Aplicando Procedência de Dados para a Geração e Análise de Perfis de Doadores de Sangue

Fernanda N. Almeida[1], Pedro P. S. B. Silva[1], Alfredo M. Júnior[2], Ester C. Sabino[2], João E. Ferreira[1]

[1]Instituto de Matemática e Estatística – Universidade São Paulo (USP)
Rua do Matão, 1010, Bloco A, Butanta, 05508-900, São Paulo – SP, Brasil

[2]Fundação Pró-Sangue – Hemocentro de São Paulo, Av. Dr. Enéas de Carvlho Aguiar, 155 – 1º andar, Cerqueira César, 05403-000, São Paulo – SP, Brasil

{falmeida,jef}@ime.usp.br, {pedro.paulo.sbs,sabinoec}@gmail.com, alfredo.mendrone@prosangue.sp.gov.br

Abstract. *Regular blood donors are frequently prevented from donating due to potential risk of anemia, and the reasons that lead to these rejections may vary according to each individual. The São Paulo Blood Center collects data from donations and also from rejections. One of the main reasons for rejection is iron deficiency. In this paper, we are interested in constructing a data provenance model and applying it to the São Paulo Blood Center database. Our contribution is the creation of a data provenance model that is able, not only to recognize potentials profiles of anemic donors, but also verify the reliability of the extracted information from the data set.*

Resumo. *O grupo de doadores regular de sangue possui um índice considerável de recusas por motivos de anemia, e as razões que levam a isto variam de acordo com o indivíduo. O Hemocentro de São Paulo registra tanto dados das dações quanto das recusas. Um dos principais motivos de recusa é a anemia ferropriva (ou perda de ferro no sangue). Neste trabalho, estamos interessados em construir um modelo de procedência de dados e aplicá-lo ao banco de dados do Hemocentro de São Paulo. Nossa contribuição é a criação de um modelo de procedência de dados que seja capaz não somente de identificar os potenciais perfis para os doadores anêmicos, mas também verificar quão confiáveis são as informações extraídas do conjunto de dados.*

1. Introdução

As tarefas dos experimentos *in silico* começaram a ser disseminadas e executadas em larga escala pela comunidade científica. Isso aumentou a necessidade da quantidade e capacidade de recursos computacionais bem como a necessidade de gerenciar dados de fontes heterogêneas e de qualidade variável.

Com o objetivo de caracterizar a qualidade dos dados nos experimentos *in silico*, optamos por usar a abordagem de procedência de dados [Buneman 2000]. Essa abordagem será empregada não somente para identificar a origem de um determinado conjunto de dados, mas também, para auxiliar a formação de uma visão mais detalhada sobre a qualidade, viabilidade e confiabilidade do dado. Uma das principais motivações

para o desenvolvimento deste trabalho é a caracterização do comportamento de doadores com grande probabilidade de desenvolver anemia. Para isso, desenvolvemos um modelo de procedência de dados baseado no contexto das doações de sangue da Fundação Pró-Sangue/Hemocentro de São Paulo (*www.prosangue.sp.gov.br*). Essa Fundação possui um banco de dados com dados referentes a cada doação de sangue, que são provenientes dos registros de triagens e doações realizadas pelos doadores. As triagens são formulários submetidos ao doador no ato da doação de sangue e que contém basicamente: dados pessoais do potencial doador, motivos da doação e informações da bolsa de sangue (quando houver). As doações, além dos dados clínicos também armazenam resultados de exames, tais como: chagas, sífilis, hepatite, HTLV (vírus t-linfotrópicos humanos) e HIV (vírus da imunodeficiência humana).

Após fazermos uma análise geral no banco de dados verificamos que alguns erros de digitação foram introduzidos, como, por exemplo, troca de sexo, falta do preenchimento de alguns campos, dados inválidos, dentre outros. Inicialmente tentamos fazer agrupamentos, relacionando sexo, idade, quantidade de doações e níveis de medição de hemoglobina no sangue (medidos pelos testes de micro-Hematócrito e hemoglobina) para identificar possíveis padrões de comportamento nos dados. No entanto, os resultados obtidos foram muito confusos. Observamos apenas nuvens de dados.

Com o modelo de procedência de dados proposto neste trabalho, pretendemos mostrar que é possível fazer investigações detalhadas no banco de dados sem que haja necessidade de se corrigir sua estrutura e seus dados. Nossa intenção é mostrar que um modelo de procedência de dados pode ser ajustado a um banco de dados e que a complexidade e/ou dificuldade de implementação está mais voltada para os erros introduzidos no conjunto de dados do que na sua estrutura. O modelo de procedência proposto aqui se aplica somente a um banco de dados de fonte conhecida. Nosso principal interesse é desenvolver um modelo de procedência que possa ser aplicado a outros bancos de dados dentro do domínio de doações de sangue.

2. Trabalhos Relacionados

O principal objetivo da procedência é recolher os indícios quanto ao tempo, lugar e, se for o caso, a pessoa responsável pela criação, produção, descoberta ou inserção do dado. Técnicas de análise comparativa, pareceres de peritos e os resultados de vários tipos de testes científicos também podem ser utilizados para este fim, mas o que institui de fato a procedência é a sua documentação.

Existem diferentes definições para o termo “procedência de dados” [Buneman 2000]. Segundo Simmhan e colaboradores (2005), procedência de dados refere-se aos processos de investigação e armazenamento da origem de uma parte do dado, bem como sua movimentação entre os bancos de dados. É a documentação complementar de um determinado dado que contém a descrição de “como”, “quando”, “onde” e “por que” ele foi obtido e “quem” o obteve [Buneman 2001]. Nesta mesma linha estão Woodruff e colaboradores (1997), afirmando que procedência não somente inclui a origem do dado (identificação, responsável pelo dado, data de criação), mas também os processos aplicados a ele (algoritmos e seus respectivos parâmetros). Já Lanter (1991), se refere à linhagem do dado derivada do um produto (empregado a Sistemas de Informação

Geográficos) como, informações que descrevem os materiais e as transformações aplicadas para produzir o dado [Lanter 1991]. Nesse caso, a procedência não somente está associada com o produto de dados, como também aos processos de criação e viabilização do dado [Buneman 2000]. Essa definição de procedência descrita por Lanter (1991) foi expandida por Greenwood (2003) que vê a utilização de um processo para registrar metadados de experimentos com *workflow*, anotações e apontamentos sobre experimentos feitos por pesquisadores [Buneman 2000].

Com as diversas definições existem sinônimos que são utilizados para denotar procedência de dados, dentre eles os mais comuns são: "linhagem" [Bose 2002], "proveniência" ou "*pedigree*" [Brown and Stonebraber 1995], e também "histórico de derivação" [Hachem 1993]. Em uma abordagem diferente proposta por Buneman (2001), o termo "procedência" é utilizado para fazer referência somente à origem do dado e o termo "*pedigree*" para fazer referência ao histórico de como aquele dado foi produzido.

Apesar da importância de se compreender a origem dos dados, poucas são as tentativas feitas para traçar um modelo de procedência em bancos de dados científicos. A maioria das pesquisas visam o armazenamento da procedência de dados provenientes da *Web* ou bibliotecas digitais.

Na pesquisa em banco de dados, a procedência auxilia no processo de rastreamento e sinalização da origem dos dados, como também sua movimentação entre as diferentes fontes de dados [Simmhan 2005]. Com o declínio dos custos de armazenamento de dados em *hardware*, tornou-se comum à integração de bancos de dados, ou seja, construir um novo banco de dados usando outros já existentes. Parte da solução para recordar a origem dos dados é anexar a procedência como componente dos bancos de dados [Buneman 2001]. O gerenciamento destes dados traz novos problemas à comunidade científica de banco de dados, uma vez que a tecnologia de banco de dados convencionais não pode ser diretamente aplicada [Buneman 2001]. Além disso, de acordo com Buneman (2001), a inserção de novos dados requer métodos de desenvolvimento mais sofisticados. Contudo, alguns esforços importantes para incluir procedência em banco de dados estão em andamento [Buneman 2001].

Como uma tentativa de solucionar parte desses problemas de procedência em Bioinformática, o projeto *myGrid* [Greenwood 2003; Zhao 2003] tem como principal objetivo a construção, administração e distribuição intensa de dados provenientes de experimentos biológicos *in silico*. Os autores acreditam que os dados provenientes de experimentos executados em laboratórios de bancada associados com experimentos *e-science* possuem pouco valor se outros cientistas não forem capazes de armazenar a procedência desses dados [Greenwood 2003]. A intensão do projeto *myGrid* é potencializar o valor dos experimentos em *e-science*, ou seja, o uso de recursos eletrônicos, como banco de dados, métodos computacionais, dentre outros. Para entender melhor o funcionamento desse projeto, suponha um dado biológico como sendo uma seqüência de DNA. Nesse caso, o *myGrid* gerará a procedência desse dado, que fornecerá: notas de onde o dado vem, informação sobre espécies a que pertencem, o responsável pela entrada dos dados, quando o dado foi inserido, sintaxe, semântica, dentre outras informações [Greenwood 2003]. Depois de gerada a procedência, ela será armazenada e compartilhada na *Web* por vários outros grupos de pesquisa. O projeto

incluiu não apenas a procedência, mas também atalhos para outros projetos (direta ou indiretamente relacionados), indicativos de páginas da *Web*, referências bibliográficas, dentre outros [Greenwood 2003].

3. Cenário de Aplicação

Mensalmente, a Fundação Pró-Sangue/Hemocentro de São Paulo envia ao nosso grupo de Banco de Dados, localizado no Instituto de Matemática e Estatística da USP, dados do seu banco transacional. Por mês, são enviadas em média 13 mil triagens, o que equivale a aproximadamente 200MB de dados submetidos e armazenados em nosso grupo de Banco de Dados. Assim, a Fundação Pró-Sangue é o hemocentro brasileiro com o maior fluxo de doações. Esses dados são conjuntos de arquivos texto contendo as informações coletadas nos procedimentos da doação. Esses procedimentos inicialmente consistem em coletar os dados pessoais do candidato à doação, o qual recebe um código que o acompanha durante todo o processo de doação. Em seguida, um teste de anemia é feito (teste de micro-Hematócrito e/ou Hemoglobina) para verificar se o candidato à doação esta apto a doar, ou seja, se possui níveis de hemoglobina dentro do aceitável. São verificados também os batimentos cardíacos, pressão arterial e o peso do indivíduo. Após estas etapas o candidato responde a uma entrevista confidencial, com o objetivo de avaliar se a doação pode trazer riscos para ele ou para o receptor, este processo é chamado de triagem clínica.

Em 1996, as triagens deixaram de ser arquivadas em repositórios comuns e passaram a ser digitalizadas. Atualmente, temos armazenado dados do período de janeiro de 1996 a outubro de 2009. Em 2007, houve uma mudança nos questionários da triagem, sendo introduzidas perguntas para se obter informações mais detalhadas sobre o candidato. Tais alterações dividiram nosso conjunto de dados em dois subgrupos distintos relacionados às informações da triagem: anteriores e posteriores a 2007.

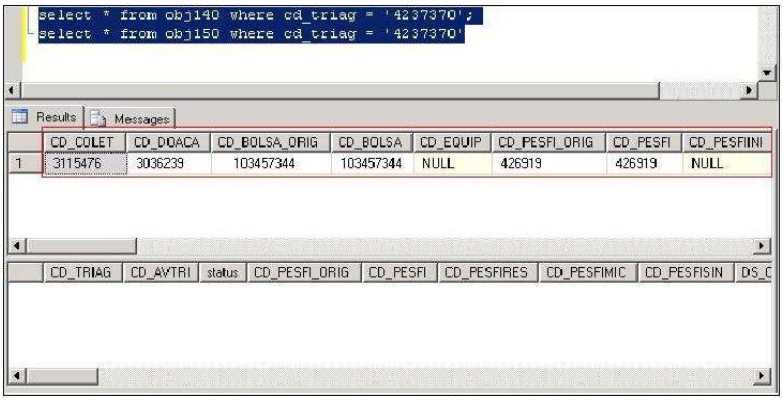

Figura 1. Exemplo de inconsistência. Neste caso, um indivíduo esta presente na tabela de doadores (*'cd_doacao'*), no entanto, não há registro desse doador na tabela de triagens (*'cd_triag'*) selecionada em azul. Isso não pode acontecer já que todos os indivíduos, obrigatoriamente, antes de efetuar a doação passam pela triagem.

Como a inserção dos dados das triagens na Fundação Pró-Sangue ainda é digitada pelos funcionários e o formulário apresenta campos de texto de livre edição, decidimos avaliar a qualidade, consistência e confiabilidade dos dados armazenados. Além disso, esses bancos de dados em sua maioria não são normalizados,

conseqüentemente mais vulneráveis a inconsistências. Devido à possibilidade de que erros de digitação e problemas de falta de consistência herdados das fontes comprometessem a confiabilidade nos dados, desenvolvemos uma seqüência de procedimentos que tem por objetivo garantir sua qualidade. Para exemplificar esses dois problemas, de confiabilidade e de inconsistência, foi utilizado um trecho real extraido dos arquivos enviados pela Fundação Pró-Sangue (Figura 1).

4. Modelo de Procedência de Dados

Como vimos na sessão II, diferentes abordagens foram desenvolvidas para atender aos requisitos individuais de procedência de dados. Entretanto, as metodologias apresentadas anteriormente abordam a procedência com a mesma finalidade, a de armazenar a origem de um dado. Aqui apresentaremos as técnicas para o desenvolvimento e implementação do modelo de procedência proposto neste trabalho, com o qual não somente iremos expor o modelo no âmbito descritivo de dados, como também, voltaremos nossa abordagem ao domínio de aplicação e, nos principais questionamentos biológicos de interesse desta pesquisa. Assim, trataremos o modelo de procedência empregado, desde a obtenção do conjunto de dados até seu tratamento e posteriores análises em que aplicaremos algumas técnicas de análise de sobrevivência que melhor se ajustarão ao nosso conjunto de dados.

A tabela de triagens disponibilizada pelo hemocentro contém 23 atributos, no entanto, a maioria deles não esta diretamente relacionada com o tema deste artigo. Assim, foram selecionados somente aqueles que têm alguma correlação com este trabalho, são eles: sexo, raça, escolaridade, tipo de doação, data da visita ao hemocentro, peso, altura e nível de hemoglobina no sangue (medidos pelos testes de micro-Hematócrito — micro-Hct e Hemoglobina — Hg). A partir disso, fizemos um cuidadoso levantamento da qualidade dos dados de cada um dos atributos listados e chegamos no resultado apresentado na Tabela 1. Com esses dados em mãos foi possível projetar e implementar rotinas de tratamento e normalização dos dados. Por exemplo, foram considerados erros do sistema entradas de triagens que tiveram porcentagem de hematócrito superior a 50% ou inferior a 30% ou que tiveram nível de hemoglobina superior a 18g/dL ou inferior a 10g/dL. Em nossa análise descritiva, observamos que para os mais de dois milhões de triagens analisadas, aproximadamente 218 mil registros tinham valores inválidos no atributo peso. Apesar disto representar apenas 7,5% do total do banco de dados (que tem mais de 2,5 milhões de dados registrados), optamos pela exclusão desta variável devido ao fato do potencial doador não passar pela pesagem, tornando esse um dado muito subjetivo. O mesmo acontece para altura, raça e escolaridade. No que diz respeito ao atributo referente ao nível de hemoglobina no sangue, a rotina gerada excluiu os campos inválidos como, por exemplo, registros de micro-Hct igual a 99 e Hg igual a 99,99. Verificamos que aproximadamente 0,89% do total apresentaram registros nulos para estas variáveis. Isto viabilizou a remoção das triagens somente quando os dois campos (micro-Hct e Hg) apresentarem os valores inválidos.

Com a classificação das variáveis do banco de dados de doadores de sangue em níveis de qualidade e confiabilidade (Tabela 1), selecionamos somente aqueles que tiveram porcentagem maior ou igual a 80%. Entretanto, o atributo '*peso*' apesar de apresentar um índice elevado de qualidade foi desconsiderado por ser uma informação

subjetiva, ou seja, o indivíduo não é pesado. Excluímos também de nossas análises o atributo '*tipo de doação*' por entendermos que ele não possui, no caso deste trabalho, interferência direta na doação de sangue. Nossa intenção foi, a partir da investigação de variáveis com índices altos de qualidade e confiabilidade tentar identificar e caracterizar quais são os diferentes grupos de doadores no Hemocentro de São Paulo. Por isso, é fundamental contarmos somente com atributos extremante confiáveis, pois deste modo podemos nos aproximar daquilo que acontece após contínuas doações.

Tabela 1. Qualidade e confiabilidade dos atributos pré-selecionados para análise.

Atributo	Qualidade	Confiabilidade
Sexo	100,00%	alta
Raça	52,00%	baixa
Escolaridade	4,50%	baixa
Tipo de doação	80,00%	alta
Dia da visita	100,00%	alta
Mês da visita	100,00%	alta
Ano da visita	100,00%	alta
Peso	92,50%	baixa
Altura	(***)	baixa
Hematócrito	90,00%	alta
Hemoglobina	81,00%	alta

(***) Não havia registros suficientes para mensurar os dados.

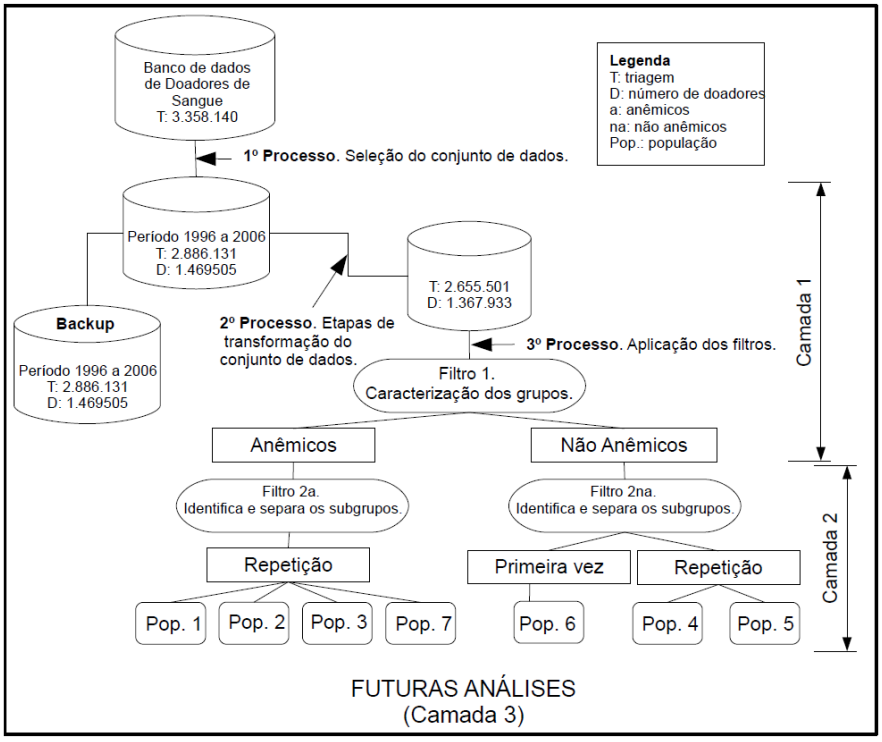

Figura 2. Modelo de procedência de dados desenvolvido especificamente para o domínio de doações de sangue da Fundação Pró-Sangue.

De mão dos atributos selecionados demos início ao desenvolvimento de filtros baseados em perguntas feitas pelos pesquisadores da Fundação Pró-Sangue. Nomeamos

essas perguntas de camadas, pois foi a partir delas que geramos os filtros e conseqüentemente novas visões do conjunto de dados que estamos analisando. Temos duas camadas principais: (*i*) qual é a população de doadores que pode ser considerada de risco para o desenvolvimento de anemia ferropriva (falta de ferro no sangue) decorrente de múltiplas doações de sangue? e; (*ii*) qual é o número de doações de sangue e o intervalo entre estas doações que aumenta a probabilidade de deixar o(a) doador(a) com risco de desenvolver anemia?. Como o foco deste trabalho é a caracterização e análise dos doadores que possuem mais chances de desenvolver anemia daqueles que não a tem, separamos os doadores em dois tipos: (i) doadores anêmicos — são aquelas pessoas que foram impedidas de doar por apresentarem nível de hemoglobina no sangue abaixo do permitido pelo Ministério da Saúde e; (ii) doadores não anêmicos — aqueles que sempre apresentarão níveis aceitáveis. Assim, construímos vários filtros tais como: identificação e separação dos sub-grupos; caracterização das populações de doadores. Detalhes da construção dos filtros e do fluxo de caracterização dos dados podem ser visto na Figura 2.

5. Discussão e Conclusão

Como primeiro resultado preliminar deste trabalho conseguimos definir, com o auxílio do modelo de procedência de dados, sete populações distintas de doadores de sangue descritas na Tabela 2. Dessas, fomos capazes de identificar as populações 3 e 4, que são as que mais se aproximam da realidade deste estudo e que estão atualmente sendo analisadas. No entanto, até o presente momento o método de procedência utilizado nos experimentos iniciais se mostrou robusto e eficiente, gerando grupos de doadores curados.

Tabela 2. Definição das populações de doadores de sangue.

Populações	**Doações**		**Doador (*N*)**
	Primeira Doação	**Próximas Doações (*n* vezes)**	
População 1	Recusado por apresentar risco de desenvolver anemia	Depois de duas ou mais doações, o doador não foi mais recusado por apresentar risco de desenvolver anemia.	13180
População 2	Recusado por apresentar risco de desenvolver anemia	Depois de duas ou mais doações, o doador voltou a ser recusado por apresentar risco de desenvolver anemia.	3690
População 3	Aceito para doação	Depois de duas ou mais doações, o doador voltou a ser recusado por apresentar risco de desenvolver anemia. *(*)*	31970
População 4	Aceito para doação	Depois de duas ou mais doações, o doador pode também ter sido recusado para doação por motivos diferente de risco de desenvolver anemia.	344980
População 5	Recusado para doação por motivos diferentes de risco de desenvolver anemia	Recusado em outras tentativas de doação de sangue por razões diferente de risco de desenvolver anemia.	11350
População 6	Aceito para doação uma única vez	O doador pode ter voltado repetidas vezes, mas foi aceito para doação de sangue uma única vez.	846930
População 7	Este grupo inclui todos os doadores que doaram no período de 1996 – 2002. Nesse período, baseado nas Leis Federais Brasileiras, o critério de corte para determinação de risco de desenvolver anemia era menor do que o considerado atualmente (a partir de 2003).		12195 *(**)*

(*) Primeiro registro de risco de anemia após a primeira doação; (**)11274 mulheres e 921 homenens.

Como próxima etapa empregaremos técnicas de análise de sobrevivência para afirmar com clareza o quão confiável são os dados gerados (a partir do nosso modelo) e quais informações poderemos obter das análises desses dados. Temos interesse em analisar quais os fatores de risco ou de prognósticos (sejam eles quantitativos ou qualitativos) no tempo de sobrevida de um doador ou de um grupo de doadores, bem como definir as probabilidades de sobrevida em diversos momentos no seguimento da doação.

References

Bose, R. A Conceptual Framework for Composing and Managing Scientific Data Lineage. In: International Conference on Scientific and Statistical Database Management, 14., July, 2002, Edinburgh, Scotland. Proceedings. 2002. p. 15-19.

Brown, P.; Stonebraker, M. A System for the Management of Earth Science Data. In: International Conference of Very Large Data Bases, 21., 1995, Zurich, Switzerland. Proceedings. 1995. p. 720-728.

Buneman, P.; Khanna, S.; Tan, W. Data Provenance: Some Basic Issues. In: Foundations of Software Technology and Theoretical Computer Science (FST TCS), December 13-15, 2000, New Delhi, India. Proceedings. Springer-Verlag, 2000. v. 1974.

Buneman, P.; Khanna, S.; Tan, W. Why and Where: A Characterization of Data Provenance. In: International Conference on Database Theory, 4-6 January, 2001, London, United Kingdom. Proceedings. 2001. p. 15. On-line.

Greenwood, M.; Goble, C.; Stevens, R.; Zhao, J.; Addis, M.; Marvin, D.; Moreau, L.; Oinn, T. Provenance of e-Science Experiments - Experience from Bioinformatics. In: UK OST e-Science second All Hands Meeting 2003. (AHM'03), 2-4 Sept, 2003, Nottingham, UK. Proceedings. 2003. p. 4.

Hachem, N.; Gennert, M.; Ward, M. A Spatio-Temporal Database System for Global Change Studies. In: AAAS Workshop on Advances in Data Management for The Scientist and Engineer, February, 1993, Boston, Massachusetts. Proceedings. 1993. p. 84-89.

Lanter, D. P. Design of the Lineage-Based Meta-Data Base for GIS. In:Cartography and Geographic Information Systems, vol. 18, 1991.

Simmhan, Y. L.; Plale, B.; Gannon, D. A Survey of Data Provenance Techniques. In: Technical Report TR-618: Computer Science Department, Indiana University, 2005.

Woodruff, A.; Stonebraker, M. Supporting Fine-Grained Data Lineage in a Database Visualization. In: International Conference on Data Engineering, 13. 7-11 Apr. 1997, Birmingham, UK. Proceedings. 1997. p. 15.

Zhao, J.; Goble, C.; Greenwood, M.; Wroe, C.; Stevens, R. Annotating, linking and browsing provenance logs for e-Science. In: Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Oct. 2003, Oxford Road, Manchester. Proceedings. 2003. p. 6.