeza não pode acontecer isoladamente, senão o problema nunca desaparecerá. As duas operações, limpeza de dados e prevenção de erros, devem ocorrer simultaneamente. Decidir limpar os dados primeiro e preocupar-se com a prevenção mais tarde, geralmente significa que a

prevenção do erro nunca é realizada de forma satisfatória e, entretanto, mais e mais erros são adicionados à base de dados.

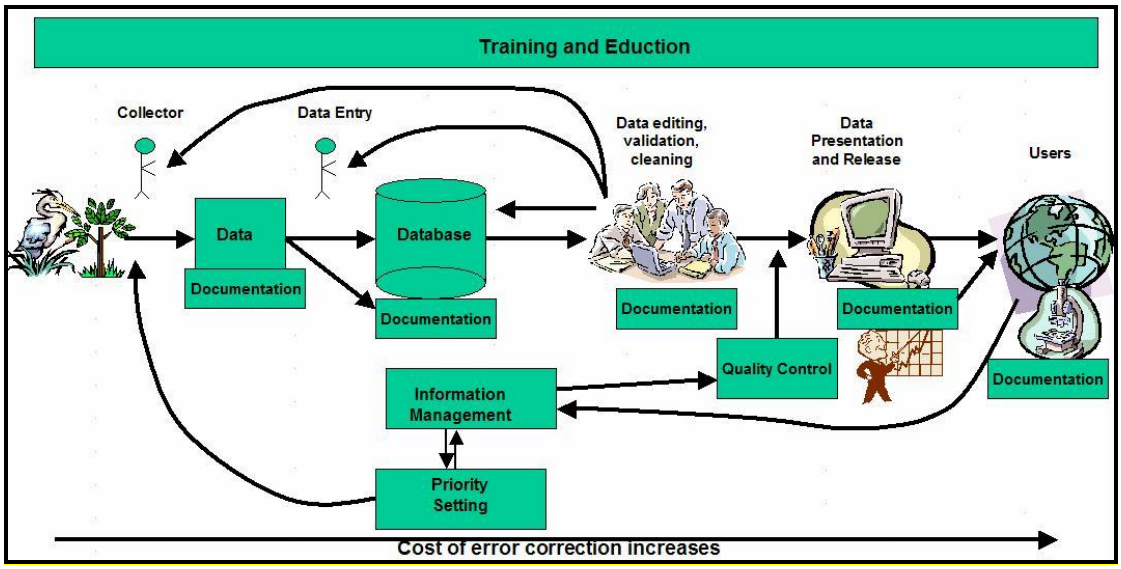

Fig. 4. *Cadeia de gestão de informação que mostra o aumento do custo da correção dos erros à medida que se avança na cadeia. Boa documentação, educação e formação estão integrados em todos os passos.*

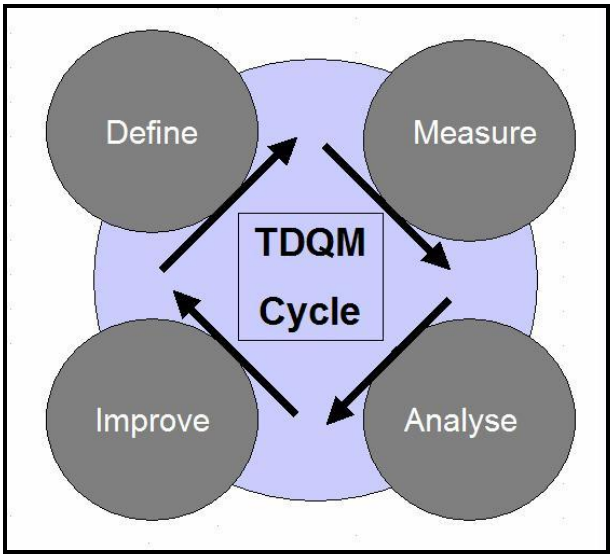

Fig. 5. *O Ciclo da gestão da qualidade dos dados, que mostra a natureza cíclica do processo de gestão de dados (segundo Wang 1998).*

Conservadores[14] e donos de dados (agências de coleções particulares como museus e herbários) são amplamente responsáveis pela qualidade dos seus dados. No entanto, aqueles que fornecem e usam os dados também têm responsabilidades.

Atribua a responsabilidade da qualidade dos dados àqueles que os criaram. Se isto não for possível, atribua a responsabilidade o mais próximo à criação dos dados (Redman 2001).

[14] No Brasil, guardiões.

2.5. O coletor tem a responsabilidade primária

A responsabilidade primária pela gestão da qualidade de dados fica com o coletor dos dados. É da sua responsabilidade ter a certeza de que:

- informação da etiqueta está correta,
- informação da etiqueta está rigorosamente registada e documentada,
- a informação de localização é tão precisa quanto possível, e tanto a exatidão como a precisão são documentadas,
- as metodologias de colheita estão totalmente documentadas,
- as etiquetas ou notas de campo estão claras e não ambíguas, e
- a informação da etiqueta é legível e de fácil leitura pelos técnicos que inserem os dados.

Se a informação da etiqueta ou no caderno de campo do coletor não estiver clara e correta, é extremamente difícil corrigir retrospetivamente. Isto é menos importante no que diz respeito à parte taxonómica dos dados em casos onde são mantidas colheitas *voucher, uma vez que,* normalmente, é verificada por especialistas em data posterior.

É também importante que as notas de localização e de informação subsidiária sejam feitas na mesma altura da colheita ou observação e não sejam deixadas para o final do dia ou até voltarem para o laboratório, como aconteceu frequentemente no passado.

A maioria dos dados chega a uma organização a partir de "fornecedores", e é mais fácil desenvolver boas práticas de colheita de dados do que corrigir erros a jusante.

2.6. O conservador ou curador tem a responsabilidade central ou de longo-prazo.

O conservador (ou guardião) dos dados (museu, herbário, universidade, agência de conservação, ONG ou individuo privado) tem responsabilidade a longo prazo para manter e melhorar a qualidade dos dados enquanto tiver a responsabilidade pelos mesmos (veja, por exemplo, a lista de responsabilidades de custódia em Olivieri *et al*. 1995, p. 623). É importante que o conservador da organização que tenha a custódia dos dados assuma responsabilidade acrescida na gestão da qualidade de dados na sua instituição, mas também é essencial que a organização tenha uma cultura de qualidade de dados de tal forma que cada indivíduo dentro da organização saiba que tem uma parte de responsabilidade na qualidade dos dados mantidos pela organização. É da responsabilidade do conservador assegurar que:

- os dados são transcritos para a base de dados corretamente e com exatidão a partir das notas do coletor,
- procedimentos de controlo de qualidade de dados são implementados e aplicados durante a sua captura,
- os dados e a sua qualidade estão adequadamente documentados e com exatidão,
- verificações de validade são feitas regularmente nos dados,
- verificações de validade estão documentadas no seu todo,

- os dados são armazenados e arquivados convenientemente (veja notas sobre armazenamento abaixo),
- versões anteriores são sistematicamente armazenadas para permitir comparações e voltar à versão "não limpa" dos dados,
- a integridade dos dados é mantida,
- os dados são disponibilizados atempadamente e de forma rigorosa com a documentação que permite aos utilizadores determinar a "aptidão para o uso",
- a responsabilidade de custodia quanto à privacidade, direitos de propriedade intelectual, direitos de autor e sensibilidades dos proprietários tradicionais / originais são mantidas,
- as condições de uso dos dados são mantidas e tornadas disponíveis em conjunto com nenhumas restrições ao uso e a áreas conhecidas de dados inadequados,
- todos os requisitos legais a respeito dos dados são honrados e respeitados,
- o *retorno de comentários* dos utilizadores acerca da qualidade de dados é realizado em tempo útil,
- a manutenção da qualidade de dados é mantida aos mais altos níveis em todos os momentos,
- todos os erros conhecidos estão documentados e são conhecidos pelos utilizadores.

A propriedade e custódia de dados não só confere direitos de gestão e controlo de acesso aos dados, como confere responsabilidades pela sua gestão, controlo de qualidade e manutenção. Quem possui a custódia dos dados tem também a responsabilidade moral de fiscalizá-los para utilização pelas gerações futuras.

2.7. Responsabilidade do utilizador

Os utilizadores de dados também têm responsabilidade na sua qualidade. Os utilizadores necessitam de dar a conhecer informações sobre quaisquer erros ou omissões que possam encontrar, erros na documentação dos dados e informações adicionais que possam ser necessárias no futuro, etc. Muitas vezes, é o utilizador, quando olha para os dados no contexto de outros dados, que pode identificar erros e valores discrepantes que de outra forma iriam passar despercebidos. Um único museu pode ter apenas um subconjunto dos dados totais disponíveis (de um estado ou região, por exemplo), e é apenas quando os dados são combinados com os dados de outras fontes que os erros podem tornar-se evidentes.

Dependendo dos objetivos da colheita de dados numa instituição, o utilizador também pode dar contribuições valiosas na definição de prioridades para o futuro no que diz respeito à colheita de dados e sua validação (Olivieri *et al*. 1995).

O utilizador também tem responsabilidade para determinar a aptidão dos dados para o uso e não usar os dados de maneira inapropriada.

Os utilizadores e coletores têm um papel importante a desempenhar em dar assistência ao conservadores na manutenção da qualidade dos dados nas coleções e ambos têm um interesse declarado de que os dados tenham a maior qualidade possível.

2.8. Criação de parcerias

A criação de parcerias para a manutenção da qualidade de dados pode ser uma medida gratificante e pode auxiliar na diminuição de custos. Isto é particularmente válido para museus e herbários, onde registos duplicados estão distribuídos por diversos museus. Muitas comunidades de bibliotecas trabalham em colaboração e estabelecem parcerias para melhorar a catalogação dos seus materiais (*Library of Congress* 2004), os museus e herbários poderiam facilmente operar de forma similar. Estas parcerias e acordos de colaboração podem ser desenvolvidos com:

- coletores de dados importantes (com o objetivo de melhorar o fluxo de informação – por exemplo, desenvolvendo padrões para colheita de dados e de formulários de relatório, fornecimento de dados de GPS, etc.),
- outras instituições que detenham dados semelhantes (e.g. duplicados de coleções),
- outras instituições afins com necessidades de qualidade de dados semelhantes e que possam desenvolver métodos de controlo de qualidade de dados, ferramentas, padrões e procedimentos,
- intermediários de dados chave (como o GBIF) que desempenham um papel na colheita e distribuição de informação a partir de inúmeros fornecedores de dados,
- utilizadores dos dados (especialmente aqueles que possam realizar testes de validação nos dados durante ou antes do da análise), e
- estatísticos e auditores de dados que podem melhorar as metodologias de gestão de dados, fluxo de dados e técnicas de qualidade de dados.

A sua instituição não é a única a lidar com a qualidade de dados.

2.9. Priorização

Para tornar os dados de elevado valor para a maioria dos utilizadores num curto espaço de tempo, pode ser necessário dar prioridade à colheita e validação dos dados (veja também os comentários sobre *Integridade*, abaixo). Para fazer isto, pode ser necessário:

- focar em primeiro lugar nos dados mais críticos,
- concentrar em unidades discretas (taxonómicas, geográficas, etc.),
- dar prioridade a espécimes tipo e *vouchers* que sejam importantes,
- ignorar dados que não são usados ou para os quais a qualidade de dados não pode ser garantida (e.g. registos com informação geográfica pobre, mas tenha em mente a importância histórica de alguns dados com pobre informação geográfica),
- considerem os dados que são de valor mais amplo, são de maior benefício para a maioria dos utilizadores e são de valor para os mais diversos usos,
- trabalhem em áreas em que grandes quantidades de dados podem ser limpos com o menor custo (por exemplo, através do uso de processamento em lote).

Nem todos os dados são criados da mesma forma, portanto foque-se nos mais importantes e se a limpeza de dados é requerida, assegure-se de que nunca terá de ser repetida.

2.10. Completude

As organizações devem esforçar-se pela completude dos dados (ou de unidades discretas do dados através da priorização, e.g. para uma categoria taxonómica, uma região, etc.) para que todos os registos elegíveis sejam usados na compilação de dados. É melhor completar a informação de uma unidade discreta e torná-la disponível, do que ter imensos dados incompletos disponíveis, dado que análises realizadas sobre dados incompletos não são compreensíveis. É também importante ter uma política de dados que defina limites de dados ausentes e as respostas correspondentes, juntamente com uma política de documentação da integridade dos dados (ver em *Documentação*, abaixo).

2.11. Validade e Atualidade

Existem três fatores chave relacionados com a atualidade e a validade dos dados:

- Em que período os dados foram recolhidos?
- Quando foram atualizados os dados para refletir mudanças no mundo real?
- Por quanto tempo é suscetível que os dados se mantenham atualizados?

A validade dos dados é uma questão frequentemente levantada pelos utilizadores. Muitos conservadores de dados tendem a usar a validade para se referirem ao período em que os dados foram originalmente recolhidos ou pesquisados. Devido ao atraso entre a colheita e a publicação dos dados (que para dados biológicos pode ser um tempo excessivamente longo) a informação publicada é uma representação de "o que era" e não de "o que é". A maioria dos utilizadores de dados de biodiversidade estão cientes disso e isto constitui um dos valores deste tipo de dados e o que os torna bastante diferentes da maioria dos outros tipos de dados.

Nos termos da gestão de qualidade de dados, validade é normalmente usada no contexto do período de tempo em que os dados são “usados até” (por vezes também chamado por atualidade) e que podem estar relacionados com a ultíma vez em que os dados foram revistos[15] ou atualizados. Isto pode ser especialmente relevante no que diz respeito ao nome ligado aos dados. Quando foi a última atualização e se estão de acordo com a última taxonomia? Onde as regras taxonómicas modernas são seguidas, se uma espécie é dividida numa série de taxa menores, um deles mantém o nome do conceito mais amplo. Pode ser importante para o utilizador saber se o nome utilizado se refere ao conceito mais extenso ou mais curto. Validade pode ser utilizada como equivalente à data “usar até” usada nos produtos alimentares, além do qual o conservador não garante a informação nomenclatural anexada ao registo.

Pode acontecer que para a maioria das bases de dados a validade e atualidade dos dados não sejam relevantes ou possíveis de incluir ou manter. Isto pode ser aplicado a grandes coleções de museus ou herbários, por exemplo. Por outro lado, pode ser importante para dados de observação ou de pesquisa onde não existam espécimes associados, ou onde não haja atualizações dos dados tendo em conta as revisões taxonómicas recentes. Também é uma questão importante para coleções secundárias, incluindo coleções que tenham sido reunidas por uma agência externa a partir de um conjunto de agências. Um exemplo pode

[15] No Brasil, revisados.

ser quando um conjunto de instituições de um país em desenvolvimento tornam os seus dados disponíveis através de uma instituição que hospede os dados para serem providos ao portal do GBIF e não são apresentados ao vivo a partir da base de dados.

2.12. Frequência de atualização

A frequência de atualização dos dados dentro de uma base de dados está relacionada com a validade e a atualidade e a necessidade de formalizar e documentar. Isto inclui a adição de novos dados, bem como a frequência de divulgação de dados corrigidos. Ambos têm efeito sobre a qualidade dos dados e são, portanto, importantes para os utilizadores. Um utilizador não quer ter o trabalho de fazer o *download* ou obter uma base de dados que está prestes a ser atualizada e melhorada.

2.13. Consistência

Redman (1996) reconheceu dois aspetos de consistência: Consistência Semântica – onde a visualização dos dados deve ser clara, inequívoca e consistente; e a Consistência Estrutural – na qual o tipo de entidades e atributos deve ter a mesma estrutura base e formato. Um exemplo simples de consistência de semântica é quando os dados estão sempre nos mesmos campos, e por isso são fáceis de encontrar - por exemplo, há campos separados para a categoria infraespecifica e nome infraespecífico de forma a que seja sempre claro que o campo do nome da infraespécie inclui só um nome ou epíteto (ver tabela 1) e não está misturado, de forma que algumas vezes inclui só o nome, e noutras inclui um prefixo de “var.” ou “subsp.” seguido pelo nome, etc. (ver tabela 2)

Género	Espécie	Infraespécie
Eucalyptus	globulus	subsp. bicostata
Eucalyptus	globulus	bicostata

Tabela 1. *A tabela mostra inconsistências de semântica no campo da infraespécie.*

Género	Espécie	Infrasp_rank	Infraespécie
Eucalyptus	globulus	subsp.	bicostata
Eucalyptus	globulus		bicostata

Tabela 2. *A tabela mostra consistência semântica no campo da infraespécie adicionando um novo campo (“Infrasp_rank”).*

Um bom desenho de uma base de dados relacional não vai permitir que muitas destas questões ocorram, no entanto, muitas das bases de dados existentes utilizadas por instituições com coleções não estão bem projectadas.

A consistência estrutural ocorre onde há consistência dentro de um campo, por exemplo o campo “Infrasp_categ” (Tabela 2) deve ter sempre subespécies registadas do mesmo modo – não umas vezes como “subsp.”, outras como “ssp.”, “subespecies”, “subspec.”, “Espécies”, etc. Isto pode ser evitado através de um bom desenho da base de dados com atributos bem estruturados.

A consistência tanto nos métodos como na documentação é importante pois permite ao utilizador saber que testes e como foram realizados, onde encontrar a informação e como interpretar importantes porções de informação. A consistência, no entanto, necessita de ser balançada com a flexibilidade (Redman 2001).

2.14. Flexibilidade

Os conservadores de dados devem manter flexibilidade no seu método de controlo da qualidade de dados, pois embora muitos dados biológicos sejam similares na natureza, diferentes abordagens à qualidade de dados podem ser adequadas para dados de diferentes regiões (por exemplo, que conjuntos de dados associadas estão disponíveis para se comparar com os dados), diferentes grupos taxonómicos (organismo aquático versus terrestres, etc.), ou diferentes métodos de colheita de dados (observação ou registos de levantamentos versus coleções de museus com espécimes, etc.).

As opiniões taxonómicas são, na realidade, hipóteses e opiniões (hipóteses) taxonómicas diferentes (válidas) podem levar a que um mesmo organismo seja classificado de forma diferente por diferentes taxonomistas e, assim, ter um ou mais nomes alternativos – cada um dos quais pode ser igualmente válido (Pullan *et al*. 2000, de Knapp *et al*. 2004). Um exemplo é quando dois taxonomistas discordam quanto à colocação de um taxa em géneros diferentes – por exemplo, alguns taxonomistas colocam certas espécies no género *Eucalyptus*, enquanto que outros acreditam que pertence ao género *Corymbia*. Na prática, e especialmente em zoologia, o ponto de vista do revisor mais recente é aceite a menos que haja uma boa razão para rejeitar essa opinião.

A flexibilidade permite a capacidade de alterar uma determinada visão acomodando uma nova ou diferentes solicitações[16]. Trabalhos recentes publicados por *Taxonomic Databases Working Group* (TDWG)[17] e outros focaram-se em estruturas de bases de dados que permitam apresentar esses conceitos alternativos (Berendsohn 1997) e, embora a natureza desta flexibilidade, deste modo, possa parecer reduzir a qualidade, na realidade, permite aos utilizadores uma maior flexibilidade na determinação da aptidão para o uso e, nesses casos pode aumentar a perceção da qualidade.

2.15. Transparência

A transparência é importante porque transmite confiança na avaliação daqueles que usam os dados. A transparência significa garantir que os erros não são escondidos, mas sim identificados e reportados, que a validação e os procedimentos do controlo de qualidade estão devidamente documentados e disponibilizados, e que os mecanismos de retorno de comentário estão disponíveis e são encorajados.

Um exemplo onde a transparência é importante é na documentação das metodologias de colheita (especialmente importante para dados de observação e de levantamento). Mais uma vez, isto apoia o utilizador a ser capaz de determinar se os dados são adequados para o uso que pretende.

[16] No Brasil, demandas.
[17] http://www.tdwg.org/

2.16. Medidas e metas de desempenho

As medidas de desempenho são uma adição válida aos procedimentos do controlo de qualidade, e asseguram que cada utilizador individual dos dados pode confiar no nível de exatidão ou qualidade dos dados. Medidas de desempenho podem incluir verificação estatística dos dados (por exemplo 95% de todos os registos estão a 1,000 metros da posição reportada), no nível de controlo de qualidade (por exemplo – 65% de todos os registos foram verificados por um taxonomista qualificado nos últimos 5 anos; 90% foram verificados por o taxonomista qualificado nos últimos 10 anos), integridade (todas as quadrículas de 10 minutos foram amostradas), etc., etc.

Medidas de desempenho ajudam a quantificar a qualidade de dados. As vantagens são que:

- a organização assegura a si própria que certos dados são de alta qualidade documentada;
- eles auxiliam na gestão dos dados e na redução da redundância, e
- eles ajudam na coordenação dos vários aspetos da cadeia da qualidade de dados e assim podem ser organizados e ser usados por diferentes técnicos.

Antes de medir os níveis de qualidade de dados, primeiro considere como os utilizadores podem usá-los e então estruture os resultados para que possam ser usados mais eficientemente.

2.17. Limpeza de dados

Os princípios de limpeza de dados serão desenvolvidos no documento associado *Princípios e métodos de limpeza de dados*. Basta dizer que um enquadramento geral de limpeza de dados conforme modificado por Maletic e Marcus (2000) é:

- Definir e determinar tipos de erros
- Pesquisar e identificar casos de erros
- Corrigir erros
- Documentar casos e tipos de erros
- Modificar os procedimentos de entrada de dados para reduzir a incidência de erros semelhantes no futuro.

Não seja seduzido pela aparente simplicidade das ferramentas de limpeza de dados. São válidas e ajudam a curto prazo, mas a longo prazo, não há substituto para a prevenção do erro.

2.18. Anómalos

A deteção de anómalos (geográficos, estatísticos e ambientais) pode providenciar um dos testes de avaliação mais úteis para encontrar possíveis erros nos dados espaciais. É importante, de qualquer modo, que os testes não apaguem indiscriminadamente dados por serem identificados como anómalos estatísticos. Em dados ambientais isso é notório, quando registos perfeitamente corretos aparecem como anómalos estatísticos. Isto pode dever-se a padrões evolutivos históricos, a regimes de alterações climáticas, vestígios

deixados por atividade humana, etc. A exclusão indiscriminada de anómalos pode remover registos valiosos do conjunto de dados e distorcer análises futuras.

Os utilizadores, por outro lado, podem decidir eliminar anómalos da sua análise se tiverem dúvidas da sua validade como bons registos. A identificação de anómalos não só ajuda os conservadores de dados a identificar possíveis erros, como pode ajudar os utilizadores a determinar se os registos individuais de dados têm ou não aptidão para o uso na sua análise.

A deteção de anómalos pode ser um bom método de validação, mas nem todos os anómalos são erros.

2.19. Estabelecer metas de melhoria

A definição de metas simples e fáceis de quantificar pode levar a uma melhoria na qualidade dos dados. Uma meta como reduzir para metade a percentagem de novos registos mal georreferenciados a cada seis meses, durante dois anos, pode conduzir à redução total da quantidade de erros em 94% (Redman 2001). Tais metas devem concentrar-se em:

- ser claras e agressivas em relação aos prazos,
- taxas de melhoria em vez de valores de qualidade reais,
- definições claras (tal como para "mal georreferenciado"),
- metas que sejam simples e atingíveis.

Metas a longo prazo podem também ser introduzidas simultaneamente com o processo de redução do tempo (não adiciona valor) para metade, a cada ano, necessário para a limpeza de dados, melhorando as técnicas de entrada e de validação de dados.

As metas de desempenho são uma boa forma de uma organização manter consistente o seu nível de verificação e validação da qualidade de dados, por exemplo 95% de metade dos registos estão documentados e validados num período de 6 meses após receção.

2.20. Auditoria

É importante para os conservadores saber que dados foram verificados e quando. Isto ajuda a evitar a redundância e perda de dados através de falhas. A melhor maneira de o fazer é manter um registo do processo de auditoria e da validação.

2.21. Controlos da edição

Os controlos da edição envolvem regras que determinam os valores permitidos para cada campo. Por exemplo, o valor no campo do mês deve estar entre 1 e 12, o valor para o dia deve estar entre 1 e 31 com o valor máximo a depender do mês, etc. Regras uni-variadas aplicadas a um único campo (e.g. o exemplo do mês, acima) e regras bivariadas aplicadas a dois campos (e.g. a combinação do dia e do mês).

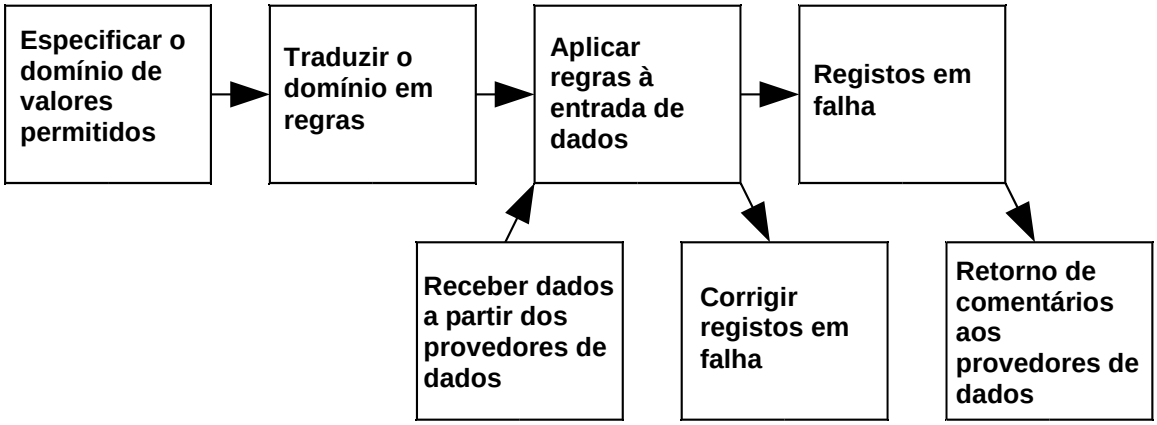

Fig. 6. *Uso do controlo de edição (modificado a partir de Redman 2001).*

Um segundo exemplo é com dados de coordenadas. Testes de alcance simples irão testar (caso os dados estejam em latitudes e longitudes) que a latitude é entre 0 e 90 graus, minutos e segundos estão entre 0 e 60, etc. Uma vez que se transforme os dados para UTM, no entanto, torna-se mais complicado. Muitas vezes, uma base de dados que inclui os dados a partir de uma pequena região que se enquadra numa zona UTM não irá incluir a zona dentro da base de dados. Isto pode parecer bastante aceitável, desde que os dados nunca sejam combinados com dados de outras regiões. Mas quando é feita uma tentativa para combinar os dados, estes tornam-se bastante inutilizáveis. Assim, os controlos de edição precisam assegurar que a Zona apropriada seja sempre incluída.

2.22. Minimizar a duplicação e reformulação[18] de dados

A experiência no mundo dos negócios tem demonstrado que o uso da cadeia de gestão da informação (ver figura 3) pode reduzir a duplicação e o refazer de dados e levar a uma redução das taxas de erro até 50%, e reduzir os custos resultantes da má utilização de dados até dois terços (Redman, 2001). Isto deve-se principalmente aos ganhos de eficiência através da atribuição de responsabilidades claras para a gestão de dados e do controlo de qualidade, minimizando assim afunilamentos e tempos de espera, duplicação através de verificações de controlo de qualidade por diferentes membros da equipa, e melhorando a identificação dos melhores métodos de se trabalhar.

2.23 Manutenção de dados originais (ou verbatim)

É importante que os dados originais sejam registados pelo coletor, ou até inseridos posteriormente por curadores, etc. e não sejam perdidos no processo de edição e de limpeza de dados. As alterações às bases dados feitas durante o processo de limpeza de dados devem ser inseridas como informação adicional, mantendo também a informação original. Uma vez a informação apagada, é difícil ou mesmo impossível recuperar. Isto pode ser particularmente importante para informação do coletor e localização. O que parece posteriormente para um curador como um erro pode não ser um erro real. As alterações do nome de uma localização para outro (e.g. de Checoslováquia para República

[18] No Brasil, retrabalho.

Checa, por exemplo), altera não só o nome mas também os limites. Pode ser importante mais tarde, saber o que foi escrito originalmente e não ter só a versão "corrigida". Veja também os comentários em *Arquivamento*.

2.24. Categorização pode levar à perda de qualidade dos dados

Categorização de dados pode muitas vezes levar à perda de dados e à redução da qualidade geral dos dados. Um exemplo pode ser com a recolha de dados com informações de localidade detalhadas (e possivelmente até mesmo georreferenciados), e de seguida armazenar os dados com base numa grelha.

É quase sempre melhor armazenar os dados na sua resolução mais fina, e de seguida, classificá-los, se isso for necessário para um determinado uso. Se um utilizador precisa de produzir um mapa presença/ausência numa grelha de 10 X 10 minutos, é fácil fazer a partir de dados armazenados como pontos, mas se os dados tiverem sido armazenados na base de dados em grelha, é impossível fazer algo com os dados numa escala mais fina.

Isto faz com que seja extremamente difícil (e talvez até mesmo impossível) combinar dados que podem ter sido classificados segundo uma escala em grelha ou de origem diferente. O mesmo é o caso com dados descritivos - se os dados são classificados em estados podem ser necessários para uma chave (por exemplo> 6m = árvore; <6m = arbusto), e novos dados são obtidos a partir de outra fonte que usou em vez de 4m ao invés de 6m na sua definição para árvore, então o que fazer com aqueles dados entre 4 e 6 metros. É muito melhor armazenar os dados em metros exatos, e preocupar-se sobre se é uma árvore ou arbusto mais tarde.

Um caso em que isto ocorre com frequência é no armazenamento com exatidão da georreferenciação. É recomendado armazenar sempre este tipo de dados em metros, mas uma grande quantidade de bases de dados fá-lo em categorias (<10m, 10-100m, 100-1000m, 1000-10000 m). Se houver um registo ao qual seja possível determinar de forma exata para 2 km, então perde-se imediatamente informação ao colocá-lo na categoria de exatidão de 10km.

2.25. Documentação

Uma boa documentação é um princípio fundamental da gestão de dados. Sem uma boa documentação, o utilizador não pode determinar a adequação dos dados para o uso que têm em mente e, portanto, não pode determinar a qualidade dos dados para o efeito. Uma discussão mais detalhada sobre a documentação é realizada na *Documentação* abaixo.

2.26. Retorno de comentários

É essencial que os conservadores de dados encorajarem o retorno de comentários dos utilizadores acerca dos seus dados e que os considerem seriamente. Como mencionado acima na *Responsabilidade do utilizador*, o utilizador muitas vezes tem uma melhor oportunidade de detetar certos tipos de erro através da combinação de dados de várias fontes, do que o conservador de dados que trabalha isoladamente.

O desenvolvimento de bons mecanismos de retorno de comentários nem sempre é uma tarefa fácil. Pode-se colocar um botão de retorno de comentários na página de interface

da pesquisa, ou num anexo enviado aos utilizadores no momento do *download* dos dados definindo a metodologia para o retorno de informação sobre os erros nos dados e como enviar comentários para os conservadores. Alguns destes métodos são desenvolvidos no documento associado sobre os *Princípios e Métodos de Limpeza de Dados*.

Os canais efetivos de retorno de comentários entre os utilizadores e fornecedores são um mecanismo fácil e produtivo para a melhoria da qualidade dos dados.

2.27. Educação e Formação

A educação e formação em todos os níveis da cadeia de informação pode levar a uma grande melhoria na qualidade dos dados (Huang *et al.*, 1999). Isso começa com a formação e educação dos coletores no uso de bons procedimentos de colheita e na implementação das necessidades dos utilizadores da informação, através da formação dos técnicos que inserem os dados e técnicos responsáveis pela gestão diária das bases de dados, através da educação dos utilizadores finais em relação à natureza dos dados, suas limitações e usos potenciais. Os aspetos de educação e formação de qualidade dos dados estão em grande parte dependentes de uma boa documentação.

Um exemplo da integração de dados de controlo de qualidade, educação e formação pode ser visto no projeto de georreferenciação MaPSTeDI (*University of Colorado*, 2003). O processo envolve a verificação de um determinado número de registos de cada técnico de georreferenciação. Com um novo técnico, os primeiros 200 registos são verificados quanto à precisão por um supervisor. Isto não só mantém a qualidade dos dados, como permite que o técnico aprenda e melhore não voltando a cometer erros. Dependendo do técnico, um adicional de 100 registos podem ser verificados à medida a que o técnico se torne mais experiente, sendo reduzida a uma seleção aleatória de 10% dos registos e, eventualmente, para cerca de 5%. Se a percentagem de erros descobertos for elevada, então os registos devem ser verificados. Procedimentos bem projetados como estes podem ajudar na educação de novos utilizadores. Por outro lado, se não há procedimentos, há pouca forma de garantir a consistência entre técnicos e entre tarefas.

A má formação está na origem de muitos problemas na qualidade de dados.

2.28. Responsabilidade

A atribuição de responsabilidade pela qualidade geral dos dados pode ajudar as organizações a alcançarem um nível consistente de controlo de qualidade, fornecer um ponto de referência para o retorno de comentários sobre os erros, e fornecer um ponto de contato para a documentação e pesquisas.

3. Dados Taxonómicos e Nomenclaturiais

> "*Dados taxonómicos fracos podem 'contaminar' as áreas de estudo relacionadas*" (Dalcin 2004).

Taxonomia é a teoria e a prática de classificar organismos (Mayr e Ashlock 1991). A maioria dos dados de espécies aqui considerados tem uma componente taxonómica (ou nomenclatural) (e.g. o nome do organismo e a sua classificação) - denominado "Domínio de dados de classificação" por Dalcin (2004). A qualidade desta parte dos dados e como esta pode ser determinada difere consideravelmente da parte espacial dos dados, que normalmente é mais abstrata e mais difícil de quantificar.

Os dados taxonómicos consistem em (nem sempre presentes):

- Nome (cientifico, comum, hierarquia, categoria)
- Estatuto nomenclatural (sinónimo, aceite, tipificação)
- Referência (autor, local e data de publicação)
- Determinação (por quem e quando foi o registo classificado)
- Campos de qualidade (exatidão da determinação, qualificadores)

Uma das maiores fontes de erros nos nomes taxonómicos são os erros ortográficos. Detetar erros ortográficos numa base de dados taxonómica pode ser uma tarefa simples quando envolve nomes taxonómicos que representam hierarquias taxonómicas tais como os nomes da Família e Género (Dalcin 2004). Nestes casos, listas autoritárias estão normalmente disponíveis para a maioria dos grupos taxonómicos. Cada vez mais, também, listas completas com nomes das espécies estão a ser tornadas disponíveis em projetos como o *Species 2000*[19] e o GBIF[20]. O uso de nomes de espécies ou epíteto sozinhos sem o género associado como ficheiro autoritário, raramente é satisfatório, uma vez que muitos epítetos podem ter variações mínimas no nome de um género para outro. Um dos métodos para verificar erros ortográficos é detetar e isolar erros no nome científico, usando algoritmos de similaridade para identificar um par de nomes científicos que tenham um elevado grau de similaridade mas que não são exatamente o mesmo (Dalcin 2004, CRIA 2005). De longe, o método mais satisfatório para reduzir a probabilidade de erros ortográficos nos nomes científicos é criar ficheiros autoritários para o processo de entrada de dados nas bases de dados utilizando listas de seleção com o nome do género e espécie, da família, etc. Numa situação ideal onde os ficheiros autoritários estão disponíveis, o uso destas técnicas devem reduzir a incidência deste tipo de erros para praticamente zero. Infelizmente, existem enormes áreas do mundo, e um conjunto de grupos taxonómicos importantes para os quais não existem ainda listas disponíveis.

Onde os ficheiros autoritários são importados de fontes externas tais como o *Catalogue of Life* ou GBIF, o Source-Id deve ser registado na base de dados para que as alterações feitas entre edições da autoridade, possam ser facilmente incorporadas na base de dados e esta

[19] http://www.species2000.org
[20] http://www.gbif.org/species

seja atualizada. Espera-se que em pouco tempo isto possa vir a tornar-se fácil através do uso dos Identificadores Globais Únicos (GUIDs)[21].

A qualidade taxonómica dos dados depende muito das capacidades taxonómicas disponíveis. O *Taxonomic Impediment* (Environment Australia 1998) e a diminuição global de taxonomistas bem formados levará a uma diminuição a longo prazo da qualidade da produção taxonómica e na qualidade resultante dos dados primários de espécies (Stribling *et al*. 2003). A *Global Taxonomic Initiative (GTI)* (CBD 2004) tem tentado diminuir o chamado "impedimento taxonómico" mas é provável que a situação se mantenha no futuro. A qualidade pode também decair com o tempo, especialmente em casos onde os espécimes *voucher* não estão disponíveis ou mantidos (por exemplo para a maioria dos dados de observação e dos dados de levantamentos) ou naquelas áreas onde a especialização taxonómica relevante não está disponível.

A capacidade de uma instituição produzir resultados taxonómicos de alta qualidade (incluindo dados primários de espécies documentados) é influenciado por (segundo Stribling *et al*. 2003):

- o nível de experiência e formação dos colaboradores,
- o nível de acesso à bibliografia técnica, referências e coleções *voucher* e especialistas taxonómicos,
- possuírem equipamento e instalações de laboratório apropriados, e
- acesso à Internet e aos recursos aí disponíveis.

3.1. Registo da exatidão da identificação, etc.

Tradicionalmente, os museus e herbários têm tido em operação um sistema de determinação em que especialistas que trabalham com grupos taxonómicos, de tempos a tempos, examinam os espécimes e determinam a sua circunscrição ou identificação. Isto é realizado normalmente como parte de estudos de revisão, ou por um especialista que visita a instituição e revê as coleções durante essa visita. Este é um método comprovado, mas que consome muito tempo e que é realizado em grande parte ao acaso. É pouco provável que se possa evitar esta necessidade, uma vez que a identificação automática por computador é improvável no curto, ou mesmo, no longo-prazo.

Uma opção pode ser a incorporação de um campo nas bases de dados que forneça um indicador da certeza da identificação. A data de determinação é normalmente incorporada na maioria das bases de dados de coleção. Esta opção seria composta por um campo de código e pode ser uma das opções (Chapman 2004):

- identificado por especialista mundial do taxa, com elevada certeza
- identificado por especialista mundial do taxa, com grau de certeza razoável
- identificado por especialista mundial do taxa, com algumas dúvidas
- identificado por especialista regional do taxa, com elevada certeza
- identificado por especialista regional do taxa, com grau de certeza razoável
- identificado por especialista regional do taxa, com algumas dúvidas
- identificado não-especialista do taxa, com elevada certeza

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

- identificado não-especialista do taxa, com grau de certeza razoável
- identificado não-especialista do taxa, com algumas dúvidas
- identificado pelo coletor com elevada certeza
- identificado pelo coletor com certeza razoável
- identificado pelo coletor com algumas dúvidas

Como classificar estas categorias é um assunto em discussão, do mesmo modo se estas categorias são ou não as melhores. Eu percebo que existem algumas instituições que têm um campo desta natureza, mas nesta fase, não fui capaz de descobrir um exemplo. O padrão HISPID versão 4 (Conn 2000) inclui uma versão simplificada – a etiqueta de nível de verificação, com 5 códigos, a saber:

0	O nome do registo não foi verificado por nenhuma autoridade.
1	O nome do registo foi determinado por comparação com outras plantas já identificadas.
2	O nome do registo foi determinado por um taxonomista ou outra pessoa competente usando elementos de herbário, bibliografia e/ou material vivo documentado.
3	O nome da planta foi determinado por um taxonomista envolvido na revisão sistemática do grupo.
4	O registo é obtido através de colheitas ou de material tipo por métodos de reprodução assexuada.

Tabela 3. *Etiquetas de Nível de Verificação do HISPID (Conn 2000).*

Muitas instituições já têm um modelo para a certeza do registo usando termos como: "aff.", "cf.", "*s. lat.*", "*s. str.*", "?". No entanto alguns destes (aff., cf.) terem definições estritas, o seu uso por indivíduos diferentes pode variar consideravelmente. O uso *sensu stricto* e *senso lato* implica variações no conceito taxonómico.

Adicionalmente, quando os nomes provêem de outras pessoas que não o especialista taxonómico, pode-se listar a fonte de nomes usados (segundo Wiley 1981):

- descrição de novos taxa
- revisões taxonómicas
- classificações
- chaves taxonómicas
- estudos de faunísticos ou florísticos
- atlas
- catálogos
- checklists
- guias
- bolsas de estudo taxonómicas/ regras de nomenclatura
- análises filogenéticas

A incerteza pode ser normalmente reduzida e a qualidade melhorada através da comparação de duas ou mais publicações ou especialistas. As diferenças de identificações entre taxonomistas, no entanto, podem não implicar necessariamente que uma das identificações seja um erro, mas pode mostrar opiniões taxonómicas diferentes (e.g. diferentes hipóteses) em relação à posição do taxon.

3.2. Precisão na identificação

De acordo com Stribling *et al.* (2003), a precisão na identificação (que é erradamente designada de precisão taxonómica) pode ser avaliada comparando resultados de uma

amostra aleatória que será processada por dois taxonomistas ou especialistas. Também se pode fazer uma avaliação comparando os nomes dados a duplicados de espécimes pertencentes (e identificados) por diferentes instituições. Isto são noções abstratas e não sei o valor que tem registar este tipo de informação.

Uma segunda parte na identificação da precisão, é, no entanto, o nível até ao qual o espécime está identificado. Uma identificação até à espécie, ou subespécie é uma identificação mais precisa que uma até à família ou género. Ao documentar um conjunto de dados talvez seja importante, os utilizadores saberem que 50% das identificações.

3.3. Enviesamento

O enviesamento é um erro sistemático que surge no desvio uniforme de valores (Chrisman,1991). Normalmente surge da aplicação consistente de uma metodologia que leva a erros cuja natureza é sistemática. O enviesamento na nomenclatura taxonómica pode surgir quando uma identificação é precisa, mas não exata. Este enviesamento pode também surgir em más interpretações de uma chave dicotómica ou estrutura morfológica, no uso inválido de nomenclatura ou publicações desatualizadas (Stribling *et al.* 2003) (e.g. usar uma flora de outra área para o estudo e que pode não ter todos os taxa relevantes para a área em estudo).

3.4. Consistência

As inconsistências podem ocorrer dentro do domínio da classificação na base de dados se dois ou mais nomes forem considerados "aceites" para representar o mesmo taxon (e.g. *Eucalyptus eremaea* e *Corymbia eremaea*). Isto pode estar relacionado com a diferença de opiniões em relação à , ou erros devido a grafias diferentes (por exemplo *Tabernaemontana hystrix, Tabernaemontana histryx e Tabernaemontana histrix* – CRIA 2005).

3.5. Plenitude

Motro e Rakov (1998 in Dalcin 2004) referem-se à plenitude como "se todos os dados estão disponíveis" e dividem plenitude dos dados em *plenitude do ficheiro* (nenhum registo em falta) e *plenitude do registo* (todos os campos são conhecidos para cada registo).

A plenitude em termos taxonómicos (e.g. com uma base de dados de nomes ou taxon) refere-se à cobertura de nomes. A base de dados inclui nomes de todos os níveis da hierarquia (e.g. até à subespécie ou só espécie)? Que porção do reino animal ou das plantas é coberta pela base de dados? A base de dados inclui sinónimos? Todas estas questões são importantes para o utilizador avaliar a aptidão para o uso dos dados para si. Dalcin (2004), por exemplo, divide plenitude em *plenitude de nomenclatura*, que representa a inclusão de todos os nomes possíveis, tendo em conta o contexto (e.g. no contexto taxonómico – uma lista de todos os nomes para um grupo taxonómico especifico; ou num contexto espacial – uma lista de todos os nomes de espécies para aquela região) e em *plenitude na classificação* que representa todos os nomes possíveis relacionados com o nome "aceite" para um dado taxon (e.g. um sinónimo completo).

Com uma base de dados de espécimes ou observações, a plenitude pode ser no sentido de " todos os campos *Darwin Core* estão incluídos" e "todos os campos do *Darwin Core* têm dados". Em base de dados de caracteres, "estão presentes caracteres para todas as fases da vida" (e.g. frutos das plantas, fases nos insetos).

3.6. Coleções de espécimes

A importância das coleções de espécimes não pode ser deixada de parte, no entanto nem sempre é possível inclui-las nas bases de dados. Muitas bases de dados de observação são construídas sem se fazer ao mesmo tempo coleções espécimes. Nem sempre é possível em determinados casos ou áreas recolher uma amostra para fazer um *voucher*, por questões políticas, legais, de conservação ou outras situações.

Quando é possível fazer espécimes *vouchers* é um exercício valioso na fase inicial de programas baseados em espécies, para desenvolver acordos de cooperação entre coletores de dados e instituições como museus ou herbários que possam dar ao depósito das coleções de espécimes de referência (Brigham 1998). Estes acordos devem também estender-se ao arquivamento apropriado e estratégias de disponibilização, incluindo tempo mínimo antes de disponibilização ou arquivamento.

4. Dados espaciais

Os dados espaciais têm frequentemente liderado a área de desenvolvimento padrões para documentação de dados (por exemplo com o desenvolvimento do *Spatial Data Transfer Standards* (USGS 2004), o programa[22] INSPIRE (Informação de informação espacial na Europa) e desde então têm estado na vanguarda do desenvolvimento de padrões para a qualidade de dados (e.g. ISO 19115 para Informação Geográfica – Metadados[23]). A natureza numérica da maioria dos dados espaciais torna-os mais aptos ao uso nos procedimentos estatísticos do que os dados taxonómicos, permitindo assim o desenvolvimento de uma série de métodos de verificação da qualidade de dados (veja o documento anexo *Princípios e métodos de limpeza de dados*).

Isto não quer dizer que todas as partes espaciais dos dados (o "domínio dos campos dos dados" " de Dalcin 2004) sejam fáceis de digitalizar ou sejam exatos. Muitas coleções históricas em museus e herbários só têm descrições textuais muito simples da localidade da colheita e isto leva a um enorme esforço para os converter em elementos georreferenciados ou em coordenadas. Isto pode ser agravado pela natureza de algumas coleções, por exemplo, coleções recolhidas numa época em que os mapas detalhados não estavam disponíveis para os coletores e onde os nomes das localidades já não são atualmente usados em índices toponímicos ou mapas. A adição aos registos históricos de dados georreferenciados, especialmente quando não existem bons índices toponímicos, pode levar demasiado tempo e resultar em baixos níveis de exatidão.

Foram desenvolvidas muitas ferramentas para ajudar os utilizadores a georreferenciar os seus dados, incluindo ferramentas e guias *on-line*. Estas serão melhor abordadas no documento associado, *Princípios e métodos de limpeza de dados*. Adicionalmente, a maioria dos coletores usam atualmente GPS (Sistema Global de Posicionamento) para georreferenciar na altura da recolha. Para uma discussão sobre a exatidão associada com o uso do GPS veja o capítulo "*Recolha de dados*".

Os testes de erros associados a georreferenciações já atribuídas envolvem:

- verificação contra informação interna ao próprio registo ou entre registos ao longo da base de dados – por exemplo estado, nome do distrito, etc.;
- verificação contra uma referência externa usando uma base de dados – o registo é consistente com as localidades de recolha do coletor?
- verificação contra uma referência externa usando um SIG – o registo está em terra em vez de no mar?
- verificação dos anómalos no espaço geográfico; ou
- verificação dos anómalos no espaço ambiental.

Todos estes métodos serão mais desenvolvidos no documento anexo, *Princípios e métodos de limpeza de dados*.

[22] http:// http://inspire.ec.europa.eu/
[23] http://www.iso.org/iso/catalogue_detail?csnumber=26020

4.1. Exatidão espacial

Como é medida a exatidão posicional de dados espaciais?

Para a maioria das camadas SIG (mapas topográficos, etc.) a fonte de "verdade" é relativamente fácil de determinar uma vez que normalmente existem fontes externas de elevada exatidão em algumas propriedades na base de dados – pontos de levantamento trigonométricos, estradas e interseção de estradas, etc. (Chrisman 1991). Muitos dos testes, no entanto, não são simples e a documentação – como no *US National Map Accuracy Standard* - complicada. Tradicionalmente, a exatidão espacial é determinada por comparação a alguns pontos "bem definidos" juntamente com níveis de erros especificados aceitáveis, como a média da raiz quadrada do desvio de zero (RMSE) uma determinada exatidão (Chrisman 1991). A RMSE não é fácil de aplicar a pontos individuais, no entanto, é mais aplicável à totalidade de um conjunto de dados ou mapas digitais. Com pontos individuais a distância até à localização verdadeira pode ser obtida usando métodos simples como o método do raio do ponto (Wieczorek *et al*. 2004) ou métodos similares são fáceis de usar. Existem dois fatores envolvidos – como é que a exatidão do ponto bem definido pode ser determinada quando se determina a exatidão do ponto a testar, o que é a exatidão e precisão da medição do ponto a testar vai adicionar ao erro. Por exemplo, se uma interseção na estrada só pode ser exata a menos de 100 metros, então o centróide do ponto de colheita é um círculo de 100 metros antes de adicionar a precisão do ponto. (veja comentários em Wieczorek 2001).

O *US Federal Geographic Data Committee* (FGDC))lançou os padrões de exatidão para o posicionamento geoespacial (GPAS) em 1998. Estes padrões incluem secções separadas para redes geodésicas e para exatidão de dados espaciais (FGDC 1998).

- "*O NSSDA usa a média da raiz quadrada do erro (RMSE) para estimar a exatidão posicional. O RMSE é a raiz quadrada da média do quadrado da diferença entre os valores das coordenadas do conjunto de dados e os valores de coordenada de uma fonte independente de elevada exatidão para pontos idênticos.*"
- "*A exatidão é reportada em distâncias no solo a um nível de 95% de confiança. A exatidão reportada a um nível de 95% de confiança significa que 95% das posições do conjunto de dados terão um erro no que respeita à posição real no solo que será igual ou menor que o valor de exatidão atribuído. O valor de exatidão atribuído reflete todas as incertezas, incluindo aquelas introduzidas pelas coordenadas geodésicas, compilação e o cálculo final das coordenadas no solo.*"

Exemplos da exatidão de mapas realizadas na Austrália usando este método, tendo em conta o produto, são:

- "*A exatidão média deste mapa é de ±100 metros na posição horizontal de detalhes bem definidos e de ±20 metros na altitude*" (Divisão Nacional de Mapas, Sheet SD52-14, Edition 1, 1:250,000).

Estas exatidões precisam de ser adicionadas a qualquer georreferenciação de uma coleção baseadas num mapa em papel ou digital. Como existe sempre incerteza na exatidão de dados espaciais, não pode ser aplicada nenhuma indicação absoluta acerca da exatidão, mas é importante que a exatidão conhecida esteja documentada. Os erros são propagados através da cadeia de informação e contribuem para as incertezas no resultado final, seja

um mapa resultado de um SIG ou um modelo de espécie usando um *software* de modelação de distribuição (Heuvelink 1998).

4.2. Projeto BioGeomancer

O projeto[24] foi recentemente financiado pela Fundação Gordon e Betty Moore para ajudar a melhorar a georreferenciação dos registos primários de espécies e avaliar, melhorar e documentar a exatidão. Este projeto deve reportar e tornar disponíveis as ferramentas desenvolvidas em 2006.

4.3. Falsa precisão e exatidão

Um fator adicional a ter em conta é a Falsa Precisão e Exatidão. Muitos utilizadores SIG não estão cientes de todas as questões que a exatidão de dados espaciais implica e assumem que os seus dados são absolutos. Normalmente, reportam níveis de exatidão impossíveis para aquele tipo de fonte de dados. Muitas instituições usam agora SIG para ajudar na georreferenciação, fazendo aproximações a níveis não suportados pelos dados (e usando casas decimais), terminando com uma precisão pouco realista. Também, com o uso de um registo GPS a localização do evento da colheita não é muitas vezes reportado para 1 ou 2 metros, quando na realidade ao usar vários aparelhos GPS de mão provavelmente possuem uma precisão de cerca de 10 metros ou menos, isto é particularmente relevante quando se usa o GPS para se determinar a altitude (veja comentários em "*Recolha de dados*", abaixo).

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

5. Coletor e dados de colheita

A informação acerca do coletor e da colheita (domínio dos dados da colheita de Dalcin 2004) inclui informação acerca da própria colheita – o coletor, colheita e informação adicional como habitat, solo, condições climáticas, experiência do observador, etc. Podem ser categorizadas como (modificado de Conn 1996, 2000):

- Autor(es) da colheita e número de coletor(es)
- Experiência dos observadores, etc.
- Período(s) / data da colheita
- Método de colheita (particularidades dos dados de observação/amostragem)
- Dados associados

Muitas destas questões variam consideravelmente de acordo com o tipo de dados que serão recolhidos – sejam para uma coleção num museu, uma observação ou resultados de uma pesquisa detalhada. Para uma coleção estática como num museu, o nome, número do coletor e a data são atributos chave, com dados associados como hábito, habitat, etc. e talvez o método de captura (animais). Para dados de observação, coisas como duração da observação, área amostrada pela observação, hora do dia (hora de inicio e fim da observação, além da data), e dados associados como as condições climáticas, sexo do animal observado, atividade, etc., são importantes. Para dados de pesquisa, informação sobre os métodos dessa pesquisa, tamanho (grelha e área total), esforço, condições climáticas, frequência, indicação de quando há recolha de espécies *voucher* e os seus números, etc. em conjunto com os dados referidos para observações.

5.1. Exatidão do atributo

Os problemas que podem entrar em conflito com a qualidade de dados no que respeita aos dados de colheita, incluem o modo como o nome do coletor, número, iniciais, etc. são inseridos (Koch 2003), a exatidão no registo da data e horas, a consistência do registo de dados no momento da colheita, como hábito, habitat, solo, tipo de vegetação, cor da flor, sexo, espécies associadas.

Um exemplo de problemas que normalmente aparecem com dados de colheita é o "número de coletor", pois alguns coletores não usam números únicos para identificar as suas colheitas. Isto pode causar perda de qualidade pois esses números são muitas vezes usados para identificar a localização da colheita, identificações, colheitas duplicadas em diferentes instituições, etc.

5.2. Consistência

A consistência no uso de terminologia em relação ao domínio da colheita é bastante irregular, e é raro que campos de dados associados, em particular, sejam consistentes ao longo de uma base de dados, e muito menos em base de dados diferentes.

5.3. Plenitude

A plenitude da informação de uma coleção é usualmente muito variável. É frequente que o habitat, número do coletor, época de floração, etc. não estejam preenchidos em muitos

registos. Isto torna um estudo de habitat, por exemplo, difícil a partir de uma única coleção.

6. Dados decritivos

O uso de base de dados descritivos está a aumentar tanto para o armazenamento de dados como métodos de publicação, substituindo muitas vezes as publicações tradicionais. Dados morfológicos, fisiológicos e fenológicos são exemplos de dados neste domínio. Dados descritivos são muitas vezes usados para gerar informação para uso em análises cladísticas e descrições geradas automaticamente por ferramentas de identificação.

O *Taxonomic Databases Working Group* (TDWG) tem uma longa história no desenvolvimento e promoção de padrões na área das bases de dados descritivas – primeiro com o seu suporte ao padrão DELTA (Dallwitz e Paine 1986) e mais recentemente com o desenvolvimento do grupo de trabalho "Estrutura dos Dados Descritivos"[25].

A qualidade dos dados descritivos pode ser variável, embora os elementos dos dados sejam usualmente medidos, na realidade a exatidão pode ser determinada por casos onde os dados não são observáveis (e.g. com dados históricos), ou não são fáceis de observar (e.g. demasiado dispendiosos) e/ou inferidos em vez de reais (e.g. avaliação subjetiva como cor, abundância, etc.).

Na maioria dos casos, os dados descritivos são arquivados ao nível da espécie em vez de ao nível do espécime sendo usualmente em média ou em amplitude. Como foi referido por Morse (1974 como referenciado por Dalcin 2004), as informações taxonómicas têm inerente um nível mais baixo de confiabilidade do que dados de observação do espécime. Independentemente disto, existe recentemente uma grande tendência em armazenar, pelo menos alguns destes dados, ao nível do espécime aumentando assim a qualidade.

A qualidade dos dados descritivos pode ser variável, embora os elementos dos dados sejam usualmente medidos, na realidade a exatidão pode ser determinada por casos onde os dados não são observáveis (e.g. com dados históricos), ou não são fáceis de observar (e.g. demasiado dispendiosos) e/ou inferidos em vez de reais (e.g. avaliação subjetiva como cor, abundância, etc.).

Na maioria dos casos, os dados descritivos são arquivados ao nível da espécie em vez de ao nível do espécime sendo usualmente em média ou em amplitude. Como foi referido por Morse (1974 como referenciado por Dalcin 2004), as informações taxonómicas têm inerente um nível mais baixo de confiabilidade do que dados de observação do espécime. Independentemente disto, existe recentemente uma grande tendência em armazenar, pelo menos alguns destes dados, ao nível do espécime aumentando assim a qualidade.

6.1. Plenitude

Ao nível do espécime, a plenitude dos registos de dados descritivos pode depender da qualidade do espécime, época do ano, etc. Por exemplo, pode não ser possível registar características do fruto ou flores do mesmo espécime. Por esta razão, muitos campos essenciais vão ser deixados em branco. Noutros casos, o atributo pode não ser relevante para a caracterização e portanto nem todos os atributos serão registados.

[25] http://www.tdwg.org/standards/116/

6.2. Consistência

Os problemas de inconsistência podem surgir entre dois itens de dados relacionados. Por exemplo, no descritor caraterísticas, duas espécies podem ser registadas como (Dalcin 2004):

- "HÁBITO=HERBACIA" e
- "USOS=MADEIRA"

Inconsistências na representação do mesmo atributo podem também afetar a qualidade, especialmente onde são utilizadas definições pobres, do atributo ou os padrões consistentes não são rigidamente obedecidos. Por exemplo (Dalcin 2004):

- "COR DA FLOR= CARMIM", e
- "COR DA FLOR=CARMESIM".

O uso de terminologias padrão pode ajudar a reduzir consideravelmente o grau de erro e más interpretações. Estas terminologias estão a ser desenvolvidas em diferentes áreas e disciplinas e a recente alteração para o desenvolvimento de bases de dados descritivas, aumentou a consistência com a qual as terminologias são usadas. O desenvolvimento dos padrões TDWG para a estrutura dos dados descritivos (EDD) (TDWG 2005) só pode auxiliar este processo.

7. Colheita de dados

Existem diversas maneiras de colheita de dados primários de espécies e dados de ocorrência de espécies, cada um com os seus níveis de precisão e exatidão, bem como com as suas fontes de erro e incerteza. Cada um deles têm diferentes impactos na "aptidão para o uso" final ou qualidade, dos dados. Muitos dos métodos usados para dados de espécies serão brevemente discutidos.

7.1. Oportunista

A maioria dos dados de ocorrência de espécies foram recolhidos de forma oportunista. Muitos destes registos estão agora armazenados como espécimes em museus e herbários. A maioria dos dados históricos só incluem uma referência textual à localização, como a 5Km NW de uma cidade, etc. e raramente são georreferenciados no momento da recolha. A georreferenciação é normalmente realizada depois e usualmente por alguém que não o coletor (Chapman e Busby 1994). Muitos registos observacionais (dados de atlas de aves, etc.) também foram recolhidos oportunistamente.

Estes dados são normalmente registados digitalmente no formato de lotes, e a georreferenciação normalmente é feita usando como referência mapas físicos. Normalmente incluem ambos baixa precisão e exatidão. A maioria destes dados não podem ter mais do que 2-10 km de exatidão.

7.2. Amostragem de campo

Dados de amostragem de campo incluem geralmente uma referência espacial, normalmente na forma de latitude, longitude ou referência UTM. A referência espacial pode normalmente ser considerada como tendo cerca de 100–250 metros de exatidão. Devem ser tomados cuidados, no entanto, ao quê que se refere essa referência espacial – pode não se referir à localização real da observação, mas sim, por exemplo, ao ponto médio do transeto, ou ao canto (ou centro) de um quadrado de uma grelha e nem sempre isto é claro. Além disso, como os registos raramente têm associados espécies *voucher* (e.g. uma coleção física construída e armazenada para referência posterior) a exatidão taxonómica não pode ser sempre invocada. Isto é particularmente verdade quanto mais tempo passa desde a recolha, podendo os conceitos taxonómicos já terem sido alterados.

7.3. Observações de longa escala

Alguns estudos biológicos só recolhem dados de uma determinada área ou célula de uma grelha. Por exemplo, um estudo das espécies num parque nacional, ou observação de aves feitas com uma grelha de quadrados de 10-minutos (e.g. Aves da Austrália 2001, 2003). A exatidão deste tipo de registo só pode ser na ordem dos 1-10 km ou maior.

7.4. Sistemas de Posicionamento Globais (GPS)

Os sistemas de posicionamento globais, ou GPSs estão cada vez mais presentes na colheita de dados de espécies. Isto inclui não só dados de pesquisa, mas também recolhas oportunistas e de observação.

A tecnologia GPS usa a triangulação para determinar a localização de uma posição na superfície terrestre. A distância medida é o intervalo entre o recetor GPS e os satélites GPS (Van Sickle 1996). Como a localização no espaço dos satélites GPS é conhecida, a posição na terra pode ser calculada. São requeridos no mínimo 4 satélites GPS para determinar a localização de uma posição na superfície terrestre (McElroy *et al*. 1998, Van Sickle 1996). Hoje, isto não é uma limitação, pois podemos receber informação de até 7 ou mais satélites na maior parte das localizações na terra, no entanto historicamente, o número de satélites dos quais se podia receber informação não era suficiente. Antes de Maio de 2000, a maioria das unidades GPS usadas por civis envolviam "disponibilidade seletiva". A sua remoção permitiu uma enorme melhoria na exatidão que pode ser esperada (NOAA 2002).

Antes da remoção da disponibilidade seletiva, a exatidão dos recetores GPS portáteis usados pela maioria dos biólogos e observadores no campo, era da ordem dos 100 metros ou pior (McElroy *et al*. 1998, Van Sickle, 1996, Leick 1995). A partir daí, no entanto, a exatidão dos recetores de GPS melhorou e hoje, a maior parte dos GPS portáteis produzidos prometem erros de menos de 10 metros em áreas abertas quando se usa 4 ou mais satélites. A exatidão pode ser melhorada averiguando os resultados de múltiplas observações numa única localização (McElroy *et al*. 1998), e alguns recetores modernos de GPS que incluem médias de algoritmos podem diminuir a exatidão para cerca de 5 metros ou talvez ainda melhor.

O uso de *GPS diferencial* (DGPS) pode melhorar consideravelmente a exatidão. A DGPS usa referenciação de uma estação de base de GPS (usualmente num ponto de controlo do estudo) com uma localização conhecida para calibrar a receção do GPS. Este funciona entre a estação de base e o GPS portátil que fazem referenciação por satélite ao mesmo tempo da posição, reduzindo assim os erros causados pelas condições atmosféricas. Desta forma os GPS portáteis aplicam a correção apropriada para a posição determinada. Dependendo da qualidade dos recetores usados, pode-se esperar uma exatidão entre 1 a 5 metros. Esta exatidão diminui quando a distância do GPS à estação base aumenta. Mais uma vez realizar médias pode melhorar estes valores (McElroy *et al*. 1998).

O *Wide Area Augmentation System* (WAAS) é um sistema de GPS baseado na Navegação e aterragem[26], desenvolvido para pilotar com precisão aeronáutica (Federal Aviation Administration 2004). A WAAS envolve uma antena terrestre com localização precisa e conhecida, podendo providenciar uma posição de grande exatidão com o uso do GPS. Também foram desenvolvidas tecnologias similares tais como a *Local Area AugmentatioSystem* (LAAS) para dar uma precisão ainda mais fina.

Grandes exatidões podem ser recebidas usando tanto o GPS diferencial em tempo real (McElroy *et al*. 1998) como o GPS estático (McElroy *et al*. 1998, Van Sickle 1996). O GPS estático usa instrumentos de elevada precisão e técnicos especializados sendo geralmente usados por pesquisadores. Estudos realizados na Austrália usando estas técnicas reportaram exatidões na faixa dos centímetros.

[26] No Brasil, aterrisagem.

Estas técnicas não são para usar extensivamente para colheita de registos biológicos devido ao custo e ausência de requisitos para estas precisões. Para obter exatidões como as descritas acima, o recetor de GPS tem de estar localizado numa área livre de obstruções e de superfícies refletoras e que tenha um bom campo de visão para o horizonte (por exemplo, não trabalham muito bem debaixo de uma floresta com uma canópia densa). Os recetores de GPS têm de ser capazes de registar sinais de pelo menos 4 satélites de GPS num arranjo geométrico. A melhor solução é ter "*um satélite diretamente acima e os outros três igualmente espaçados à volta do horizonte*" (McElroy *et al*. 1998). O recetor de GPS tem também de estar configurados para um datum apropriado à área e o datum utilizado de ser registado.

Altitude GPS. A maior parte dos biólogos sabe pouco sobre a determinação da altitude usando um GPS. É importante ter em conta que a altitude dada pelo recetor de GPS é na verdade a altitude em relação ao datum central da terra (e está assim relacionado com a estrutura elipsoidal da terra) e não a uma altitude relacionada com o nível médio do mar ou com um datum padrão de altitude como o datum de altitude da Austrália. Na Austrália, por exemplo, a diferença entre a altitude dada por um recetor de GPS e o nível médio do mar pode varia de -35 a +80 metros e tende a variar de uma maneira imprevisível (McElroy *et al*. 1998, Van Sickle 1996).

8. Entrada e Aquisição de dados

(Recolha de dados eletronicamente)

> "*A aquisição e recolha de dados é inerentemente sujeita a erros tanto simples como complexos*" (Maletic e Marcus 2000).

8.1. Captura básica de dados

O primeiro passo na captura de dados é usualmente a recolha de informação a partir da etiqueta do espécime, revista científica ou caderno de campo, livro de registos ou ficheiro em papel. Isto pode ser feito através de técnicos de entrada de dados especializados ou não, ou através da digitalização eletrónica da informação. O nível de erros devido à entrada dos dados pode ser sempre reduzido através da dupla digitalização, usando software de aprendizagem e treino associado à digitalização e usando peritos e supervisores para levarem a cabo testes às entradas ou a uma amostra base (veja o guia MaPSTeDI, mencionado abaixo).

8.2. Interface do utilizador

O desenvolvimento de uma interface para um utilizador específico de entrada de dados também pode ser uma forma de diminuir os erros na entrada de dados. Muitas instituições usam pessoas não qualificadas ou voluntários como técnicos de entrada de dados desenvolvendo uma interface simples (não técnica) do utilizador onde os técnicos se sintam confortáveis e com a qual possa aumentar a exatidão de inserção de dados. Este tipo de interface pode ajudar na inserção de dados, pois é capaz de rapidamente procurar campos de preenchimento obrigatório, entradas existentes na base de dados, outras bases de dados relacionadas e até usar motores de busca como o Google que podem ajudar o operador a decidir na inserção correta de um nome ou terminologia onde possa ter dificuldade na leitura da etiqueta, ou a determinar o que deve ou não ir para um determinado campo. Em alguns casos isto pode ser aplicado ao longo da construção da base de dados incorporando tabelas autoritárias ou menus de seleção impedindo os técnicos de terem de tomar decisões acerca de nomes, localidades ou habitats.

8.3. Georreferenciação

Os mapas são uma das formas mais efetivas de comunicar informação o que, por si só, justifica o recente aumento da inserção em base de dados e da georreferenciação dos dados de espécimes de museus e herbários, juntamente com o aumento da captura de informação de observação já georreferenciada. A capacidade de melhorar os dados com mapas permite-nos um melhor estudo, identificação, visualização, documentação e correção de erros e de incertezas nos dados (Spear *et al*. 1996). Também proporciona um método poderoso para visualizar e comunicar incertezas em relação aos dados, e permitir aos utilizadores no presente determinar a qualidade ou a aptidão para o uso dos dados.

A captura eletrónica de dados e a anexação de informação geográfica (e.g. georreferenciar os dados) pode ser uma tarefa difícil e consumir muito tempo. Os resultados do projeto MaPSTeDI (University of Colorado 2003) sugerem que um técnico

competente pode georreferenciar um registo a cada 5 minutos. Outros estudos (Armstrong 1992, Wieczorek 2002) mostraram que a georreferenciação pode demorar bastante mais – por exemplo a base de dados MANIS sugere uma taxa de 9 por hora para os Estados Unidos, 6 por hora para não norte americanos dos Estados Unidos e 3 por hora para localidades não norte americanas (Wieczorek 2002).

MaNIS/HerpNet/ORNIS
Georeferencing Guidelines - http://manisnet.org/manis/GeorefGuide.html
MaPSTeDI
Georeferencing in MaPSTeDI - http://mapstedi.colorado.edu/geo-referencing.html

Foram desenvolvidos um conjunto de métodos e guias excelentes instruções de assistência aos gestores de dados durante a georreferenciação. As instruções de georreferenciação desenvolvidas por John Wieczorek do *Museum of Vertebrate Zoology* em Berkeley (Wieczorek 2001) e pelo MaPSTeDI (*Mountains and Plains Spatio-Temporal Database Informatics Initiative*), (University of Colorado 2003) são dois dos estudos mais abrangentes na área até à data e cujas orientações vos aconselho a ler. Estes guias cobrem a determinação da exatidão e precisão de um ponto derivado de uma localização textual, incertezas provenientes do uso de diferentes datums, efeitos do uso de diferentes escalas de mapas, etc. São compilações abrangentes sobre o tema e espero que os leitores deste documento possam considerá-las complementares a este documento.

Existe também uma série de ferramentas *on-line* que podem ajudar na determinação dos dados geográficos – por exemplo para lugares a uma dada distância e direção a partir de uma localidade já conhecida. Este tema será mais desenvolvido no documento associado *Princípios e Métodos de Limpeza de Dados*.

geoLoc - Reference Centre for Environmental Information -
http://splink.cria.org.br/tools/

8.4. Erro

Ferramentas como as mencionadas anteriormente são ferramentas poderosas para reduzir o erro e aumentar a qualidade. Mas nenhum método de georreferenciação consegue eliminar totalmente o erro. Como referido nas instruções do MaPSTeDI:

> "*Dado que a georreferenciação não é uma ciência exata e nenhuma coleção pode ser georreferenciada correctamente a 100% , a verificação da qualidade pode aumentar drasticamente a percentagem de coleções georreferenciadas corretamente. Todos os projetos devem ter isto em conta quando planeiam a sua georreferenciação*" (University of Colorado 2003).

Uma fonte de erros na georreferenciação é o uso indiscriminado de índices toponímicos eletrónicos. Em alguns casos estes índices foram desenvolvidos através de projetos para a publicação de mapas em papel, e as localizações dos pontos dados pelos índices são do canto inferior esquerdo de onde o nome será escrito no mapa e não da localização do ponto a que se refere (e.g. o índice toponímico anterior a 1998 desenvolvido pelo grupo *Australian Land Information*). Com sorte, a maioria dos índices foram corrigidos, mas podem existir dados georreferenciados em bases de dados de museus e herbários,

baseados nestes valores. A exatidão deste tipo de registos devem ser verificados aleatoriamente nas localidades contrapondo os dados do índice com um mapa preciso de grande escala.

Muitas vezes é mais rápido e mais eficiente realizar georreferenciação como uma atividade separada e após a digitalização da informação da etiqueta. Isto permite que a base de dados seja usada para procurar colheitas por localidade, coletor, data, etc. e permite um uso mais eficiente de mapas para obter a informação geográfica. Isto também ajuda a evitar a ocorrência de duplicação de georreferenciação de múltiplos registos para a mesma localidade, etc.

9. Documentar dados

> "*Os metadados são dados acerca dos dados. São uma descrição das características dos dados que foram recolhidos para um propósito específico*" (ANZLIC 1996a).

A boa documentação dos dados ocorre tanto ao nível da base de dados como ao nível do registo dos dados.

Os metadados fornecem informação acerca de um conjunto de dados tais como conteúdo, extensão, acessibilidade, exatidão, plenitude, aptidão para o propósito e aptidão para o uso. Quando os metadados são fornecidos, o utilizador pode ganhar uma maior compreensão da qualidade da base de dados e determinar a adequação da base de dados antes de a utilizar. Bons metadados permitem uma melhor troca, pesquisa e recuperação dos dados. Os metadados normalmente referem-se a todo o conjunto de dados, no entanto alguma documentação pode ser vista ao nível do registo (como seja o registo da exatidão) como sendo metadados ao nível do registo. Independentemente do nome que lhes possamos dar, uma boa documentação, tanto ao nível do conjunto de dados como ao nível do registo é importante.

Todos os dados incluem erros – não há como escapar a isso! O importante é saber o que o erro é, e saber se este erro está dentro dos limites aceitáveis para o uso pretendido dos dados. É aqui que os metadados se revelam ainda mais importantes para a base de dados como um todo, tornando-se assim pertinente na área do desenvolvimento dos metadados a definição de "aptidão para uso". O conceito de aptidão para o uso não ficou totalmente reconhecido na área da informação geográfica até ao início dos anos 90 e não o foi até meados dos anos 90 quando começou a aparecer na literatura da área (Agumya e Hunter 1996).

O registo de informação só ao nível do conjunto de dados, no entanto, nem sempre fornece a informação que o utilizador necessita. Registar erros ao nível do registo, especialmente com dados de espécies, pode ser extremamente importante para determinar a aptidão desse registo para o uso. Quando esta informação está disponível, um utilizador pode pedir, por exemplo, apenas os dados que sejam melhor que um certo valor métrico - e.g. melhor que 5000 metros. É também importante que as ferramentas automáticas de georreferenciação incluam o cálculo de exatidão com um campo no resultado final.

É também importante que os utilizadores dos dados percebam o conceito de aptidão para o uso. Muitas vezes os dados de ocorrência de espécies são extraídos de uma base de dados num formato "registo n.º, x,y" sem que tenha associado nenhum valor de exatidão. As próprias coordenadas representam sempre um ponto, mas raramente se refere ao ponto verdadeiro. Alguns registos foram introduzidos na base de dados com um ponto arbitrário (por exemplo uma recolha que só tenha "América do Sul" na etiqueta) e lhe atribuem uma exatidão de 5000000 metros no campo da exatidão. Existem algumas bases de dados que o fazem! Extrair o registo e usar este ponto arbitrário será extremamente enganoso. Os utilizadores precisam de estar cientes de que existe um campo para a exatidão caso esteja presente, e ser aconselhado sobre o uso. Em casos onde os fornecedores de dados

desenvolvem relatórios de dados padrão, devem tornar o campo da exatidão obrigatório quando o dado é inserido

Os dados devem ser documentados com metadados suficientemente detalhados para permitir o uso por terceiros sem referência à origem dos dados.

Documentar a exatidão, precisão e erros em dados espaciais é essencial para que os utilizadores possam ser capazes de determinar a qualidade destes dados para o seu objetivo de uso.

Esta documentação deve incluir (no mínimo):

- título do conjunto de dados
- fonte dos dados
- linhagem dos dados (operações realizadas nos dados desde a sua recolha ou derivação)
- exatidão (posicional, temporal e atributo)
- consistência lógica
- data e expetativa de durabilidade dos dados (exatidão dos dados e estado, frequência de atualização)
- definições dos campos de dados
- metodologia de colheita
- plenitude
- condições e restrições ao uso (e.g. direitos de autor, licença de restrições, etc.)
- custódia dos dados e informação de contacto

Vale a pena definir alguns destes termos pois nem todos os detentores de dados estão cientes deles. Muitos destes termos referem-se a uma coleção de dados numa base de dados em vez dos registos de colheita individualmente.

9.1. Exatidão posicional

A exatidão posicional refere-se a quão perto a descrição das coordenadas do recurso se compara à localização real (Minnesota Planning 1999). Quando é possível e conhecido, o Datum Geodésico usado para determinar as coordenadas da posição.

É também recomendável que as bases de dados incluam um campo para o valor da exatidão posicional de cada registo individualmente. Existem várias maneiras de o fazer. Algumas bases de dados usam códigos, no entanto, é preferível que seja um simples valor métrico que seja usado para estimar a exatidão do registo (Chapman e Busby 1994, Conn 1996, 2000, Wieczorek *et al*. 2004). Isto pode ser importante para os utilizadores extraírem os dados para um propósito particular – por exemplo, eles podem só querer os dados cuja exatidão seja melhor que 2000 metros. Algumas vezes, pode ser também importante incluir um campo ao nível do registo de como a informação geográfica foi determinada. Por exemplo:

- uso de GPS diferencial
- GPS portátil afetados pela disponibilidade seletiva (e.g. antes de 2002)

- Um mapa de referência de 1:100 000 e obtido por triangulação usando recursos facilmente reconhecíveis.
- Referência de mapa usando uma conta inoperável
- Referência de mapa obtida remotamente (e.g. num helicóptero)
- Obtido automaticamente usando *software* de georreferenciação através do método ponto-raio.
- Uso de índices toponímicos incluindo nome, data e a versão do índice.

9.2. Exatidão do atributo

A exatidão do atributo refere-se à avaliação de quão corretas e fiáveis são descritas as características dos dados em relação aos valores na realidade. Idealmente, deveria incluir uma lista de atributos e informação para cada um. Por exemplo:

> "*Os registos são fornecidos por observadores experientes. Obteve-se uma exatidão adicional testando a correção dos atributos contra espécimes voucher depositados num museu ou herbário para verificação por peritos. Aproximadamente 40% dos registos de plantas estão verificados com espécies voucher, 51% anfíbios, 12% mamíferos, 18% répteis e 1% para aves*" (SA Dept. Env. & Planning 2002).

9.3. Linhagem

A linhagem refere-se à fonte dos dados, juntamente com os processos/alterações realizados na base de dados até ao estado atual. Pode incluir o método de recolha (e.g. "dados recolhidos numa grelha de 10 X 10 metros") e a informação dos testes de validação que foram realizados nos dados. A história das etapas dos processos podem incluir:

- os método(s) de captura dos dados
- quaisquer passos e métodos intermédios
- os métodos usados para gerar o produto final
- qualquer passo de validação realizado nos dados.

Por exemplo:

> "*Os dados foram obtidos usando quadrados fixos de 20 metros x 20 metros. Foram efetuadas contagens de espécies, e também foram recolhidos dados acerca da estrutura e outras características do habitat. Os dados foram classificados usando o Twinspan por grupos compostos de espécies semelhantes*".

9.4. Consistência lógica

A consistência lógica proporciona uma breve avaliação das relações lógicas entre itens nos dados. Embora para a maioria dos dados recolhidos aqui (dados de museus e herbários) alguns dos itens podem não ser relevantes, no entanto podem-no ser para dados de observação (*checklists* de espécies de um parque nacional ou bioregião, etc.) e alguns dados de investigação. Para dados espaciais onde os dados são arquivados digitalmente, os testes de consistência lógica podem ser realizados automaticamente. Coisas como:

- Todos os pontos, linhas e polígonos têm legenda e algum tem legenda em duplicado?
- As linhas intersetam-se nos nós ou cruzam-se involuntariamente?

- Os limites dos polígonos estão fechados?
- Todos os pontos, linhas e polígonos estão topologicamente relacionados?

A consistência lógica pode também ser aplicada no caso de conjuntos de dados onde haja outra relação lógica entre os itens e objetos nos dados. Nestes casos deve ser incluído uma descrição de qualquer teste sobre as relações que tenha sido feito. Podem ser exemplos datas que ocorram em diferentes campos – se uma data dada num campo diz que o projeto foi realizado entre os anos "a" e "b" mas a data de registo de um atributo noutro campo está fora desse intervalo, então é logicamente inconsistente; ou registos que estão fora do intervalo geográfico – se num campo se regista que os dados foram recolhidos no Brasil e outro campo inclui registos de latitudes e longitudes para o Paraguai, então há inconsistências lógicas entre os dois campos. A documentação sobre verificações realizadas é uma parte importante dos metadados. As verificações podem incluir testes como "pontos no polígono" que são utilizados para este propósito no mundo SIG. Veja um desenvolvimento dos métodos no artigo associado *Princípios e Métodos de Limpeza de Dados*.

9.5. Plenitude

A plenitude refere-se tanto à cobertura temporal e espacial dos dados ou conjunto de dados como uma porção da extensão total possível dos mesmos. A documentação acerca da plenitude é uma componente essencial para determinar a qualidade dos dados. Os exemplos podem incluir:

> "*Completa para áreas a norte dos 30ºS, registos esparsos só entre 30º e 40º S.*"
>
> "*Este conjunto de dados abrange apenas registos recolhidos de modo oportunista antes de 1995, na sua maioria da Nova Gales do Sul, mas inclui alguns registos de outros estados.*"

Da perspetiva do utilizador, a plenitude está relacionada com "todos os dados de que precisa" (English 1999). Isto é, o utilizador precisa de saber se a base de dados inclui todos os campos de que necessita para a sua análise e precisa de saber a "plenitude" de alguns desses campos. Por exemplo, o utilizador pode querer fazer um estudo comparativo de atributos ao longo do tempo, mas se a base de dados só incluir dados até um determinado ano, pode não ser utilizável para a sua análise (veja o segundo exemplo acima).

9.6. Acessibilidade

Para os dados terem valor para os utilizadores necessitam de estar acessíveis. Nem todos os dados estão acessíveis on-line e para ter acesso a alguns utilizador pode ter de contactar o conservador e pedir permissão para aceder aos dados, ou para obter uma copia do que necessita em CD. Documentação sobre as condições de acesso (e uso) são importantes para os utilizadores poderem aceder aos dados e por isso é um aspeto de qualidade de dados. A documentação de acessibilidade pode incluir:

- Morada de contacto para os dados
- Condições de acesso
- Método de acesso (se os dados estiverem acessíveis eletronicamente)

- Formato dos dados
- Advertências
- Informação de direitos de autor
- Custos, se aplicável
- Restrições ao uso

9.7. Exatidão temporal

A exatidão temporal refere-se à exatidão da informação no tempo. Por exemplo: "*a exatidão dos dados dura um mês*". Isto pode ser importante para bases de dados onde o campo "dia" não possibilita valores nulos e nos casos onde não há informação disponível, automaticamente coloca "1" nesse campo. Isto pode levar a falsas imprecisões da exatidão. Isto é ainda mais importante onde só se conhece o ano do registo e a base de dados automaticamente coloca o 1º de Janeiro. Se um utilizador estiver a estudar o período de floração de uma planta ou o padrão migratório de uma ave, por exemplo, precisa de saber esta informação para que possa excluir estes registos (para o seu objetivo) como sendo de baixa qualidade e não "aptos para o uso".

9.8. Documentar procedimentos de validação

Uma das chaves para saber que erros existem é a documentação. É de pouca utilidade para qualquer pessoa se as verificações de qualidade de dados realizadas e correções feitas, não tiverem sido documentadas na sua totalidade. Isto é especialmente importante quando estas verificações estão a ser realizadas por outros que não o produtor original dos dados. Há sempre a possibilidade dos erros descobertos não serem afinal erros, e que alterações que sejam feitas adicionem novos erros. É também importante que as verificações não sejam feitas repetidamente. Não podemos dar-nos ao luxo de desperdiçar recursos desta maneira. Estes registos podem ser verificados e considerados perfeitamente bons e genuínos anómalos. Se esta informação não for documentada no registo, numa fase seguinte, alguém poderá realizar mais verificações de qualidade dos dados e identificar outra vez os mesmos registos como sendo suspeitos. Esta pessoa pode então excluir o registo da sua análise, ou gastar mais tempo valioso a reverificar toda a informação. Isto é gestão básica de risco e deve ser realizada rotineiramente por todos os conservadores de dados e utilizadores. O valor e a necessidade de uma boa documentação não pode ser forçada em demasia. Ela ajuda os utilizadores a conhecer o que os dados são, qual a qualidade e quais os propósitos dos dados devem estar aptos. Também auxilia os curadores e conservadores de dados a manter o controlo e a qualidade dos dados e não desperdiçar recursos na reverificação dos supostos erros.

9.9. Documentação e desenho de uma base de dados

Uma das maneiras de ter a certeza de que os erros estão devidamente documentados é inclui-los no plano inicial de *design* e construção da base de dados. Campos adicionais de qualidade de dados/exatidão podem ser posteriormente adicionados. Campos como exatidão posicional e geográfica, fonte de informação para a georreferenciação e elevação, campos para quem adiciona informação – foram os dados da coordenada adicionados pelo coletor usando um GPS, ou um operador de entrada de dados numa data posterior usando

um mapa numa escala particular, foi a elevação gerada automaticamente a partir da DEM, se sim, qual foi a sua fonte, a sua data e escala, etc. Todas estas informações vão ser avaliadas mais tarde quando se determinar a informação e tem valor para um uso em particular ou não, e o utilizador dos dados pode então decidir.

> "os utilizadores dos dados necessitam de ter em atenção quando baseiam avaliações biológicas em conjunto de dados taxonómicos, que não tenham presentes documentação especifica pelo menos de algumas características de desempenho." (Stribling et al. 2003).

10. Armazenamento de dados

O armazenamento de dados pode afetar de diferentes modos a qualidade dos dados. Alguns destes não são óbvios, mas têm de ser considerados tanto na conceção da forma de armazenamento (base de dados) bem como nos itens da cadeia de qualidade de dados.

O tema da seleção e desenvolvimento de uma base de dados é demasiado extenso para ser abordado aqui e deve ser objeto de estudo em separado. Um estudo encomendado pelo GBIF examinou o S*oftware* de Gestão de Coleções Referência (Berendsohn et al. 2003), que recomendo aos leitores que consultem.

Esta secção examina alguns dos princípios do armazenamento de dados que dizem respeito à relação com a qualidade de dados.

10.1. Cópia de segurança dos dados

Realizar uma cópia de segurança dos dados regularmente ajuda a garantir níveis de qualidade consistentes. É essencial que as organizações mantenham procedimentos de recuperação e cópia de segurança contra desastres. Sempre que os dados sejam perdidos ou corrompidos, há uma consequente perda de qualidade.

10.2. Arquivamento

O Arquivamento (inclui obsolescência e descarte) de dados é uma área da gestão de dados que necessita de maior atenção. O arquivamento dos dados, em particular por universidades, ONG's ou pessoas a título individual deve ser um assunto prioritário na gestão de dados. As universidades têm uma elevada mobilidade de pessoas e muitas das vezes os dados são armazenados de forma repartida – normalmente no PC do próprio investigador ou num arquivador. Se não estiverem bem documentados, estes dados podem perder muito facilmente a sua utilidade e acessibilidade. Normalmente, são descartados algum tempo depois do investigador ter deixado a instituição uma vez que ninguém sabe o que são ou ninguém tem o trabalho de os manter. É por este motivo que as universidades em particular, necessitam de estratégias de documentação e arquivamento.

Os investigadores individuais, que trabalhem fora de uma instituição necessitam de garantir que os seus dados são mantidos e/ou arquivados após a sua morte ou após deixarem de ter interesse neles. Do mesmo modo, as ONG's que podem não ter financiamento a longo prazo para o arquivamento de dados, necessitam de entrar em acordo com organizações apropriadas que tenham uma estratégia de gestão de dados a longo prazo (incluindo para arquivamento) e que possam ter interesse nos dados.

O arquivamento dos dados tornou-se mais fácil nos últimos anos devido ao desenvolvimento dos protocolos do DiGIR/Darwin Core e BioCASE/ABCD[27]. Estes protocolos proporcionam uma forma simples para uma instituição, departamento de uma universidade ou individuo, exportar a sua base de dados num destes formatos e armazená-los em formato XML, ou no seu *site*, ou encaminhar para uma instituição de acolhimento.

[27] http://www.tdwg.org/

Este é um modo de arquivar dados facilmente e a longo-prazo e/ou torná-los disponíveis através de procedimentos de pesquisa distribuídos como o portal de dados do GBIF.

A limpeza, descarte e arquivamento de dados são alguns dos problemas com dados na *World Wide Web*. As páginas de Internet que são abandonadas pelos seus criadores, ou que contenham dados antigos e obsoletos deixam o ciberespaço literalmente cheio de detritos digitais (várias referências). As organizações necessitam de construir uma estratégia de arquivamento dos seus dados na sua cadeia de gestão de dados. O arquivamento físico de dados é um tópico demasiado longo para ser incluído aqui, no entanto, foi publicado um documento recente sobre arquivamento de dados através do uso de CDs e DVDs pelo *Council on Information and Library Resources* e pelo *United States National Institute of Standards and Technology* (Byers 2003). Este documento é um resumo importante desta tecnologia e os leitores poderão estar interessados em consultá-lo.

"Os dados que já não são necessários (por razões legais ou outras) não devem ser destruídos, ou colocados em risco sem explorar todas as possibilidades – incluindo o arquivamento" (NLWRA 2003).

10.3. Integridade dos dados

A integridade dos dados refere-se às condições em que os dados foram ou não alterados ou destruídos de um modo não autorizado e as condições em foram ou não alterados, destruídos ou modificados de modo malicioso ou acidental (como por vírus ou picos de tensão).

Os dados mudam frequentemente – por exemplo, quando a informação taxonómica de um registo é atualizada após uma redeterminação – mas os utilizadores esperam que o sistema do computador mantenha a integridade dos dados e que o próprio sistema não altere inadvertidamente e de modo incorreto um valor. A corrupção de dados dá-se quando a integridade falha e ocorre inadvertidamente uma alteração incorreta.

A integridade dos dados é preservada através de uma boa gestão, armazenamento, cópias de segurança e arquivamento.

10.4. Padrões de erros

As bases de dados taxonómicas e de ocorrência de espécies, - tal como todas as bases de dados, são vulneráveis a padrões de erro no seu conteúdo. English (1999) reconhece os seguintes erros padrões a que chamava, defeitos dos dados. Dalcin (2004) adotou estes para o uso em base de dados taxonómicas. Os valores aqui são de English (1999) com exemplos citados a partir de Chapman (1991) e de bases de dados provenientes da *Australian Virtual Herbarium*[28] e *speciesLink*[29] do Brasil:

- **Valores redundantes no domínio** – existem valores não padronizados ou valores sinónimos quando dois ou mais valores ou códigos têm o mesmo significado. A redundância é muito típica em dados descritivos, se estes não seguirem as

[28] http://www.chah.gov.au/avh/
[29] http://splink.cria.org.br/

terminologias padronizadas ou onde a compilação de dados de diferentes fontes for mal controlada.

- **Dados com valores em falta** - Não existe valor num campo de dados que deve conter valores. Isto diz respeito aos campos de preenchimento obrigatório e aos campos que não é obrigatório preencher na captura de dados, mas que serão necessários no processamento posterior. Temos como exemplo valores de georreferenciação ou de coordenadas (latitude e longitude).
- **Dados com valores incorretos -** Isto pode ser causado pela transposição ao datilografar, inserção de dados no lugar errado, má interpretação do significado dos dados recolhidos, a impossibilidade de ler corretamente a etiqueta ou o catalogador não saber o valor a colocar nos campos de preenchimento obrigatório. Dados com valores errados são o erro mais óbvio e comum e podem afetar todos os valores dos dados em todos os campos. Os erros ortográficos em nomes científicos é um erro padrão comum associado a valores incorretos nos dados em bases de dados taxonómicas ou nomenclaturiais (veja discussão na outra secção), bem como a inserção do valor zero em campos da georreferenciação, etc.
- **Dados com valores não-atomizados**[30] - Ocorrem quando mais do que um valor é inserido no mesmo campo (e.g. género, espécie e autor no mesmo campo, ou o *rank* e o nome intraespecífico). Este tipo de padrão de erros resulta normalmente de um *design* da base de dados mal pensado. Este tipo de padrão de erros pode causar problemas sérios na integração dos dados.

Género	Espécie	Infraespécie
Eucalyptus	globulus	subsp. bicostata
Família	**Espécie**	
Myrtaceae	Eucalyptus globulus Labill	

Tabela 4. *Exemplos de valores de dados Não-atomizados.*

- **Esquizofrenia do domínio** - Campos usados para inserir valores diferentes daqueles para o qual o campo foi projectado que acabam por se incluir dados de natureza diferentes.

Família	Género	Espécie
Myrtaceae	Eucalyptus	globulus?
Myrtaceae	Eucalyptus	? globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?

[30] No Brasil, Não atômicos.

Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	To be determined

- **Ocorrências duplicadas** – Registos múltiplos que representam uma única entidade. Os casos mais comuns ocorrem quando existem grafias ou nomenclaturas alternativas válidas. Isto pode criar dificuldades aos utilizadores quando pesquisam por um nome, ou quando tentam combinar dados de diferentes bases de dados. Exemplos:
    - *Phaius tancarvilleae*
    - *Phaius tankervilliae*
    - *Phaius tankarvilleae*
    - *Phaius tankervilleae*
    - *Phaius tankervillae*
    - Brassicaceae/Cruciferae (equivalentes exatos; ambos são permitidos pela *International Botanical Code*).
- **Dados com valores inconsistentes** – Ocorrem quando os dados de bases de dados relacionadas podem ser atualizados de forma inconsistente ao mesmo tempo ou em momentos distintos em ambas as bases. Por exemplo, entre a base de dados da coleção viva e de um herbário, ou entre as bases de dados de colecções de um museu e a base de dados de imagens relacionadas.
- **Contaminação da qualidade da informação** – Resulta da combinação de dados exatos com dados não exatos. Por exemplo a combinação de dados com informações ao nível da subespécie numa base de dados que inclui apenas dados até ao nível de espécie.

10.5. Dados espaciais

O armazenamento de dados espaciais abrange a informação sobre a localização (informações textuais da localidade) bem como informação de coordenadas (dados georreferenciados) normalmente em pares de coordenadas (uma abcissa e uma ordenada). Muitas bases de dados estão agora a começar a incluir dados de localização analisados ou atomizados, como o nome do local mais próximo, distância e direção, adicionalmente a informação em texto livre da localização. Muitos projetos já estão em andamento para melhorar a análise da informação que está presente nos textos livres sobre a localização, para criar um campo automático e para auxiliar no processo de georreferenciação. O projeto BioGeomancer criado pela Fundação Gordon e Betty Moore é um desses projetos.

A informação georreferenciada (ou coordenadas) é normalmente inserida nas bases de dados como latitude e longitude (sistema de coordenadas esféricas) ou em coordenadas UTM (ou relacionado) (sistema de coordenadas planimétricas). Um sistema de coordenadas esféricas tal como latitude e longitude está baseado no globo e para serem representadas em papel tem de ser esticados de modos pouco usuais, conhecidas por projeções. Neste tipo de sistema, as áreas não são iguais e a distância entre um grau de latitude e o próximo, por exemplo, podem variar consideravelmente dependendo se a pessoa está mais

próxima do equador ou dos polos. Os sistemas de coordenadas planimétricos são próximos às projeções das áreas e podem ser usados para medir ou fazer cálculos de áreas.

Muitas instituições estão agora a começar a introduzir dados em graus, minutos e segundos ou em graus e minutos decimais (como reportado por muitas unidades de GPS), onde a base de dados converte para graus decimais para armazenamento. Para transferir ou usar em SIG é normalmente mais adequado armazenar em graus decimais pois facilita a sua transferência e proporciona a maior exatidão possível.

O armazenamento de dados em coordenadas UTM normalmente ocorre em instituições em que os dados são restritos a uma zona UTM. É a vantagem de ser baseada numa área como discutimos em cima, cada grelha é um quadrado (ou retângulo) e permite a sua fácil representação num mapa no plano ou para calcular distâncias e áreas. No entanto, quando se armazena dados através do sistema de coordenadas UTM (ou relacionados) é importante que a zona também seja armazenada, caso contrário surgem dificuldades na combinação de dados de outras áreas ou instituições.

10.6. Graus decimais

O armazenamento em graus decimais em muitas bases de dados pode levar a *precisões falsas* como foi mencionado acima. A precisão com que cada dado está armazenado (e é disponibilizado) deve ser tido em conta. A base de dados não deve permitir reportar uma precisão maior do que a precisão mais alta atribuída aos dados nela contida. Para a maioria dos dados biológicos, estes valores serão de cerca de 4 casas decimais (cerca de 10 metros).

10.7. Datums

Existem muitos datums geodésicos. A terra não é uma verdadeira esfera, mas sim um elipsóide e as dificuldades surgem quando se tenta encaixar o sistema de coordenadas na superfície deste elipsóide (Chapman *et al*. 2005). Para resolver isto, foi criado o conceito de "datum". Um datum é um conjunto de pontos usados para referenciar uma posição na esfera para um elipsóide de revolução. Historicamente, foram criados diferentes sistemas de referência para as diferentes partes da terra, e foram os avanços[31] na área dos satélites que permitiram criar um verdadeiro sistema de referência global ou datum à medida que os satélites foram usados para fixar o centro da terra.

As diferenças de latitude e longitude de uma posição na terra usando diferentes datums geodésicos podem ser de 400 metros ou mais (Wieczorek 2001).

Por causa desta diferença, é importante que as bases de dados registem os datums utilizados senão quando os dados são combinados, o erro resultante entre dois registos para a mesma localização pode ser muito significativo.

[31] No Brasil, advento

11. Manipulação de dados espaciais

Existem várias modos de manipular dados espaciais. Muitos deles não têm qualquer efeito na exatidão dos dados espaciais, enquanto outros têm. Alguns dos métodos que afetam a exatidão posicional dos dados espaciais são:

11.1. Conversão do formato de dados

Possivelmente, a conversão de dados mais comum é a conversão das coordenadas decimais em graus/ minutos/ segundos (DMS para DD) efetuada por aqueles que estão envolvidos no armazenamento ou no uso dos dados de espécies ou de ocorrência de espécies, ou de coordenadas UTM para graus decimais (UTM para DD) de uma coleção. Outras alterações incluem converter de milhas para quilómetros em descrições textuais de localidades, a conversão de pés para metros e de registos de profundidades, altitude, etc.

Todas estas conversões são relativamente simples, mas podem levar a falsas impressões de exatidão devido ao mau uso da precisão. Por exemplo uma coleção que dê uma altitude de 250 pés (o que o coletor pode ter querido dizer entre 200 e 300 pés) quando convertido para valores métricos será 76,2 metros (uma casa decimal) ou talvez 76 metros se arredondado. Seria melhor registar a conversão do valor para 80 metros e seria ainda melhor incluir uma exatidão de campo de talvez 20 metros (±). O uso de precisões falsas pode levar ao que parece ser um aumento de exatidão, mas na realidade é uma perda de qualidade.

11.2. Datums e Projeções

A conversão dos dados de um datum geodésico para outro pode levar a erros bastante significativos pois as conversões não são uniformes (veja Wieczorek 2001 para uma discussão de datums e o seu efeito na qualidade dos dados). Muitos países ou regiões estão agora a converter a maioria dos seus dados para um padrão da sua região - ou o *World Geodetic Datum* (WGS84), ou datums que se aproximam muito de perto (como o *Australian Geographic Datum* (AGD84), na Austrália, que varia do WGS84 à volta de 10 cm, ou o EUREF89 na Europa que varia do WGS84 à volta de 20 cm, são dois exemplos). A conversão da posição de um datum para outro, por exemplo, provavelmente não é necessária se a exatidão do dado for aproximadamente de 5 ou 10 km. Se estiver a lidar com dados com exatidão de cerca de 10-100, no entanto, a alteração de datums pode ser muito importante e significativa (em algumas áreas acima de 400 m ou mais – Wieczorek 2001).

De forma similar, quando os dados estão mapeados em polígonos (e.g. colheitas de um parque natural), é necessário estar atento a erros que possam surgir na conversão de uma projeção para outra (e.g. Albers para Geographic). Estão disponíveis fórmulas padrão para se calcular o erro que surgirá ao fazer essas conversões, e os metadados que acompanham os dados devem refletir essa informação.

11.3. Grelhas

Sempre que um dado é convertido de um formato vetorial para raster ou grelha, há perda de exatidão e precisão. Isto deve-se ao tamanho das células da grelha no ficheiro raster que é usado para aproximar o dado vetorial (Burrough e McDonnell 1998). A precisão e exatidão não podem ser recuperadas reconvertendo os dados para o formato vetorial. Para uma discussão mais alargada dos problemas encontrados no uso e na conversão de dados raster e de problemas de escala veja Chapman *et al.* (2004).

11.4. Integração de dados

Os conjuntos de dados geográficos são difíceis de integrar quando têm inconsistências entre eles. Estas inconsistências podem envolver tanto atributos espaciais como características dos dados e pode ser necessário usar várias medidas de correção, que por vezes são consumidores de tempo (Shepherd 1991). As inconsistências podem resultar de:

- Diferenças na recolha ou técnica de medição (e.g. tamanho da área e períodos de tempo nos dados de observação), métodos de recolha (tamanho da grelha, largura do transeto) ou categorias de dados (e.g. diferentes definições de categorias para dados categóricos).
- Erros nas medições ou métodos de amostragem (e.g. erros na transcrição, registo de dados, identificações).
- Diferenças de resolução (espacial, temporal ou atributo).
- Definições vagas e imprecisas.
- Falta de precisão dos objetos (e.g. limites do solo ou vegetação, identificações onde algumas são até à espécie, outras até à sub-espécie e outras só até ao género).
- Diferenças no uso ou na interpretação da terminologia e nomenclatura (e.g. uso de taxonomias diferentes).
- Diferenças nas propriedades do GPS (datum, sistema de coordenadas, etc.).

Estes problemas de integração são maiores onde os dados são:

- De diferentes tipos (e.g. dados de espécimes de um museus misturados com dados de pesquisa e de observação).
- De diferentes jurisdições (e.g. onde os métodos de pesquisa podem ser diferentes).
- Obtidos de múltiplas fontes.
- Consiste em diferentes tipos de dados (mapas, espécimes, imagens, etc.).
- De diferentes períodos de tempo.
- Armazenado em diferentes tipos de base de dados, meios, etc. (e.g. alguns programas de bases de dados não permitem valores “nulos”).
- Analisados de várias maneiras (e.g. quando uma base de dados inclui o nome cientifico completo no mesmo campo ou outras têm-no separados em diferentes campos como género, espécies).

A integração de dados produz resultados de maior qualidade quando os provedores e detentores dos dados seguiram e usaram de forma consistente os padrões de armazenamento.

12. Representação e Apresentação

> "*Os Métodos devem ser sempre desenvolvidos para tornar mais eficiente o uso dos dados existentes, qualquer que seja a sua qualidade. No entanto, para que os dados sejam fiáveis, também têm de ser validados ou acompanhados por informação que indique o seu nível de fiabilidade*" (Olivieri et al. 1995).

No seu papel de entender, explicar, quantificar e avaliar a biodiversidade, os cientistas e as instituições científicas são cada vez mais reconhecidos como provedores de informação. Este reconhecimento está baseado na capacidade de fornecer informação fiável e usável por decisores, gestores, público em geral e outros. Informação ambígua, confusa, incompleta, contraditória e errada, resultante de uma pobre gestão da base de dados, pode afetar a sua reputação como provedor de informação e autoridade científica (Dalcin 2004).

Uma das principais finalidades da manipulação de dados digitais nas ciências biológicas é fornecer aos utilizadores, informações com baixo custo de consulta e utilização da mesma. Nesse sentido, o seu sucesso é determinado pela medida em que ele pode fornecer ao utilizador de um modo exato uma visão do mundo biológico. Mas o mundo biológico é infinitamente complexo e precisa de ser generalizado, aproximado e abstraído para ser representado e compreendido (Goodchild *et al*. 1991). A maneira de fazer isto é através do uso de sistemas de informação geográfica, ferramentas de modelação ambiental e sistemas de apoio à decisão. No entanto, para usar essas ferramentas, é essencial que essa variação seja amostrada e medida, e que os erros e incertezas sejam descritas e visualizadas. É nesta área que temos ainda um longo caminho a percorrer até alcançar ao que se poderia chamar boas práticas.

A Biologia foi uma das primeiras disciplinas a desenvolver técnicas de reportar[32] erros usando barras de erro e estimativas estatísticas. O reportar do erro não é visto como uma fraqueza pois os erros fornecem informação crucial para uma correta interpretação dos dados (Chrisman 1991). Na entrega de dados de espécies, é necessário desenvolver e usar técnicas semelhantes de deteção e de reportar erros, para que os utilizadores dos dados os possam interpretar e usar corretamente.

Programas de qualidade de dados eficazes ajudam a prevenir constrangimentos para as instituições e para indivíduos – tanto internamente como publicamente.

12. 1. Determinar as necessidades dos utilizadores

Determinar as necessidades dos utilizadores não é um processo simples, pois é difícil desenvolver requisitos detalhados e depois estruturar os dados de acordo com eles. Mas é importante localizar utilizadores chave e trabalhar com eles para conhecer as suas necessidades e requisitos. Bons requisitos de utilizadores pode levar a uma melhor recolha e gestão de dados e no geral uma melhor qualidade dos mesmos.

[32] No Brasil, relatorizar

12.2. Relevância

A relevância relaciona-se de perto com a "qualidade" e refere-se à relevância dos dados para a sua utilização. Pode estar relacionado com coisas tão simples como tentar usar a Flora de uma área, noutra área para a qual não foi elaborada ou para dados que possam estar numa projeção diferente daquela que era suposto, requerendo assim a um trabalho considerável para os tornar úteis e "relevantes".

12.3. Credibilidade

A credibilidade é uma dimensão dos dados em que são considerados pelo utilizador como sendo credíveis (Dalcin 2004). Está muitas das vezes sujeita à perceção ou avaliação do utilizador tendo em conta a adequação dos dados para o seu propósito e pode ser baseado em experiência anterior ou comparação com padrões conhecidos (Pipino *et al*. 2002). A reputação de um conjunto de dados pode por vezes depender da perceção da credibilidade dos dados pelos utilizadores (e como tal a usabilidade), mas é algo que pode ser melhorado com uma boa documentação.

Wang et al. (1995) incluiu um diagrama que relaciona muitos destes tópicos numa representação hierárquica e mostra a relação entre entidades tais como credibilidade e reputação, etc.

12.4. Viver com incerteza em dados espaciais

A incerteza, especialmente nos dados espaciais, é um facto, mas na maioria das vezes a incerteza nos dados não está bem documentada e nem sempre é óbvia para os utilizadores. A proliferação de sistemas de mapeamento simples de usar, permitiu que pessoas não profissionais em SIG conseguissem facilmente visualizar e analisar relações espaciais nos seus dados, mas na maioria dos casos é feito usando escalas inapropriadas (Chapman *et al*. 2005), e sem ter em conta o erro espacial e incerteza inerente aos dados (Chapman 1999). Em alguns casos isto pode levar a um uso errado de dados, com consequências trágicas, ocasionalmente (Redman 2001). Recentemente, houve um aumento no número de serviços de mapas *online* que permite os utilizadores ver e analisar dados espaciais como se fosse no SIG tradicional, mas permite ao publicador do serviço controlar as camadas de dados e a escala do conjunto de dados que aparecem. Num futuro próximo isto vai ser expandido com o desenvolvimento de *Web Mapping Services* (WMS) funcionais. O controlo de camadas de dados e de escala pelos publicadores do mapa (e.g. permitir que diferentes camadas possam ser tornadas disponíveis ou não, com o zoom escolhido pelo utilizador) reduz alguns dos erros de amostra simples que de outro modo poderia não ser feito.

É essencial que a incerteza nos dados seja documentada, primeiro através do uso de bons metadados e, segundo, através da visualização e apresentação. Uma das áreas de investigação que necessita de continuar a desenvolver técnicas para visualizar a incerteza é a de dados de espécies e de ocorrência de espécies – por exemplo, para mostrar as marcas da exatidão. Em vez dos registos da coleção serem representados como um ponto de latitude e longitude há a necessidade de incluir a exatidão associada ao registo e assim ficar ligada aos passos da localização – um círculo, uma elipse, etc. e talvez até incluir nos níveis de probabilidade. (Chapman 2002).

É importante que quem conheça os dados e as suas limitações em relação à exatidão posicional e/ou de atributos dêem assistência aos utilizadores, documentando e tornando disponível essa informação para que estes possam orientar os utilizadores a determinar a aptidão dos dados para o seu uso.

12.5. Visualização do erro e incerteza

Ainda há um grande caminho a percorrer para se desenvolver bons métodos de visualização de erros para dados de espécies, apesar de já terem sido desenvolvidos novos e emocionantes métodos (e.g. Zhang e Goodchild 2002). Talvez o método mais simples seja através do uso de uma camada de erro como uma sobreposição adicional no SIG. Esta técnica tem sido usada na cartografia mundial onde uma camada pode proporcionar o sombrear de diferentes intensidades para mostrar a fiabilidade das diferentes partes do mapa. Outras técnicas podem envolver o uso de símbolos diferentes (uma linha a tracejado em oposição a uma linha sólida, pontos de tamanho e intensidade diferente, etc. para indicar dados de menor qualidade ou exatidão). O uso destas sobreposições também pode dar pistas de como os erros foram originados e esta pode ser uma ferramenta valiosa para a validação e verificação dos dados.

O uso de uma matriz de classificação de erros em que as linhas proporcionam os resultados esperados, e as colunas os resultados observados, é útil quando tais cálculos estatísticos são possíveis. Nestes casos, os erros ao longo das linhas são erros de omissão e erros ao longo das colunas erros de comissão (Chrisman, 1991). Estes métodos geralmente não se prestam para utilização com dados de ocorrência de espécies, mas podem ser importantes, por exemplo, com registos de dados de amostragem, onde a presença / ausência são observadas durante um período de tempo.

12.6. Avaliação do Risco

Os decisores preferem um clima de certeza, no entanto, os sistemas naturais são inerentemente variáveis e raramente estão em conformidade com este desejo. As técnicas de avaliação de risco oferecem cada vez mais aos decisores e gestores ambientais uma estimativa de certeza e risco, para que as decisões possam ser tomadas com maior segurança. No caso das espécies, em que o conhecimento da sua ocorrência exata é muitas vezes insuficiente, as áreas de "ocorrência provável" podem ser utilizadas como substituto. Dentro de grandes áreas de 'ocorrência provável', no entanto, pode haver áreas que são mais "prováveis" do que outras (Chapman, 2002).

O conceito de risco geralmente pode ser visto como tendo dois elementos - a probabilidade e a magnitude de algo acontecer e as consequências se e quando o evento acontecer (Beer e Ziolkowski 1995). Num contexto de dados de espécies, a avaliação de risco pode ser estendida desde o risco de um incêndio local destruir os dados se os procedimentos de *backup* não tiverem sido implementados, até ao risco de uma decisão ambiental estar errada devido ao uso de dados de má qualidade. Um exemplo disso pode ser o custo envolvido na proibição de desenvolvimento de uma área por causa da informação de que uma espécie ameaçada lá ocorre. Em algumas situações ambientais, os governos estão a aumentar a fiscalização na aplicação do *princípio da precaução* na tomada de decisões ambientais importantes.

12.7. Responsabilidades legais e morais

Podem surgir uma série de questões legais e morais nas diferentes áreas em relação a dados de espécies. Estas incluem:

- *Copyright* e Direitos de propriedade intelectual;
- Privacidade;
- Veracidade da etiquetagem;
- Apresentação restrita de qualidade para taxa sensíveis;
- Direitos indígenas;
- Responsabilidade;
- Advertências e isenção de responsabilidade.

Na maioria dos casos os ***Copyright e Direitos de propriedade intelectual***de dados podem estar comtemplados na documentação que acompanhe os dados. Quando estes variam de registo para registo, os direitos devem ser registados ao nível do registo, ou caso contrário essa informação pode estar nos metadados.

Vários países introduziram recentemente legislação de ***privacidade***, e os detentores de dados devem estar cientes das implicações da legislação referida. Isto pode ser particularmente relevante em dados que estão a ser transferidos através de fronteiras políticas ou disponibilizados através da Internet. Em alguns países, as informações sobre os indivíduos não podem ser armazenadas em base de dados ou disponibilizados sem a sua autorização expressa. Como isso pode afetar a informação associada aos dados de ocorrência-espécie não está claro, no entanto, os detentores de dados devem estar cientes do problema e prevenir, sempre que necessário.

Boas medidas de controlo de qualidade, juntamente com bons metadados normalmente levam a existir concordância com o conceito de "verdade na etiquetagem". Até agora, pelo menos na legislação, "verdade na etiquetagem" tem estado restrito a produtos alimentares. No entanto, é mencionado em trabalhos relacionados com o desenvolvimento de uma infra-estrutura global de dados espaciais (Nebert e Lance 2001, Lance 2001), infra-estrutura de dados espaciais nacional de dados espaciais para os EUA (Nebert 1999) e uma infra-estrutura para a Austrália e Nova Zelândia (ANZLIC 1996b). Na publicação da Global SDI (Lance, 2001), recomenda-se que a Câmara de Dados Espaciais deva incluir "*um método de publicidade gratuita para fornecer acesso mundial às existências sob o princípio da "verdade-na-etiquetagem"*, e para citar o australiano e documento da Nova Zelândia:

> "*Padrões de qualidade de dados terrestres e geográficos podem ser descritivos, prescritivos ou ambos. A norma descritiva é baseada no conceito de "verdade na etiquetagem", que obriga os produtores de dados a relatar o que sabem sobre a qualidade dos dados. Isso permite aos utilizadores dos dados fazer uma avaliação informada sobre a "adequação à finalidade" dos dados.*"

Apresentação restrita de qualidade para taxa sensíveis pode ocorrer onde a informação sobre a localidade é "difusa" - por exemplo, para restringir o conhecimento do local exato de espécies ameaçadas , comércio de espécies sensíveis, etc. Isto é uma redução na qualidade dos dados publicados, quando acontece deve estar documentado de

uma forma clara para que os utilizadores que estão a usar os dados possam decidir se estes servem ou não para a utilização que necessitam.

Direitos indígenas podem afetar também a qualidade de dados, pois pode haver casos onde alguma informação tem de ser restrita, pois algumas informações podem ser sensíveis para os povos indígenas. Deve ser incluída documentação para referir que "alguns dados são restritos por forma a respeitar os direitos indígenas".

Em 1998, Epstein *et al.* analisou o assunto da responsabilidade legal em relação ao uso de informação espacial. Alguns pontos chaves principais são:

- *Existe agora "**considerável potencial**" para o litígio e para a perda de reputação e integridade tanto da pessoa como da organização decorrente de erros na informação espacial.*
- *Os tradicionais avisos de isenção de responsabilidade podem não ser uma forte defesa em caso de litígio.*
- *A fim de limitar a responsabilidade, as organizações podem ser obrigadas a manter um alto nível de qualidade na documentação que legendem adequada e verdadeiramente os seus produtos para a "**melhorar a sua capacidade e conhecimento**".*

Advertências e avisos de isenção de responsabilidade são uma parte importante da documentação da qualidade de dados. Devem ser escritos de um modo não exclusivo para a organização detentora dos dados, mas também que dê ao utilizador alguma ideia da qualidade de dados e o que esperar dessa qualidade.

> *A maioria das agências e grupos envolvidos com a produção de dados serão julgados de acordo com a facilidade com a qual os dados e a informação sejam disponibilizados, e pela qualidade da informação. Aqueles que conseguirem publicar, partilhar, acede[33], integrar e utilizar a informação, serão aqueles que mais irão beneficiar* (NLWRA 2003).

12.8. Certificação e Acreditação

Podem e devem os dados de ocorrência espécies serem certificados? Com o tornar público dados de diferentes agências, os utilizadores querem saber quais as instituições em que podem confiar, e que sigam os procedimentos de controlo de qualidade documentados. Deverão os dados depender só de instituições bem conhecidas ou existem instituições menos conhecidas também com dados fiáveis? Quais os dados disponibilizados pelas instituições mais conhecidas que são fiáveis e quais não.

A *Reputação* por si só pode ser o fator decisivo em que o utilizador se baseia para escolher as suas fontes de dados, mas a reputação é um conceito subjetivo e frágil para fundamentar ações e decisões (Dalcin, 2004). É isso que queremos na nossa disciplina? Bons metadados e documentação de procedimentos de qualidade de dados, muitas vezes, podem transformar um fator subjetivo, tal como a reputação, em algo que os utilizadores possam basear uma avaliação mais científica e fundamentada. Talvez devêssemos desenvolver processos de certificação e acreditação que informe os utilizadores de organizações que estão em conformidade com as mínimas normas e procedimentos de documentação de qualidade de dados.

[33] No Brasil, acessar

O desenvolvimento da certificação de qualidade acordada poderá levar a uma melhoria na qualidade geral dos dados e a uma maior segurança entre os utilizadores sobre o valor dos dados. Este, por sua vez, poderia levar a um melhor financiamento para organizações certificadas. Dalcin (2004) sugere que "*a certificação de qualidade de dados taxonómicos pode envolver três aspetos: fontes de dados primários (matéria-prima), a cadeia de informação (o processo) e a base de dados (o produto)* ".

12.9. Revisão por pares de bases de dados

Um sistema de revisão por pares para bases de dados pode ser introduzido para base de dados de espécies. Este processo de revisão poderia contribuir para um processo de certificação como examinado acima, e pode envolver questões como procedimentos de controlo de qualidade, documentação e metadados, atualização e mecanismos de retorno de comentários, etc.

13. Conclusão

> "*Um dos objetivos de qualquer especialista de informação é evitar o erro desnecessário. Ao reconhecer diretamente o erro, pode ser possível limitá-lo para limites aceitáveis. No entanto, o erro não pode ser sempre evitado de uma forma barata ou fácil*" (Chrisman 1991).

Nunca é demais salientar a importância da qualidade de dados e verificação de erros. Como referido ao longo deste documento, é essencial que os dados tenham valor real para o desenvolvimento de resultados que levarão a melhores decisões e gestão ambiental. A qualidade de dados é uma questão importante para todos os dados, sejam eles de um museu ou dados de colheita de herbário, registos de observação, dados de amostragem, ou *checklists* de espécies. Há uma exigência de agregação por muitos governos ao redor do mundo para que os dados sejam de alta qualidade e melhor documentados. Por exemplo:

- Existe uma insistência do *Australian Federal, State and Territory Governments* para melhorar os serviços e fazer um uso mais eficiente dos recursos, incluindo dados e recursos de informação.
- Há um crescente reconhecimento de que os dados recolhidos à custa de fundos públicos devem ser devidamente geridos para torná-los acessíveis ao público, de modo a perceber o seu potencial e justificar os custos consideráveis de produção e manutenção envolvidos.
- Há uma crescente pressão dos clientes para que o acesso aos dados seja mais fácil e rápido e que as informações sejam corretas e que eles sejam fornecidos com pouco ou nenhum custo.
- Há um foco maior dentro dos governos para a necessidade de racionalizar e combinar dados, a fim de melhorar a eficiência e agregar valor.
- Existe uma exigência crescente de que os dados sejam relevantes. Isso aplica-se a novas coleções, novas pesquisas, para a gestão e publicação de dados.

A necessidade de dados de qualidade não está em questão, mas muitos gestores de dados supõem que os dados contidos e retratados no seu sistema estão absolutamente livres de erros ou que os erros não são importantes. Mas o erro e a incerteza são inerentes a todos os dados, e todos os erros afetam o uso final a que os dados possam ser sujeitos. Os processos de aquisição e gestão de dados para melhorar a sua qualidade são parte essencial da gestão de dados. Todas as partes da cadeia da qualidade da informação precisam de ser examinadas e melhoradas por organizações responsáveis de dados de ocorrência-espécie e a documentação é a chave para os utilizadores serem capazes de conhecer e compreender os dados e para serem capazes de determinar a sua "aptidão para o uso" e assim, a sua qualidade.

> "*O fator humano é potencialmente a maior ameaça para a exitadão e fiabilidade das informações espaciais. É também um fator que pode garantir tanto a confiabilidade, além de gerar um entendimento, das fraquezas inerentes a qualquer conjunto de dados espaciais*" (Bannerman, 1999)

Agradecimentos

Muitos colegas e organizações à volta do mundo contribuíram para este documento de uma forma ou de outra. Alguns directamente, alguns por estarem envolvidos em discussões com o autor durante mais de 30 anos, e alguns indirectamente através de artigos publicados ou apenas por fazerem com que a sua informação estivesse disponível para o mundo.

Em particular, gostaria de mencionar de modo particular os colaboradores, tanto os passados como presentes, do CRIA (Centro de Referência de Informação Ambiental) em Campinas, Brasil, e o ERIN (Environmental Resources Information Network) em Canberra, Austrália, os quais contribuíram com ideias, ferramentas, teorias e um enquadramento sonante que ajudaram o autor a formular as suas ideias. A sua discussão do erro e precisão em informação ambiental ao longo dos anos e trabalho pioneiro realizado por eles, pelo CONABIO no México, a Universidade do Kansas, CSIRO na Austrália, a Universidade do Colorado, o Peabody Museum em Connecticut, e a Universidade da Califórnia, em Berkeley, assim como outros demasiado numerosos para mencionar, ajudaram a trazer-nos para o estado em que estamos hoje em gestão de qualidade de dados de espécies. Eu agradeço-lhes pelas suas ideias e crítica construtiva. Adicionalmente, as discussões com Town Peterson e outros na Universidade do Kansas, Barry Chernoff na Wesleyan University no Connecticut, Read Beaman na Yale University, John Wieczorek e Robert Hijmans na Universidade da Califórnia, Berkeley, Peter Shalk e outros no ETI, em Amesterdão, Stan Blum na Academia da Califórnia e o Academy and the pessoal do GBIF em Copenhaga que me deram ideias e desafios que levaram a algumas das ideias expressas nesta publicação. Quaisquer erros, omissões ou controvérsias são, no entanto, da responsabilidade deste autor.

Eu gostaria também de agradecer a aqueles que fizeram criticas, comentários e sugestões durante a edição deste documento, e em particular aos seguintes membros do Comité GBIF para a Digitalização de Dados de Colecções de História Natural: Anton Güntsch, Botanic Garden and Botanical Museum Berlin-Dahlem, Alemanha; Francisco Pando, Real Jardín Botánico, Madrid, Espanha; Mervyn Mansell, USDA-Aphis, Pretoria, África do Sul; A. Townsend Peterson, University of Kansas, EUA; Tuuli Toivonen, University of Turku, Finlândia; Anna Wietzman, Smithsonian Institution, EUA assim como a Patricia Mergen, Belgian Biodiversity Information Facility, Bélgica.

O Lany Speers do GBIF foi instrumental enquanto encarregado do relatório e acompanhamento em todas as suas fases.

Em conclusão, gostaria de agradecer ao projecto FAPESP/Biota no Brasil em possibilitar-me a oportunidade e suporte para expandir as minhas ideais sobre gestão de qualidade de dados durante a minha estadia no Brasil em 2003-2004, e à organização GBIF por suportar e encorajar a produção deste relatório

Referências

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Minnesota Planning. 1999. Positional Accuracy Handbook. Using the National Standard for Spatial data Accuracy to measure and report geographic data quality. Minnesota Planning: Land Management Information Center. http://www.mnplan.state.mn.us/pdf/1999/lmic/nssda_o.pdf [Accessed 14 Jul. 2004]

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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