1
Introdução

Não há ciência sem proveniência. Não é possível dar crédito a um experimento se não há como repetí-lo. Em ciência, assim como em muitos outros campos, as informações relativas às metodologias e mecanismos pelos quais os resultados são obtidos podem ser tão importantes quanto os resultados propriamente ditos.

A identificação dos agentes envolvidos também é essencial. Por exemplo, em um leilão de obras de arte, as informações sobre o atual proprietário, bem como os proprietários anteriores, são extremamente relevantes para os potenciais compradores. Analogamente, em uma caixa postal abarrotada de mensagens eletrônicas, o remetente de uma mensagem é sem dúvida um filtro que determina a ordem de leitura.

Esses e outros dados relacionados estão presentes ao longo do ciclo produtivo acadêmico ou de negócios e são de fundamental importância porque descrevem a origem dos dados. São metadados e representam a proveniência desses dados. O rastro desses dados gera um conjunto de metadados de proveniência intrínsecos ao ciclo produtivo. A rastreabilidade dos dados aparece como um requisito cada vez mais presente nos sistemas de informação atuais, refletindo a crescente demanda por conformidade e confiabilidade das fontes produtivas.

Por outro lado, a interoperabilidade também é característica essencial e necessária ao planejamento de um sistema de informação. A capacidade de interoperar está diretamente ligada à estratégia adotada para a concepção do respectivo modelo do banco de dados. Se um sistema de informação é bem sucedido, isso significa que atende qualitativamente e quantitativamente a seus usuários. Mas, para ser estratégico, deve ser capaz de interoperar com outros sistemas existentes e futuros. Nesse caso, a adoção de esquemas (de exportação dos dados) padronizados seria uma solução mais plausível do que o alinhamento à posteriori dos esquemas. Porém, surpreendentemente, existe um considerável número de padrões que a comunidade de banco de dados parece ignorar durante a construção de novos modelos.

Neste capítulo, apresentamos inicialmente a motivação desta dissertação (seção 1.1). Em seguida, descrevemos o problema de pesquisa abordado que combina as questões de proveniência e interoperabilidade e, um resumo da

solução proposta para o problema estabelecido (seção 1.2). Detalhamos a metodologia para esta pesquisa (seção 1.3), apresentamos o objetivo (seção 1.4) e, por fim, a estrutura desta dissertação (seção 1.5).

1.1. Motivação

A Figura 1 (Bechini et al., 2005) ilustra a cadeia produtiva de lotes (de um produto qualquer) e as atividades relacionadas à produção dos lotes. Os pequenos retângulos verticais são atividades e os círculos são lotes. Cada par (atividade, lote) pertence a um agrupamento que identifica uma etapa da cadeia produtiva. Esses agrupamentos, localizados no topo da Figura 1, são: Fornecedor, Produtor, Transformador ou Distribuidor. Neste exemplo, uma atividade precede sempre um lote que pode estar ligado também à outra atividade da cadeia produtiva. Interessa-nos rastrear os lotes que estejam relacionados a um incidente com um determinado lote, destacado no canto superior direito.

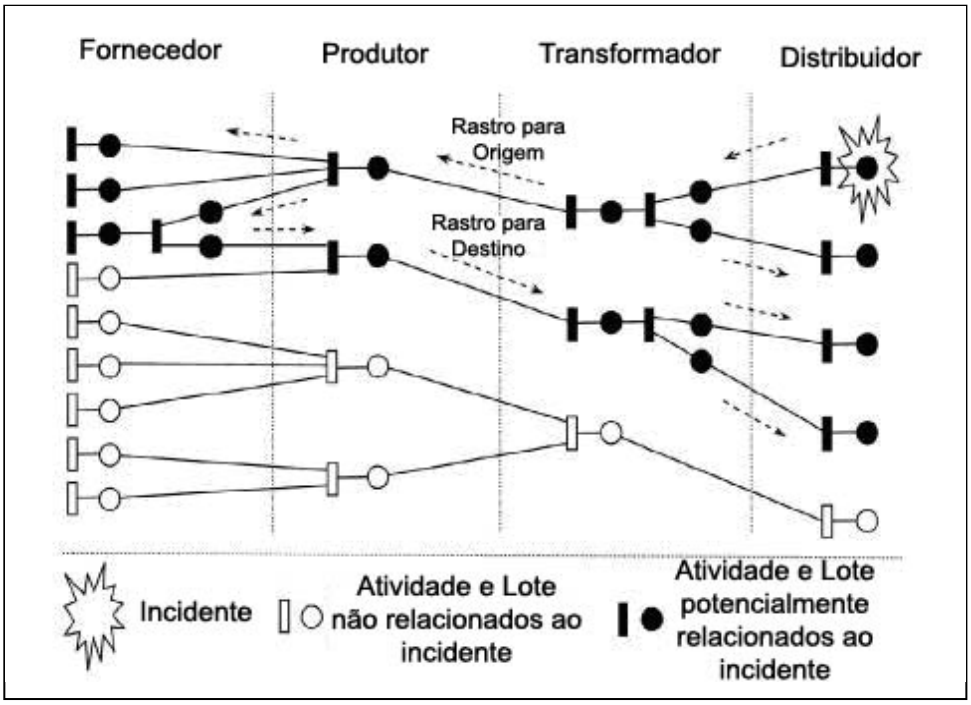

Figura 1: Rastreabilidade na Cadeia Produtiva baseado em (Bechini et al., 2005)

Rastrear um lote em um ponto específico da produção é determinante na identificação de outros lotes de materiais que foram utilizados como insumos.

Isso se torna crítico para a análise do impacto ocasionado, por exemplo, pela contaminação de uma cadeia de produção de alimentos, ou mesmo pelo *recall* de veículos que apresentam problemas no sistema de freio. Dentro dessa mesma cadeia é indispensável a identificação de participantes (determinantes ou aderentes) como pessoas, organizações, agentes de software entre outros e, permite associá-los a atividades, eventos ou processos. Pode ser utilizada para estabelecer níveis de confiança para as transformações dos dados.

Na literatura, é comum encontrarmos o conceito de rastreabilidade aplicado a objetos físicos, como é o caso na Figura 1, que trata de rastreabilidade de alimentos (*food traceability*). Por outro lado, o termo proveniência é usualmente utilizado para descrever a origem de dados. Interessa-nos estabelecer aqui que a capacidade de rastrear dados depende diretamente da existência da proveniência. Para compreender essa dependência, faremos uma analogia entre proveniência de dados e rastreabilidade de lotes físicos de materiais.

Se ao longo da cadeia produtiva há o registro de todas as atividades que aconteceram, é possível rastrear cada lote relacionado ao incidente. Se juntamente a essas atividades há também o registro de dados que descrevem a origem de cada lote, então rastrear os dados de um determinado lote é consultar dados que descrevem a origem de cada lote.

Interpretamos *rastreabilidade* como a capacidade de realizar consultas para rastrear a origem (*trace*) e rastrear o destino (*track*) (Dorp, 2004). Na Figura 1 as setas pontilhadas identificam a direção dessas consultas na cadeia produtiva. Ao seguir as setas a partir do lote que simboliza o incidente é possível rastrear os dados de todos os lotes relacionados.

Portanto, rastrear dados é realizar consultas a proveniência. Essas consultas não são uma exclusividade da área industrial e estão presentes também em outras áreas. Por exemplo, na área de gerência de configuração de software (Staa, 2003), especificamente no versionamento de software, localizar o rastro para a origem identifica possíveis causas de um componente defeituoso. De forma complementar, identificar o rastro para o destino viabiliza a análise de impacto de componentes construídos a partir de um componente potencialmente defeituoso. Assim, aqui também é essencial consultar a proveniência sobre as diferentes versões de componentes existentes.

1.2. Problema

Nesta seção, introduziremos o problema de modelar proveniência de um ponto de vista comum a diferentes áreas de aplicação. De acordo com a motivação exposta (seção 1.1), a rastreabilidade dos dados é uma necessidade essencial nos dias de hoje e depende da capacidade de realizar consultas a proveniência. Essas consultas por sua vez dependem da coleta e do armazenamento dos metadados de proveniência ao longo de uma cadeia produtiva.

Não há atualmente um modelo de proveniência, proposto a partir de padrões existentes, que descreva uma forma genérica para esse armazenamento. Por outro lado, há diferentes áreas que se beneficiariam de um modelo deste tipo. Entre elas estão:

- Engenharia de Software, na demanda por rastreabilidade de requisitos, por exemplo, identificando a proveniência de artefatos que não possuam requisitos;
- e-Science, na necessidade de repetição de experimentos, por exemplo, com a verificação dos métodos e valores utilizados e obtidos durante o processamento de um fluxo de experimentos;
- Industrial, na imposição de conformidade através de legislações vigentes, por exemplo, com a capacidade de consulta a proveniência dos eventos de uma cadeia produtiva atendendo a normas governamentais para o setor de produção de alimentos;
- Gestão de Conteúdo Digital, na auditoria e histórico de *logs*, por exemplo, na identificação do editor e outros dados de proveniência que descrevam o contexto das alterações de um documento digital;
- Preservação Digital, na preservação de documentos arquivísticos a longo prazo, por exemplo, na identificação dos eventos de arquivamento para garantir e creditar a autenticidade da fonte produtora que repassa esses documentos para arquivamento;
- Gestão de Projetos, na análise de tarefas realizadas contra as planejadas, por exemplo, através de consultas a proveniência de eventos considerados particularmente importantes (*milestones*) e rastro dos dados relativos às tarefas que contribuíram para o seu alcance.

Um modelo é uma representação do que existe no intelecto humano como solução para um determinado problema. Como cada ser humano tem uma forma diferente de interpretar o mundo, existem diferentes modelos para o mesmo tipo de problema (Sowa, 1999).

Dois modelos diferentes que propõem soluções para a representação de proveniência são sugeridos pelos projetos PASOA (Moreau & Ibbotson, 2006) e Data Provenance (Ram, 2007). Há ainda diversos outros projetos, avaliados em (Simmhan et al., 2005) e (Bose & Frew, 2005), que também estudam soluções para o problema de proveniência e propõem modelos distintos daqueles dois.

Conforme discutido em (Simmhan et al., 2005), a maioria dos projetos pesquisados possui protocolos proprietários para gerenciar a proveniência e não fundamentam a sua coleta, representação, armazenamento e consulta em padrões abertos, o que, por sua vez, dificulta a interoperabilidade.

É justamente essa realidade que se apresenta como pano de fundo para a construção de novos modelos. A necessidade de interoperabilidade entre sistemas é um desafio para muitas áreas, mas é a comunidade de banco de dados, a partir da adoção de uma maior disciplina, que pode oferecer uma solução a priori.

Adiantando um pouco os objetivos apresentados mais adiante (seção 1.4) descrevemos que, um resumo da solução para esse problema, deve partir de um estudo de padrões para construir um modelo de proveniência genérico. Acrescentamos que a utilização de padrões é uma das formas de promover a interoperabilidade dos sistemas de informação porque torna o problema de alinhamento de esquemas (*schema matching*) um problema tratável (Casanova et al., 2007).

Inicialmente aprofundaremos a compreensão do conceito de proveniência e a identificação de outros conceitos que auxiliam a sua representação. Adaptamos a metodologia para a construção de padrões de projeto conceitual de ontologias para projetar um modelo genérico de proveniência, criado com base no alinhamento de recortes de ontologias de alto nível, padrões internacionais, projetos e propostas de padrões que tratam direta ou indiretamente de conceitos relacionados à proveniência.

A partir da identificação de um conjunto de termos que capturam o conceito de proveniência faremos uma análise de cobertura e uso desses termos nas fontes selecionadas. O resultado obtido apresenta-se como um conjunto de conceitos unificadores que capturam a noção de proveniência.

A solução proposta se atém à construção e representação do modelo de proveniência e a avaliação desse modelo para casos concretos (capítulo 4). Não faz parte deste trabalho, apresentar uma solução computacional para a coleta automática de proveniência. No entanto, uma sugestão para essa coleta é apresentada em aplicações para desktop semântico e em um estudo preliminar de um centro de informações (seção 4.4).

1.3. Metodologia

A metodologia adotada nesta pesquisa divide-se em três aspectos: metodologia de construção, de representação e de avaliação do modelo de proveniência. Cada uma delas utiliza uma mesma ferramenta cognitiva, baseada em perguntas, que (Al-Yahya, 2006) afirma que tem o objetivo de enriquecer o desenvolvimento de pensamentos críticos. As perguntas sugerem metaclasses, que convencionamos chamar de conceitos abstratos de proveniência ou simplesmente conceitos abstratos: *What, How, Why, Who, Where, When e Which.*

A metodologia para construção do modelo de proveniência descrita aqui é adaptada da metodologia de construção de padrões conceituais (Gangemi, 2005) e (Blomqvist, 2005; Blomqvist & Sandkuhl, 2005), cujos passos em alto nível são:

1. Seleção de fontes de conhecimento: ontologias, normas, vocabulários, estruturas lingüísticas, teorias cognitivas, metodologias, melhores práticas entre outras;
2. Pesquisa transversal de invariâncias presentes nas fontes selecionadas;
3. Identificação de aproximações sintáticas e semânticas de conceitos, e fragmentos (grupo de conceitos relacionados) que cubram o conceito de proveniência, utilizando a ferramenta cognitiva;
4. Seleção e importação de fragmentos ou aproximações;
5. Anotações ao projeto do modelo, identificando a origem das importações;
6. Codificação e consolidação de conceitos;

A metodologia de representação do modelo será reusar a nomenclatura da principal fonte de importação e preservar a nomenclatura das demais fontes.

Para todas as classes e propriedades sem identificador único, um padrão similar ao da principal fonte deve ser definido. Para a representação final do modelo, os conceitos abstratos de proveniência devem ser eliminados, restando apenas os conceitos importados ou criados, temporariamente, associados aos conceitos abstratos. Após o descarte dos conceitos abstratos, é necessário identificar quais serão as classes primárias e quais movimentações de classes devem ser realizadas para satisfazer a disjunção de classes. Adicionalmente, a metodologia de representação também lança mão, ao longo da construção do modelo, de uma ontologia parcial de proveniência, utilizada meramente para facilitar a evolução do raciocínio, e não tem nenhum outro propósito além deste único.

A metodologia de avaliação será a de especialização do modelo. Para a avaliação de casos concretos, devem ser utilizados dados reais. Na ausência total ou parcial deles, devem ser produzidos dados sintéticos de forma a permitir a representação do domínio em análise. Para as avaliações deve ser apresentado o mapeamento dos principais conceitos, e preferencialmente exemplos de instâncias. Nesse processo, a ferramenta cognitiva e conceitos não unânimes podem ser novamente utilizados para a construção da analogia.

A Figura 2 contém a ontologia parcial para proveniência que utilizaremos para acumular os resultados parciais ao longo do caminho e que enfatiza apenas os conceitos abstratos de nossa ferramenta cognitiva. Ao longo do texto faremos três prévias (seções 2.1.6, 2.3.6 e 2.4.4), utilizando a ontologia parcial, listando apenas as classes mais relevantes encontradas durante a pesquisa.

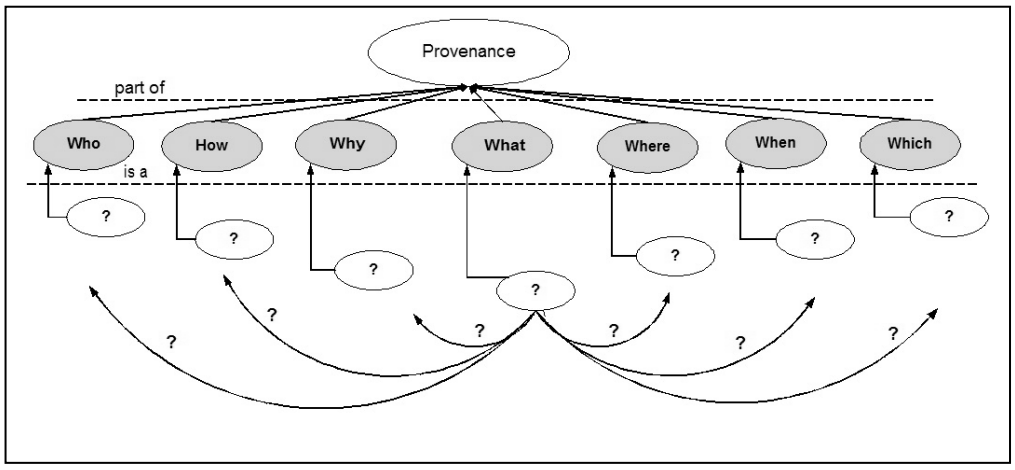

Figura 2: Ontologia parcial para proveniência

1.4. Objetivo

O foco de trabalho está em criar um modelo genérico para proveniência baseado em classes e relacionamentos já existentes em outros modelos conceituais. Os princípios adotados - proveniência, completude e reuso – exigem que:

- A solução para o modelo de proveniência registre sua origem;
- Leve em conta um amplo contexto de proveniência para que sirva como generalização para diferentes domínios;
- Seja construída a partir de conceitos e idéias de fontes que são referências mundiais.

Somando-se esses três pontos, o desafio do modelo é ser capaz de prover uma solução global que facilite a interoperabilidade entre sistemas que o adotam. Descrevemos os objetivos alinhados com tais princípios, como uma lista ordenada que será referenciada em nossas conclusões (capítulo 5):

O1. Identificar caminhos que alavanquem o reuso e aumentem as possibilidades de integração.

O2. Construir um modelo a partir do uso e da avaliação de ontologias de alto nível entre outras fontes para seleção de invariantes, fragmentos e identificação de conceitos não unânimes que cubram a noção de proveniência.

O3. Identificar um padrão internacional que possa ser utilizado como referência principal para modelagem de proveniência e que tenha influenciado direta ou indiretamente outros projetos.

O4. Explorar os modelos propostos em outros projetos e seus respectivos alinhamentos com o padrão internacional identificado, como forma de compreendê-los a partir de seus relacionamentos uns com os outros. De forma complementar, é preciso que os modelos ofereçam uma cobertura estrutural para proveniência.

O5. Concentrar esforços nos estudos de diferentes modelos conceituais para a construção e avaliação do modelo conceitual genérico para proveniência.

A partir da compreensão desses objetivos, definiremos quais reusos são possíveis e quais caminhos precisam ser explicitados, consolidando a visão como um modelo conceitual genérico para proveniência. Não é alvo deste trabalho o desenvolvimento de uma interface para a entrada de dados manual, tampouco para coleta automática da proveniência.

1.5. Organização do Texto

Este texto está organizado da seguinte forma. No Capítulo 2 descrevemos inicialmente o conceito de proveniência (seção 2.1), em seguida, discutimos aspectos de projetos de proveniência (seção 2.2) e apresentamos uma análise de cobertura da noção de proveniência utilizando ontologias de alto nível (seção 2.3). Por fim, elucidamos os resultados obtidos com a avaliação análoga de projetos e padrões (seção 2.4).

No Capítulo 3 apresentamos as preliminares à construção do modelo (seção 3.1). Então, discutimos a estratégia (seção 3.2) e a tática (seção 3.3) do projeto do modelo de proveniência. Destacamos a proveniência do modelo de proveniência (seção 3.4) e descrevemos o modelo de proveniência (seção 3.5). Por fim, apresentamos as considerações finais (seção 3.6).

No Capítulo 4 avaliamos o modelo de proveniência a partir de quatro diferentes domínios de aplicação. Inicialmente, as preliminares destacam a importância de uma API de serviços para proveniência e como consultas à proveniência poderiam ser estruturadas (seção 4.1). Em seguida, utilizamos dois domínios como motivação: o primeiro domínio analisa a aplicação do modelo sob a perspectiva de aplicações semânticas para desktops (seção 4.2) e o segundo domínio propõe a discussão do modelo como uma possível generalização a ser considerada para *Design Rationale* (seção 4.3). Em seguida, no terceiro domínio realizamos um estudo preliminar (seção 4.4) que sugere a adoção do modelo conceitual de proveniência para o desenvolvimento de um centro de informações que tem atribuições de registrar a história de um empreendimento. Apresentamos nossa avaliação com o quarto domínio (seção 4.5), que utiliza dados reais de um projeto de catálogo, a partir do mapeamento da ferramenta adotada pela equipe de desenvolvimento para o gerenciamento de configuração de software. Por fim, concluímos com as considerações finais (seção 4.6).

No Capítulo 5 apresentamos nossas conclusões, destacamos as principais contribuições deste trabalho de dissertação e apontamos possíveis direções para trabalhos futuros.