Qualidade dos dados e Interoperabilidade em SIG

CLAUDIA BAUZER MEDEIROS, ALEXANDRE CARVALHO DE ALENCAR

IC - UNICAMP - CP 6176 13081-970 Campinas SP Brasil

Abstract. Interoperability in GIS is an issue of growing importance, due to the increase in number and volume of available data sources and to the exponential expansion of new applications and systems. Research in this area involve solutions directed towards the different layers of an information system (interoperability based on common interface design, process interoperability or interoperability through data). The goal of this paper is to point out issues concerning interoperability at the data level. In particular, the text analyses issues related to the quality of geographic data as an additional dimension that must be taken into consideration in the cases of data migration and integration.

Resumo. A interoperabilidade em SIG é uma questão cada vez mais importante, tendo em vista o aumento em número e volume das fontes de dados disponiveis e o crescimento exponencial de novos sistemas e aplicações. Os trabalhos na área envolvem soluções voltadas a diferentes camadas de um sistema de informação (interoperabilidade a partir de projeto de interfaces comuns, interoperabilidade visando processos ou interoperabilidade através dos dados). O objetivo deste artigo é explorar a interoperabilidade em SIG do ponto de vista de dados. Em especial, o texto analisa questões relativas à qualidade de dados geográficos como u ma dimensão a ser necessariamente considerada quando se trata de migração e integração de dados.

1 Introdução e motivação

A proliferação de sistemas e dados motivou o aparecimento da área de pesquisa em *interoperabilidade*. Como ressaltado em um relatório de um workshop do NSF [Int97], o termo interoperabilidade significa várias coisas para várias pessoas - denota sistemas abertos, ou a habilidade de intercambiar e transmitir dados, ou ainda uniformidade na interação do usuário e a construção de interfaces customizáveis. Estudos nesta área tratam de tres tipos de problema geral: interfaces (para trabalho cooperativo entre usuários), processos (para troca de mensagens e dados entre sistemas) e dados (para reuso e intercâmbio). Estes estudos visam não apenas diminuir os custos de execução de aplicações e de uso de dados, mas também reduzir os custos de treinamento em novos sistemas, de implementação e de manutenção de aplicações.

Em alguns casos, pesquisa relativa à interoperabilidade em SIG pode tomar como ponto de partida os estudos realizados sobre interoperabilidade de sistemas em geral: por exemplo, a interoperabilidade em SIG no nível de processos é abordável sob a ótica de sistemas distribuídos, em que a preocupação é sincronizar operações e padronizar parâmetros. (Vide, por exemplo, o trabalho de [YB98], que trata a interoperabilidade de SIGs como algo gerenciável no nível de chamadas RPC.)

As caracteristicas peculiares a dados e sistemas geográficos criaram novos problemas de interoperabilidade, não encontrados em sistemas em que os dados não têm componente espacial. De fato, dados geográficos são heterogêneos por natureza e dão margens a distintas interpretações. Desta forma, a interoperabilidade em SIG frequentemente se ocupa de atividades de *integração de dados* ou de *padronização de dados*, com menos ênfase em aspectos de processos.

Este trabalho discute questões de interoperabilidade em SIG do ponto de vista de gerenciamento de dados e, em especial, bancos de dados. Sob este prisma, pode-se identificar de uma forma geral três tipos de abordagem principal: (i) questões de modelagem, (ii) integração a partir de federações e (iii) propostas de padrões. O artigo propõe que, como parte deste contexto, considere-se explicitamente características de *qualidade* dos dados geográficos. Isto oferece novas perspectivas do ponto de vista de interoperabilidade, pois permite quantificar determinados aspectos que, por sua vez, indicarão quando é apropriado intercambiar dados e quais dados podem ser reusados.

Na área de gerência de qualidade da fabricação de produtos ou prestação de serviços qualidade é comumente definida como: conformidade com especificações [Cro79], nível de satisfação do cliente [Dem86] e adequação ao uso. A partir destas definições pode-se afirmar que o foco da abordagem sobre qualidade evoluiu a partir de uma preocupação inicial apenas com o produto para uma abordagem que cada vez mais preocupa-se com o usuário e a utilização que este fará do produto.

A abordagem sobre qualidade em se tratando de dados geográficos também evoluiu segundo esta linha de pensamento. A qualidade em SIG era antes retratada apenas pela reputação de quem produzia os dados, por experiências resultantes do uso do produto e por uma declaração do órgão

produtor informando que os dados produzidos estavam de acordo com um padrão cartográfico de precisão. Requisitos atuais demandam bem mais que uma simples precisão posicional, exigida anteriormente pelos cartógrafos e suficiente como indicador de qualidade [Gup98].

A definição aqui adotada para qualidade de dados geográficos é a proposta por Chrisman [Chr84]. Segundo ele, qualidade representa o quanto um conjunto de dados se ajusta às necessidades de determinados usuários considerando-se suas aplicações geográficas. Apesar de geral, esta definição de qualidade como *"fitness for use"* tem sido adotada por diversos autores como uma boa interpretação de qualidade de dados.

O resto deste artigo está organizado da seguinte forma. A seção 2 dá exemplos de abordagens típicas de interoperabilidade em SIG no nível de dados. A seção 3 discute parâmetros para avaliação da qualidade de dados geográficos. A seção 4 ressalta como estes parâmetros devem ser considerados no processo de obtenção e conversão de dados geográficos. Finalmente, a seção 5 apresenta conclusões e extensões.

2 Interoperabilidade em SIG no nível de dados

Esta seção utiliza exemplos de abordagens de interoperabilidade em SIG, do ponto de vista de dados, segundo enfoques de modelos [Kel99], sistemas federados [SM99], padrões e metadados [Com94]. Os trabalhos citados foram escolhidos por representarem, cada um, uma faceta do problema. Modelos permitem compartilhamento de dados, em um nível semântico; sistemas federados visam conciliar aspectos de modelos e processamento; e padrões e metadados contemplam o intercâmbio dos dados.

Modelos. O trabalho de [Kel99] trata da criação de um sistema federado, na Suiça, para permitir interoperabilidade entre SIGs heterogêneos, gerenciados por diferentes cantões. Este sistema federado utiliza vários tipos de dados, mas está voltado principalmente a problemas urbanos. O trabalho advoga projeto e desenvolvimento de aplicações SIG usando orientação a objetos como forma de naturalmente induzir a interoperabilidade. Neste sentido, a interoperabilidade seria permitida pelo encapsulamento de objetos e a consequente implementação de processos via trocas de mensagens entre objetos/aplicações, a partir da especificação das suas interfaces. A proposta também aborda a questão de padrões, propondo um padrão denominado INTERLIS, que não parece ser um padrão de intercâmbio, mas sim um conjunto de regras de construção de sistemas.

Sistemas federados. Outra solução para interoperabilidade em SIG é o uso de federações em bancos de dados, considerando-se que estes contêm dados geográficos. Um *sistema de bancos de dados federado* pode ser definido como uma coleção de sistemas de bancos de dados independentes, cooperativos, possivelmente heterogêneos, que são autônomos e que permitem o compartilhamento de todos ou alguns de seus dados, sem afetar as suas aplicações locais [SL90].

Sistemas federados caracterizam-se pela heterogeneidade, presença de dados distribuídos e autonomia de cada componente, ou seja, de cada banco de dados que compõe a federação. A *heterogeneidade* pode ser identificada em diversos níveis e uma das principais causas refere-se às diferenças entre os bancos de dados componentes, como estrutura de dados, nomes, interpretações semânticas dos atributos e restrições.

Há vários outros fatores a considerar, como por exemplo questões relativas à *distribuição* e *localização* dos dados, às técnicas de *recuperação de informação* e aos mecanismos de *segurança*. Vários trabalhos buscam soluções para estes problemas, e no contexto de sistemas geográficos pode-se citar [Agu95, WD91]. As principais soluções estão relacionadas ao uso das seguintes ferramentas: mediadores, tradutores/adaptadores e visões.

Um *mediador* é um software usado para permitir a interoperabilidade entre dois ou mais SGBDs. Com a utilização de mediadores, o acesso aos dados heterogêneos é efetuado através de consultas que são submetidas ao mediador, que por sua vez as transforma em subconsultas a serem enviadas aos SGBDs componentes, onde são executadas. Seus resultados são a seguir combinados e retornados – o que, de novo, apresenta problemas em uma federação onde há vários SIG diferentes, pois raramente é possível integrar dados produzidos por SIG distintos. Sempre que há dados geográficos envolvidos nas consultas locais, o problema fica mais complexo, devido a questões de processamento e otimização de consultas espaciais (por exemplo, [NGS97]).

Já os *Tradutores/adaptadores* convertem os dados fonte para um modelo de dados comum e convertem consultas de aplicações em consultas específicas das fontes de informação envolvidas na consulta [VL97]. Alguns mediadores podem usar tradutores/adaptadores como ferramentas para resolver uma parte específica da conversão de dados. Finalmente, *visões* são usadas como um mecanismo que auxilia a integração dos componentes de um sistema federado, mas seu uso não vem sendo explorado em trabalhos de SIG.

Padrões e metadados. Padrões (por exemplo [Com94]) visam estabelecer formatos e regras de armazenamento que permitam o intercâmbio de dados. Na mesma linha, metadados fornecem notação descritiva para dados armazenados e visam facilitar a construção de consultas e permitir estabelecer correlações de dados em um nível mais abstrato.

Cada vez mais os metadados têm sido incorporados a aplicações em SIG, embora ainda de uma maneira bas-

tante rudimentar, em geral restringindo-se a informações sobre o esquema do banco de dados geográfico. Metadados são frequentemente encontrados em aplicações de bibliotecas digitais, inclusive geográficas, visando não apenas acelerar consultas mas possibilitar reuso [FFLS96]. Outras tendências no gerenciamento de metadados incluem a descrição, em mais detalhes, tanto da história (linhagem) quanto da qualidade de conjuntos de dados e de suas fontes, inclusão de mais elementos espaciais, descrição de modelos e algoritmos.

Em todas estas abordagens visando compartilhamento de dados e inter-operação, o questionamento da qualidade dos dados manipulados é relegado a um segundo plano. O resto deste texto irá detalhar este ponto.

3 Parâmetros para a qualidade de dados geográficos

Dados geográficos comumente podem ser caracterizados a partir de três componentes fundamentais: posição, tema e tempo [Aal96]; ou, de forma equivalente, espaciais, não espaciais e temporais [CCH+96].

As características que afetam a qualidade dos dados geográficos podem ser agrupadas segundo [Aro89] em três categorias: componentes de nível macro, componentes de nível micro e componentes de uso.

Os componentes do nível macro – completude, atualidade e linhagem – consideram dados usando um nível de abstração alto, e têm uma especificação subjetiva. São definidos a partir de exame manual (no caso da completude) ou através de relatórios sobre a obtenção dos dados (no caso da atualidade e linhagem).

Os componentes de nível micro concernem dados individuais e são normalmente avaliados por testes estatísticos confrontados a uma fonte independente de informação de maior qualidade comprovada. Este grupo inclui: precisão posicional, precisão de atributo, consistência lógica e resolução.

Finalmente, os componentes de uso são aqueles que dizem respeito aos recursos de uma organização particular, indicando a adequação dos dados a uma outra organização ou aplicação. Neste grupo podem ser incluídos vários componentes dentre os quais acessibilidade e custo. Componentes de uso são restritos a cada organização ou aplicação e portanto não serão abordados no artigo.

3.1 Componentes do nível macro

A *completude* de um conjunto de dados é avaliada segundo três categorias: completude de cobertura, classificação e verificação. A completude de cobertura é a porcentagem dos dados disponíveis em uma área de interesse (em função de um total estimado). A completude de classificação e a de verificação ajudam a determinar a conveniência de um conjunto de dados para uma dada aplicação. A completude de verificação refere-se à quantidade e distribuição das medidas de campo ou outras fontes independentes de informação utilizadas para validar os dados. A avaliação de completude é usualmente limitada à informação de cobertura. Normalmente informações de completude de classificação e verificação são ignoradas quando se considera a completude de dados geográficos.

A *atualidade* (data ou período de validade) é um fator crítico para a informação geográfica. Ela é normalmente considerada como a data em que o material fonte foi obtido e é também conhecida como precisão temporal. Embora o tempo seja um dos componentes fundamentais dos dados geográficos, a atualidade não é considerada um componente de qualidade micro e sim macro pelo fato deste componente temporal ser normalmente mantido fixo para que se possa avaliar os demais.

A *linhagem* de um dado geográfico é a história de como foi criado, passando por sua coleta e pelos passos de processamento necessários para produzí-lo até o seu armazenamento em um banco de dados geográfico. Assim como a atualidade, a linhagem também não pode ser medida.

Linhagem e atualidade são frequentemente componentes de padrões de metadados geográficos.

3.2 Elementos de nível micro

A *precisão posicional* corresponde à relação entre a posição real de um objeto geográfico e a posição registrada. É usualmente testada pela seleção de uma amostra específica de pontos pré-determinados e comparação destas coordenadas de posição com uma fonte independente de qualidade conhecida. Há dois componentes básicos para avaliação da precisão posicional: o desvio e a precisão. O desvio refere-se às discrepâncias sistemáticas entre a posição representada e a real, sendo normalmente medido pela média dos erros posicionais de cada ponto da amostra. A precisão concerne a dispersão dos erros posicionais dos dados, sendo comumente estimada pelo cálculo do desvio padrão dos erros de posição dos pontos de teste selecionados. Outra forma de medição deste erro, usual em se tratando de coletas de campo e fotogrametria, é o RMS (*Root Mean Square Error*). Ele é calculado a partir da aplicação da função de mínimos quadrados aos erros posicionais relativos aos pontos de teste. Esta medida não faz distinção entre os componentes desvio e precisão da precisão posicional.

A *precisão de atributo* trata dos componentes não espaciais dos dados geográficos. O método para avaliar precisão de atributos cujo domínio é contínuo é similar ao discutido para precisão posicional, enquanto que para variáveis discretas avalia-se a precisão de classificação (que associa a cada faixa de valores de atributos uma determinada classe). As dificuldades na avaliação da precisão de classificação advêm do fato das medidas de precisão serem significativa-

mente afetadas por fatores tais como o número de classes, a forma como os pontos de teste são selecionados e de algumas classes serem confundidas com outras.

A *consistência lógica* se refere à manutenção de regras de cconsistência entre os objetos geográficos, algumas das quais ligadas à semântica de uma aplicação. Um tipo especial de consistência é a que trata de relacionamentos topológicos entre os dados [CDvO93]. Dois conjuntos de dados podem estar corretos quanto ao nível de precisão e assim mesmo não possuírem consistência lógica. Por exemplo, se polígonos adjacentes forem digitalizados por pessoas ou métodos diferentes, a fronteira comum pode ser mapeada com posições ligeiramente diferentes e ainda manter precisão posicional. Quando estes polígonos são integrados em uma mesma aplicação (ou banco de dados) esta diferença pode ser acentuada. Não há padrão para medir consistência lógica, mas normalmente são utilizadas regras baseadas nas restrições de integridade das entidades envolvidas. Estas regras podem avaliar não só os relacionamentos topológicos como também os componentes não espaciais dos dados. Os resultados da avaliação deste parâmetro podem ser informados através de percentuais de satisfação às regras.

O componente espacial é muitas vezes também avaliado por outro parâmetro, a *resolução*, que se refere à menor unidade discernível ou apresentável dos dados. Em se tratando de dados de sensoriamento remoto, usa-se também o termo *resolução espacial*.

4 Qualidade e tratamento dos dados

A seção anterior apontou parâmetros para avaliar a qualidade de dados geográficos e algumas formas de medi-los. No entanto, é preciso levar em conta que os dados armazenados passam por uma sequência de processos, onde sua qualidade pode ser comprometida. Assim sendo, é preciso identificar as fontes de erro durante esta sequência: seleção das fontes de dados, coleta, conversão e armazenamento.

A verificação da qualidade dos dados geográficos começa a partir da verificação da qualidade das fontes de dados (por exemplo, dados em papel ou imagens obtidas de sensoriamento remoto) e técnicas de coleta (por exemplo, em um censo, a metodologia utilizada). Para esta verificação outros parâmetros além dos já citados são utilizados. Alguns deles são: credibilidade, nível de confiança, conveniência, condições físicas, legibilidade e precedência da fonte.

A seguir, é necessário analisar a conversão. A conversão dos dados inclui outros parâmetros específicos como a precisão relativa e a preocupação com a legibilidade e com a simbologia. Nesta etapa, trabalha-se com controle e garantia de qualidade [Hoh98].

O controle de qualidade é um processo de monitoração da qualidade de dados após a conversão e ações corretivas para assegurar que os dados atinjam os padrões predefinidos em projeto. O ideal é projetar o processo para que os erros sejam antecipados e planejar métodos para corrigí-los. O controle de qualidade envolve deteção e correção de erros e subsequente verificação das correções.

A garantia de qualidade é a atividade de verificação final dos dados convertidos antes de serem carregados no banco de dados geográfico. Os dois principais objetivos da garantia de qualidade são monitorar o nível de qualidade final dos dados convertidos e assegurar que todo o processo de controle de qualidade esteja sendo desenvolvido apropriadamente.

Vale ressaltar que padrões de metadados especificam atributos relativos à qualidade (envolvendo em geral linhagem e atualidade e, no caso de imagens de satélite, a cobertura de nuvens – vide [Com94]). Através destes metadados, os usuários podem estimar de forma grosseira o erro de análises espaciais envolvendo os dados correspondentes. Apesar de alguns trabalhos criticarem o uso de metadados para esta finalidade (é o caso de [Har98]) há um consenso de que favorecer o uso de metadados é fator crucial para melhorar tanto a qualidade quanto a disponibilidade dos dados.

5 Conclusões e trabalho em andamento

Este trabalho visa introduzir a noção da importância de medidas de qualidade quando se cuida da interoperabilidade de dados geográficos. A garantia da qualidade dos dados geográficos oferece novas perspectivas sob este ponto de vista, pois permite quantificar determinados aspectos que, por sua vez, indicarão condições necessárias ao intercâmbio e reuso de dados.

Este artigo apresentou de forma resumida alguns dos itens que devem ser levados em consideração ao se dimensionar a qualidade de dados geográficos. Estes parâmetros devem ser levantados e associados a estes dados, de forma a permitir estimar a precisão do resultado de análises espaciais que utilizam estes dados. Além disto, a quantificação de qualidade (por mais subjetiva que seja) permitirá estabelecer critérios mínimos para integração de conjuntos de dados.

Este trabalho é parte de um projeto de mestrado na área de qualidade em bancos de dados geográficos. Atualmente, está sendo desenvolvido um sistema que permite associar diferentes indicadores de qualidade a dados geográficos através de seus metadados. Com isto, usuários de SIG podem avaliar a validade dos resultados obtidos em suas aplicações.

Agradecimentos

Este trabalho foi desenvolvido com apoio de projetos financiados pelo CNPq, pelo projeto MCT-PRONEX SAI (Sistemas Avançados de Informação), pelo convênio CNPq/NSF em interoperabilidade em SIG e pela Marinha do Brasil.

Referências

[Aal96] H. J. G. L. Aalders. Quality Metrics for GIS. In M. J. Kraak and M. Molenar, editors, *Advances in Gis Research II*, pages 5B1–5B10. 7th International Symposium on Spatial Data Handling, Delft University, August 1996.

[Agu95] C. D. Aguiar. Integração de Sistemas de Banco de Dados Heterogeneos em Aplicações de Planejamento Urbano. Master's thesis, UNICAMP, march 1995.

[Aro89] S. Aronoff. *Geographic Information Systems*. WDL Publications, Canada, 1989.

[CCH+96] G. Camara, M. Casanova, A. Hemerly, G. Magalhaes, and C. Medeiros. *Anatomia de Sistemas de Informação Geográfica*. 10 Escola de Computação, 1996.

[CDvO93] E. Clementini, P. DiFelice, and P. van Oosterom. A Small Set of Formal Topological Relationships Suitable for End-user Interaction. In *Proc Third Intl. Symp. Spatial Databases - SSD*, pages 277–295, 1993.

[Chr84] N. R. Chrisman. The role of quality information in the long term functioning of a gis. *Cartographica*, 21(2):79–87, 1984.

[Com94] Federal Geographic Data Committee. Content Standards for Digital Geospatial Metadata. http://geochange.er.usgs.gov/pub/tools/metadata/standard/metadata.html, 1994.

[Cro79] P. B. Crosby. *Quality ie Free - The Art of Making Quality Certain*. McGraw-Hill, 1979.

[Dem86] W. E. Deming. *Out of Crisis*. Center for Advanced Engineering Study - MIT - Cambridge, 1986.

[FFLS96] C. Fischer, J. Frew, M. Larsgaard, and T. Smith. "alexandria digital library:rapid prototype and metadata schema". *Lecture Notes In Computer Science*, 1082:221–241, 1996.

[Gup98] S. C. Guptill. Building a Geospatial Data Framework - Finding the Best Available Data. In Michael Goodchild and Robert JeanSoulin, editors, *Data Quality in Geographic Information - From Error to Uncertainty*, chapter Quality Concepts and Models, pages 31–36. Hermes, 1998.

[Har98] F. Harvey. Quality needs more than standards. In Michael Goodchild and Robert JeanSoulin, editors, *Data Quality in Geographic Information - From Error to Uncertainty*, chapter Quality Concepts and Models, pages 37–42. Hermes, 1998.

[Hoh98] P. Hohl. *GIS Data Conversion; Strategies, Techniques and Management*. Onword Press, 1998.

[Int97] Interop. International Conference and Workshop on Interoperating Geographic Information Systems. Web site http://www.ncgia.ucsb.edu/conf/interop97, 12 1997. Site address valid as of 07/99.

[Kel99] S. Keller. Modeling and sharing geographic data with INTERLIS. *Computers and Geosciences*, 25(1):49–59, 1999.

[NGS97] A. Newton, B. Gittings, and N. Stuart. Designing a Scientific Database Query Server using the World Wide Web: the Example of Tephrabase. In Zarine Kemp, editor, *Innovations in GIS 4*, pages 251–265. Taylor and Francis, 1997.

[SL90] A. Sheth and J. Larson. Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases. *ACM Computing Surveys*, 22(3):183–236, 1990.

[SM99] H. R. Soares and C. B. Medeiros. Integrando Sistemas Legados a Bancos de Dados Heterogeneos. In *Anais, XIV Simposio Brasileiro de Bancos de Dados*, 1999.

[VL97] V. M. P. Vidal and B. F. Lóscio. Especificação de mediadores para acesso e atualização de múltiplas bases de dados. *Anais XII Simpósio Brasileiro de Banco de Dados*, 1997.

[WD91] M. Worboys and S. Deen. Semantic Heterogeneity in Distributed Geographic Databases. *ACM Sigmod Record*, 20(4):30–34, 1991.

[YB98] P. Yates and I. Bishop. The Integration of Existing GIS and Modelling Systems: with Urban Applications. *Computers, Environment and Urban Systems*, 22, 1998.