2
Contexto para Proveniência

O contexto para proveniência proporciona um olhar além das especificidades de domínios e sugere a adoção da modelagem disciplinada. A partir de análises de ontologias de alto nível, projetos e padrões é possível estabelecer uma base para identificação de padrões de modelagem. As análises transversais desses conteúdos heterogêneos permitem a extração de conceitos comuns para a construção de um modelo de proveniência de valor superior a uma terminologia, que pode evoluir para uma ontologia.

O estudo do contexto também não tem apenas o objetivo de fornecer um conjunto de classes genéricas que atenda a diferentes domínios de aplicação, que Rocha & Edelweiss (2001) definem como *framework* conceitual. Os resultados esperados deste capítulo são, de fato, a identificação e análise de invariantes que capturam os conceitos de proveniência, permitindo a representação do conhecimento acerca desse tema, e a construção de uma base conceitual para sugerir um padrão de projeto que seja uma abstração para problemas que envolvam proveniência.

Neste capítulo, 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 de proveniência utilizando ontologias de alto nível (seção 2.3). Por fim, elucidamos os resultados obtidos com a análise análoga de projetos e padrões (seção 2.4).

2.1.
O Conceito de Proveniência

Até o momento utilizamos a definição intuitiva de que proveniência refere-se à origem. Mas a modelagem do conceito de proveniência está exposta a uma diversidade semântica. Moreau (2006) apresenta mais de uma centena de formas diferentes de utilização do termo proveniência. Faz-se necessária uma clara, correta e ampla definição do conceito de proveniência.

Exploramos definições de dicionários (seção 2.1.1), seguida de uma apresentação do termo proveniência em alguns padrões de metadados (seção 2.1.2). Mapeamos os conceitos de História em conceitos de proveniência (seção 2.1.3) e analisamos a proveniência e o ciclo de vida da informação (seção 2.1.4).

Por fim, destacamos a sétima pergunta de proveniência (seção 2.1.5) e concluímos com uma ontologia parcial para proveniência (seção 2.1.6).

2.1.1. Definição de Dicionário

Inicialmente introduzimos a definição intuitiva de proveniência. Sua epistemologia tem a raiz no verbo Francês '*provenir*'. O Moderno Dicionário da Língua Portuguesa Michaelis apresenta a seguinte definição para proveniência:

Definição 2.1.1.1: (i) lugar de onde alguma coisa provém. (ii) fonte, origem, procedência.

O novo dicionário Aurélio da Língua Portuguesa define proveniência como procedência:

Definição 2.1.1.2: (i) ato ou efeito de proceder. (ii) lugar donde alguém ou algo procede. (iii) origem. (iv) proveniência.

O dicionário Infopédia[1] da Língua Portuguesa define proveniência como:

Definição 2.1.1.3: (i) lugar donde uma coisa provém. (ii) procedência. (iii) origem. (iv) fonte. (do lat. provenientìa, part. pres. neut. pl. subst. de proveníre, «provir»)

Faz-se necessário um esclarecimento mais específico porque há uma sobreposição dos conceitos de proveniência e procedência. Dessas três primeiras definições, inicialmente podemos dizer que proveniência seria: fonte, origem, exatamente como se fez intuitivamente (seção 2.1). O Arquivo Nacional (2005) apresenta a noção de procedência de forma diferente da noção de proveniência.

[1] http://www.infopedia.pt

Definição 2.1.1.4 - Procedência: termo geral empregado para designar a origem mais imediata de um conjunto de documentos, quando se trata de entrada de documentos à instituição ou serviço com finalidade de custódia, efetuada por entidade diversa daquela que o gerou. Conceito distinto do de proveniência.

Definição 2.1.1.5 - Proveniência: termo que serve para indicar a entidade que age de maneira organizada e é identificada por um nome específico, produtora de um conjunto de documentos.

O intuito de destacar as duas últimas definições (2.1.1.4 e 2.1.1.5) é única e exclusivamente o de chamar atenção para o fato de que a noção de procedência é diferente da noção de proveniência. Os dicionários de língua portuguesa consultados definem que proveniência é procedência. Como a grande maioria das fontes pesquisadas encontra-se em língua inglesa e também como forma de esclarecer a definição de proveniência – apresentada pelos dicionários de língua portuguesa – optamos por complementar a seção com duas definições adicionais em inglês, mantendo sua transcrição em formato original.

O dicionário inglês Oxford fornece a seguinte definição para o termo proveniência (*provenance*):

Definição 2.1.1.6: *(i) the fact of coming from some particular source or quarter; origin, derivation. (ii) the history or pedigree of a work of art, manuscript, rare book etc.; concr., a record of the ultimate derivation and passage of an item through its various owners.*

O dicionário americano *online* Merriam-Webster[2] define proveniência da seguinte forma:

Definição 2.1.1.7: *(i) the origin, source; (ii) the history of ownership of a valued object or work of art or literature.*

Todas as definições são compatíveis porque apresentam uma descrição sobre a origem, mas consideramos que as definições ainda apresentam certa especialização. Em busca de uma abstração mais adequada, exploramos o termo proveniência como descrição de um dado e, portanto, um metadado (seção 2.1.2). Nas definições também aparece a noção de história agregando uma semântica importante ao conceito de proveniência, por isso, exploramos um mapeamento de conceitos de história nos conceitos de proveniência (seção 2.1.3).

De agora até o final desta dissertação adotaremos o critério de apresentar todas as figuras em língua inglesa, preservando o conceito definido na fonte referenciada, para evitar a introdução de novos termos em português.

2.1.2. Proveniência como Metadado

Buneman et al. (2000) argumenta que o termo proveniência de dados (*data provenance*) é usado para se referir a fontes de consultas ou a serviços baseados no processamento de resultados.

Em determinadas áreas de conhecimento, é comum encontrarmos o termo linhagem de dados (*data lineage,* ou apenas *lineage*) (Melton et al., 1995), aplicado no lugar de proveniência de dados como seu sinônimo. Nesses casos, se refere à história de processamento de um conjunto dados.

Bose & Frew (2005) identificam outros termos que aparecem na literatura que poderiam ser interpretados como *data provenance* ou *data lineage*: *filiation*, *data genealogy*, *data set dependence*, *data archeology*, *audit trail* e *derivation history*.

Descrevemos a definição do termo que se refere à proveniência, conforme apresentado, respectivamente, nos padrões de metadados Dublin Core (2.1.2.1), Warwick Framework (2.1.2.2) e ISO 19115:2003 (2.1.2.3).

[2] http://www.m-w.com/

2.1.2.1. Dublin Core

O primeiro Workshop de Metadados (Metadata Workshop) ou DC-1 (Dublin Core 1) aconteceu em março de 1995 na cidade de Dublin, Ohio, EUA. O seu resultado foi um conjunto enxuto, por limitação de escopo, de apenas quinze elementos julgados em consenso e definidos como descritores de recursos eletrônicos que passou a ser chamado de Dublin Core.

Durante outros eventos com o propósito de aprofundar o estudo desse conjunto inicial ficou evidente que seria necessário acomodar outros descritores mais específicos porque os elementos originais não eram suficientes para descrever os recursos eletrônicos em todos seus aspectos.

O termo proveniência (*provenance*) é parte dessa acomodação. Inicialmente declarado como uma propriedade do DC (Dublin Core) em 20 de setembro de 2004[3], o termo *provenance* foi atualizado mais recentemente em 14 de janeiro de 2008.

A definição oficial mais atual do termo *provenance* é: uma declaração de qualquer mudança de propriedade ou custódia de um recurso, desde sua criação, que seja significante para resguardar autenticidade, integridade e interpretação. Esse termo foi declarado como necessário porque descreve informações sobre essas mudanças e pode ajudar a selecionar um recurso ou a interpretá-lo.

2.1.2.2. Warwick Framework

Um ano após o Workshop em Dublin (seção 2.1.2.1), o segundo Workshop de Metadados acontecia na cidade de Warwick, UK. O Warwick Framework é resultante desse novo encontro e descreve uma arquitetura de *containers* para agregar logicamente, ou até fisicamente, pacotes distintos de metadados.

A limitação de escopo do DC (Dublin Core) impõe naturalmente também uma limitação na cobertura do seu conjunto de descritores. Isso estimulou a identificação de outros metadados necessários a definir infra-estruturas de dados.

[3] http://www.ukoln.ac.uk/metadata/dcmi/collection-provenance acessado em 15/01/2008.

Entre esses metadados, está o de proveniência, destacado no Workshop em Warwick como um conjunto de dados que define a fonte da origem de algum objeto de conteúdo, por exemplo, a localização de algum artefato físico que teve seu conteúdo digitalizado. Opcionalmente, inclui um resumo de todas as transformações que foram aplicadas ao objeto referido. (Lagoze, 2006)

2.1.2.3. ISO 19115:2003

A ISO 19115:2003 propõe um padrão de metadados para dados e serviços geográficos digitais. A norma define elementos de metadados, entre eles, o conceito de linhagem (*lineage*).

A classe *lineage* é definida em ISO 19115 (2003) como informações sobre eventos ou dados de origem usados na construção do dado.

2.1.2.4. Linhagem

Widom (2005) destaca que cada banco de dados tem uma relação (lógica) de linhagem (*Lineage Relation*). Acrescenta que linhagem pode ser representada como uma tupla (*tupleID, derivation-type, time, how-derived, lineage-data*). Nessa tupla, *tupleID* é a chave primária, *derivationt-type* corresponde ao tipo de derivação, *time* é tempo, *how-derived* representa como o dado foi derivado e *lineage-data* é o dado propriamente dito.

Bose & Frew (2005) argumentam que o termo linhagem pode ser aplicado a itens que sejam a evolução de um produto de dados. Acrescentam ainda que há duas formas de navegação que implicam em: mover-se para a trás (*backward*) para descobrir produtos ou transformações ancestrais, ou mover-se para frente (*forward*) para descobrir descendentes. No início desta dissertação (seção 1.1) utilizamos como motivação a idéia de rastro para origem e rastro para o destino. De forma análoga temos respectivamente *backward e forward lineage* (linhagem para trás e para frente) (Bose & Frew, 2005).

2.1.2.5. Anotação

Proveniência é informação sobre a origem de um item de dado (digital). Braun et al. (2006) argumenta que nos casos onde a proveniência é adicionada após a criação do dado, ela pode ser considerada como uma anotação. Diz-se

então que o dado foi enriquecido com uma anotação. A anotação é, portanto, uma das formas de representação de proveniência. Esse e outros aspectos de proveniência são destacados mais adiante (seção 2.2).

2.1.3. Mapeamento de Conceitos de História

Nos padrões de metadados estudados (seção 2.1.2) também se encontra o conceito de evento associado ao conceito de proveniência. Somando-se as definições de dicionários (seção 2.1.1), a essas colocações, exploramos a partir de agora, o conceito de proveniência associado ao conceito de história.

De acordo com Bunge (1977), uma investigação sobre um episódio ou período histórico baseia-se em evidências, ou documentos, de eventos (Groth et al., 2006b) que aconteceram no passado. Para elucidar o conceito de proveniência, investigamos as noções de história.

Quando traduzidas para o contexto de dado (digital), essas noções podem ser mapeadas em perguntas simples que induzem conceitos abstratos de proveniência, semelhantes aos utilizados no modelo de proveniência proposto por (Ram, 2006a). A Tabela 1 ilustra esse mapeamento. A assim chamada *matriz de perguntas* (Wiederhold, 1993) tem o objetivo de ajudar a construção de questionários elucidando a proveniência dos dados. Yang et. al. (2003) destaca que essas são perguntas essenciais na construção de sistemas de respostas baseados em eventos.

Tabela 1: Mapeamento entre conceitos de História e Proveniência (Ram, 2006a)

Noção	**Pergunta**
Evento (*Event*)	O que (*What*)
Espaço (*Space*)	Onde (*Where*)
Tempo (*Time*)	Quando (*When*)
Ação (*Action*)	Como (*How*)
Agente (*Agent*)	Quem (*Who*), Qual (*Which*)

As perguntas de proveniência sobre o dado (usado no singular no seguinte sentido):

- Quem (*Who*) são os criadores, publicadores ou colaboradores? Quem tem permissões de acesso, quais são elas? Seria essa entidade confiável?

- Quando (*When*) o dado foi criado, acessado, modificado? Essa pergunta está ligada a quem acessou ou modificou o dado.
- Onde (*Where*) o dado foi reportado, ou onde está armazenado? Quais são as localizações físicas (no caso de mais de uma cópia)?
- Como (*How*) o dado foi derivado ou transformado? Essa pergunta relaciona-se à qual computação foi aplicada para transformá-lo.
- Quais (*Which*) aplicações, configurações de software ou de ferramentas foram usadas para a criação do dado. Essa questão está ligada às condições do ambiente.

Qual é o dado? Qual é o evento? A Figura 3 identifica o fato gerador (evento) que desencadeia as evidências sobre o episódio ou período (*creation* até *destruction*) a serem registrados. A elipse ilustra o conjunto de todas as situações possíveis (*state of affairs*[4]) no mundo.

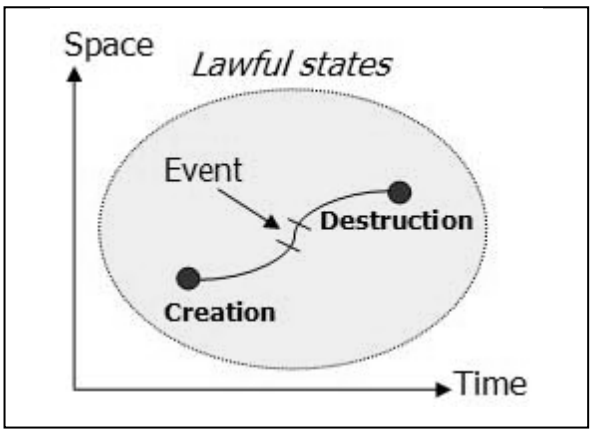

Figura 3: Visão de História de Mario Bunge (Ram, 2006a)

Para ilustrar como essas perguntas nos ajudam na captura do que aconteceu, considere o desenvolvimento de uma biblioteca de *software* para prover análise de navios com a proposta de simular o comportamento de embarcações ao serem submetidas a ondas, ventos e correntes marítimas.

Do ponto de vista de projeto a avaliação de características e respectiva atribuição de valores e pesos são fundamentais para a seleção do pacote a ser

[4] http://www.thefreedictionary.com/state+of+affairs define ***state of affairs*** como: "the general state of things; the combination of circumstances at a given time".

utilizado por um protótipo. O registro dos dados de proveniência de cada pacote avaliado é parte dos dados que ajudam a qualificar os fabricantes.

A equipe de desenvolvimento preparou um protótipo a partir do pacote selecionado de simulação de oceanos de diferentes fabricantes.

O fabricante americano foi descartado porque o pacote não havia sido adotado por nenhum concorrente de mercado. Outro fabricante também foi eliminado porque seu pacote tinha menos de um ano de vida. Nesses casos, as perguntas Quem (*What*), Onde (*Where*) e Quando (*When*) foram fundamentais na qualificação de potenciais fornecedores.

Agora, suponha que a equipe de modelagem tenha descoberto um novo padrão utilizado para construção de navios. Como o padrão foi recentemente lançado, não havia exemplos de como ele poderia ser aplicado. Em um determinado momento a equipe se deu conta que outro grupo de desenvolvedores havia utilizado o padrão em um projeto similar (com requisitos semelhantes). A equipe consultou a proveniência desses projetos para descobrir quais eventos teriam o registro de como e porque o padrão foi utilizado, para avaliar a experiência disponível anterior. Aqui é claro que perguntas de Quem, Como e Por que (*Who, How* e *Why*, respectivamente) facilitam o reuso e o compartilhamento de lições aprendidas.

A partir de então, considere que dois engenheiros de testes sênior fossem responsáveis por homologar o uso da API. Ambos realizaram o mesmo teste para diferentes casos de uso. O primeiro engenheiro explorou a API para simular ondas de nível médio por um período de tempo extenso, enquanto que o segundo computou os testes registrando os resultados aplicando ondas gigantes durante um curto espaço de tempo. Esse cenário coloca em evidência a necessidade de registrar cada caminho. Usando a pergunta Como (*How*) para capturar os diferentes procedimentos, seria então possível compará-los.

Finalmente, imagine que um analista de banco de dados tenha detectado um sério problema de desempenho relacionado a uma classe do módulo de persistência. Suponha que o erro existisse desde outubro de 2007. Ele ou algum outro recurso do projeto poderia então desejar localizar todos os artefatos que utilizaram tal classe desde então. Se fosse possível responder a perguntas Quais e Quando, haveria a possibilidade de identificar os artefatos que necessitassem ser avaliados e eventualmente corrigidos.

A ilustração descrita até aqui reforça que o registro histórico dos eventos de um projeto é essencial para capturar a proveniência dos dados. Neste exemplo, identificamos várias formas de uso para proveniência: qualidade dos

dados, auditoria, replicação, atribuição e descoberta de dados de caráter geral. Os dados de proveniência podem ser utilizados não apenas para uma boa gestão de um projeto, mas também para apoiar a conformidade com processos e normas. Em seguida, exploraremos proveniência do ponto de vista de ciclo de vida (seção 2.1.4).

2.1.4. Proveniência e Ciclo de Vida

Considere que proveniência é informação. Mills (2006) destaca que o ciclo de vida da informação usualmente se inicia com a captura e registro automáticos de eventos a partir de interações entre usuários e programas, seguido de um monitoramento e análise da informação por diferentes fontes e em diferentes formatos.

A Figura 4 adaptada de (Ram, 2006a) ilustra tipos (taxonomias) de eventos que podem ocorrer em um ciclo de informações no contexto de projeto e desenvolvimento de produtos. Optamos por manter os tipos em inglês (figura original) propositalmente para evitar a introdução de novos conceitos a partir das respectivas traduções para o português.

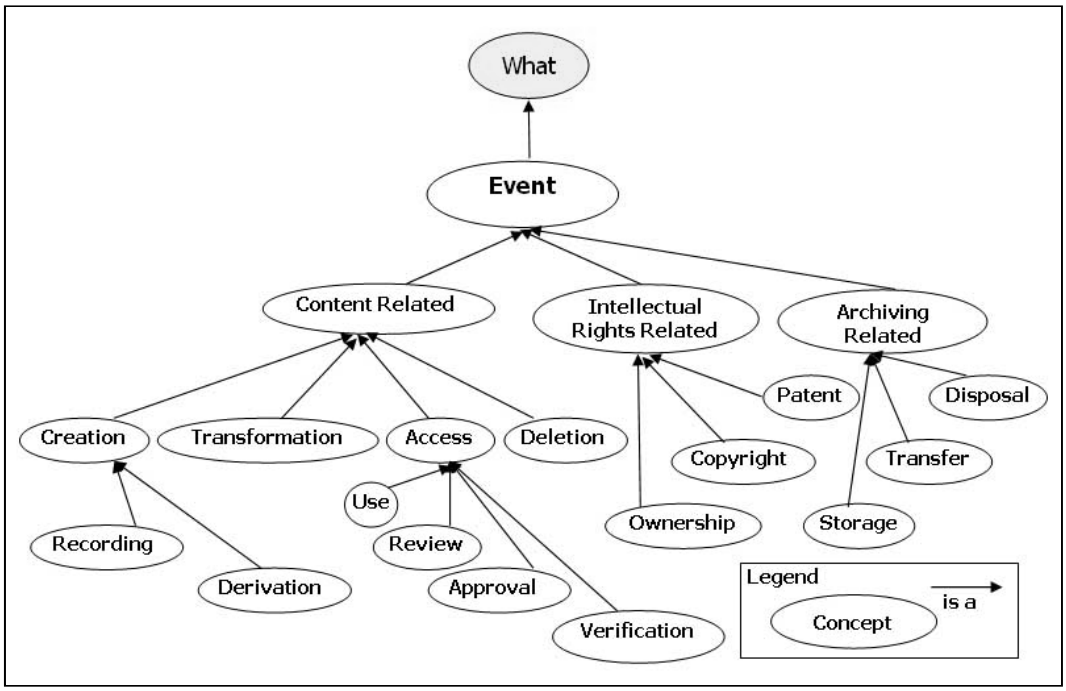

Figura 4: Ciclo de vida da informação (Ram, 2006a)

Boeuf (2006) define um evento como “algo (*What*) que acontece no espaço e no tempo e provoca alguma mudança no mundo”. Um evento pode envolver: pessoas, grupos, entre outros, que potencialmente desempenham um papel ativo ou passivo, por exemplo, provocando o evento, testemunhando-o ou submetendo-se a ele. Esse papel pode se referir a uma participação mais abstrata se estiver associado a criações do intelecto. Adicionalmente, um evento ocorre no espaço e no tempo e, portanto, possui duração (Boeuf, 2006). Ancoramos o ciclo de vida de proveniência na noção de evento.

Groth et al. (2006a) argumenta que o ciclo de vida de proveniência pode ser modelado como um processo com quatro fases: criação (*creation*), registro (*record*), consulta (*query*) e gerenciamento (*manage*). A primeira fase é a responsável pela descoberta de significado e criação da proveniência.

A segunda fase foca no armazenamento dos metadados de proveniência para uso futuro. A estratégia de armazenamento de longo prazo para guardar os metadados de proveniência demanda um componente chamado armazenador de proveniência (*provenance store*) (Groth et al., 2006a). Esse componente provê persistência e gerenciamento de acesso aos metadados de proveniência.

A terceira fase compreende a consulta ao armazenador de proveniência por usuários ou aplicações que desejam obter os metadados de proveniência. Em sua forma básica, o resultado da consulta seria o conjunto de metadados de proveniência. Uma consulta avançada poderia retornar dados mais elaborados derivados com a ajuda dos dados de proveniência.

Finalmente, a quarta fase trata do gerenciamento do armazenador de proveniência. A principal característica dessa fase é manter os metadados de proveniência sincronizado com os respectivos dados. Esse gerenciamento cobriria o arquivamento, exclusão e descarte da proveniência. Um modelo de proveniência deve levar em conta todas essas questões (Groth et al., 2006a).

2.1.5. A Sétima Pergunta de Proveniência

Ao longo do mapeamento de conceitos de história (seção 2.1.3), uma pergunta que ainda não havia aparecido no mapeamento das noções de história que faz toda a diferença saber respondê-la é: por que (*Why*) o evento ocorreu? Por que um dado sofreu uma suposta transformação? Essa pergunta está intuitivamente ligada à capacidade de rastrear a sequência de idéias relacionadas.

É essencial que o conceito de proveniência capture o conceito de razão (*Reason*), que inclui motivações como crença, desejo e intenção. Todos estes três últimos são fatores significativos e podem ser representados através do modelo *Belief-Desire-Intention* (Georgeff, 1999), onde crença representa o conhecimento sobre o mundo, desejo é o objetivo associado a um agente e intenção é o comprometimento de um agente com um determinado objetivo. De fato, é crucial rastrear as razões para capturar respostas para "o porquê" (*Why*) e registrar a proveniência que explica o motivo de uma dada criação ou transformação.

Para eleger quais classes fariam parte de nosso modelo conceitual de proveniência, mais adiante (seção 2.3), aprofundaremos o estudo das perguntas de proveniência consultando ontologias de alto nível. Por hora apresentaremos a primeira prévia da ontologia parcial para proveniência (seção 2.1) que consideramos como conclusão dos assuntos explorados até aqui.

2.1.6. Ontologia Parcial para Proveniência (prévia 1)

A Figura 5 apresenta um modelo conceitual parcial que utiliza a ferramenta cognitiva adotada.

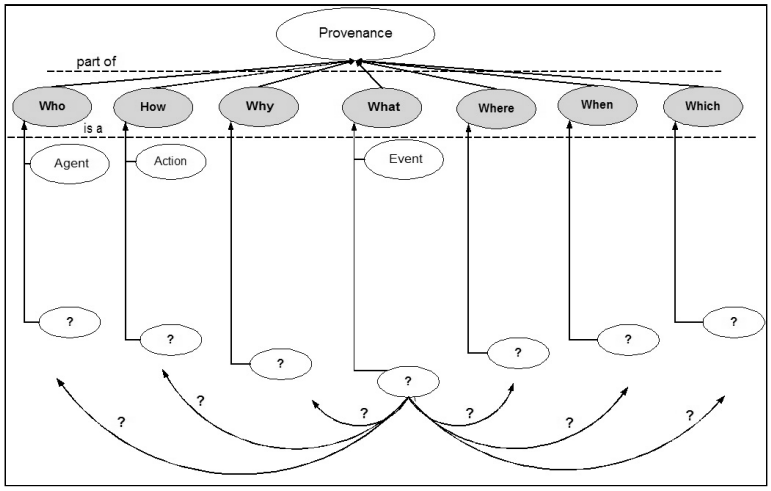

Figura 5: Ontologia parcial para proveniência (prévia 1)

O modelo W7 (Ram et al., 2006b) também utiliza as classes abstratas que estão presentes em nossa ferramenta cognitiva. Contrapomos que as classes abstratas no modelo que estamos construindo são apenas metaclasses, enquanto que no modelo W7 (Ram et al., 2006b) as classes abstratas são

classes primárias. Representamos inicialmente na Figura 5, a primeira prévia da ontologia parcial para proveniência que é parte de nossa metodologia de representação.

As classes são ovais e as propriedades de objetos são arcos unidirecionais. As classes estão estratificadas em três níveis:

- O nível de topo, ou 1º nível, contém apenas a classe Proveniência (*Provenance*);
- O nível intermediário, ou 2º nível, contém as classes que capturam as *Wh-Questions* (Cheng, 1997) que representam conceitos abstratos de proveniência e são parte de nossa ferramenta cognitiva inspirada em (Wiederhold, 1993);
- O 3º nível contém as classes que capturam os detalhes da proveniência.

Fundamentados nas seções anteriores, inicialmente assumimos que o conceito *Event* (uma especialização da pergunta “o que” ou em inglês *What*) é uma classe central do 3º nível, juntamente com *Agent* e *Action* e demais classes ainda não discutidas, representadas por círculos e identificadas com a marca de interrogação “?” na Figura 5. As classes do 3º nível deveriam idealmente ser importadas de padrões conhecidos.

Os conceitos abstratos de proveniência fazem parte da ferramenta cognitiva definida para apoiar a construção do modelo de proveniência. Analogamente, apresentaremos outras prévias da ontologia parcial para proveniência (seções 2.3 e 2.4). A seguir, descrevemos aspectos de projetos de proveniência (seção 2.2). No capítulo 3, revisaremos os conceitos associados aos conceitos abstratos de nossa ferramenta cognitiva, para organizar as classes importadas para 3º nível da ontologia.

2.2. Aspectos de Projetos baseados em Proveniência

A taxonomia de técnicas para a proveniência, Figura 6, sugere uma classificação para os esforços de pesquisa de proveniência para a área de e-Science. Destacamos que essa taxonomia oferece uma visão dos aspectos de projeto de sistemas de proveniência: porquê registram a proveniência, o que descrevem, como representam e armazenam, quais formas de disseminação de proveniência são exploradas, quais formas de uso, entre outros. Simmhan et al.

(2005) afirma que a síntese resultante da pesquisa pode ajudar projetistas de sistemas de gerenciamento de metadados a compreender os projetos estudados. Destacamos aqui os aspectos que consideramos relevantes para o trabalho desenvolvido nesta dissertação. Optamos por manter os tipos em inglês na Figura 6 (figura original) propositalmente para evitar a introdução de novos conceitos a partir das respectivas traduções para o português. Ao longo do texto desta seção nos referenciamos a essa figura utilizando os termos em inglês da taxonomia apresentada. Destacamos também que as setas dessa figura indicam a direção da especialização. Os dois pequenos círculos são recursos visuais para simbolizar o nível da hierarquia das classes que estão ligadas abaixo deles e, nada mais além disso.

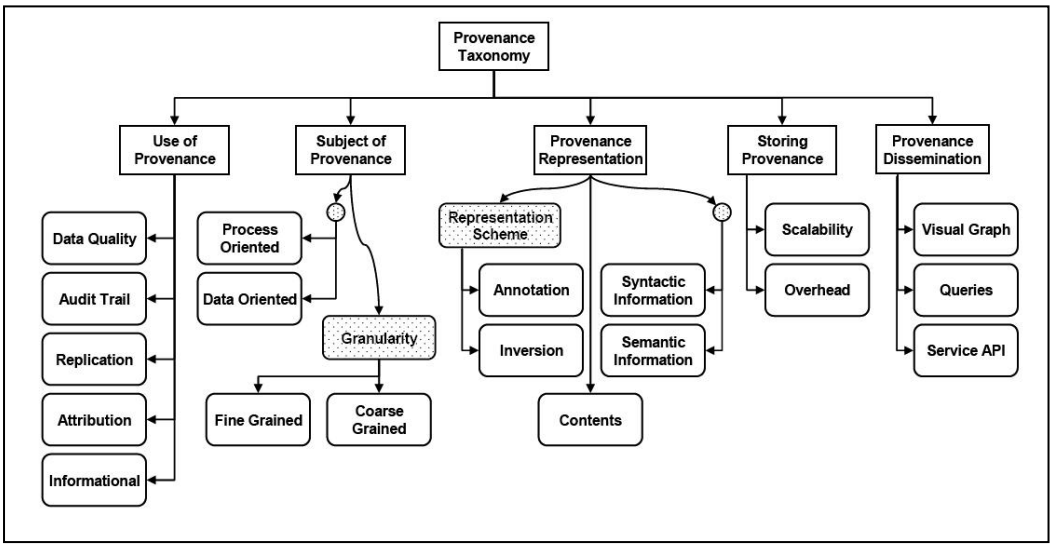

Figura 6: Taxonomia de Aspectos de Proveniência de Dados (Simmhan et al., 2005)

A pesquisa sobre proveniência (Simmhan et al., 2005) contempla projetos da área de e-Science e cobre cinco diferentes domínios (ciências): geografia, astronomia, biologia, química e meteorologia. Com exceção do domínio de biologia, que teve dois projetos estudados, todas as demais ciências tiveram apenas um projeto representante. Adicionalmente, a pesquisa classifica ainda, dois projetos como genéricos da área de e-Science.

Iniciamos resumindo a distribuição dos tipos de arquitetura de banco de dados adotados pelos projetos (Simmhan et al., 2005), onde: quatro projetos

apresentam arquitetura orientada a serviço[5], três são relacionais, um é baseado em script e outro é baseado em processamento de comandos.

Simmhan et al. (2005) destaca diferentes usos de dados de proveniência, como: qualidade de dados (*data quality*), auditoria (*audit trail*), replicação de experimentos (*replication*), atribuição de autoridade (*attribution*) e descoberta de dados (*data discovery*). A qualidade de dados pode ser estimada baseada em sua origem e nas transformações que sofre. Auditoria permite a identificação do uso de recursos e a detecção de erros. Replicação permite a repetição das derivações de dados. Atribuição estabelece propriedade, custódia, direitos e deveres. Por fim, descoberta de dados apóia a análise de contexto, que ajuda a interpretação do dado.

Nos nove projetos avaliados (Simmhan et al., 2005), há duas estratégias: orientação a dado (*data oriented*) ou orientação a processo (*process oriented*). Contabilizamos cinco projetos para a primeira, três para a segunda e um que mescla ambas. Outro ponto destacado por (Simmhan et al., 2005) refere-se à granularidade de coleta da proveniência, onde apontamos que está relacionada essencialmente ao foco selecionado no "intervalo" definido por eventos complexos ou atômicos.

Outro aspecto tratado é a representação da proveniência como anotação (*annotation*) ou como consulta invertida (*inverse query*). Na primeira, a história da derivação é representada como uma anotação associada ao dado. Por outro lado, na segunda as derivações são resultados de consultas invertidas, onde conhecendo-se o resultado, procura-se o dado de entrada que o produziu. Simmhan et al. (2005) afirma que a representação da proveniência como anotação permite um conteúdo mais rico e previamente processado, e contrapõe que a representação por consulta invertida é mais compacta, mas é processada dinamicamente. Por fim, acrescenta que a representação sintática pode adotar diferentes linguagens, por exemplo, XML ou OWL. Como praticamente metade dos projetos avaliados tem arquitetura de banco de dados orientada a serviço, nesses casos, o formato XML é uma escolha natural para a troca de dados. Mas a codificação do conhecimento semântico possibilita um uso mais elaborado da

[5] Adicionalmente às operações típicas de um banco de dados relacional, em um banco de dados orientado a serviços, pode-se invocar funções definidas pelos usuários que implementam chamadas a procedimentos armazenados (*stored procedures*)

proveniência, sendo uma base para inferências e processos de prova. (Simmhan et al., 2005)

Simmhan et al. (2005) chama a atenção para o fato da proveniência poder crescer a ponto de ficar maior do que o dado que ela descreve. Por isso, apresenta questões sobre escalabilidade e *overhead*, que devem ser analisadas quando o volume de dados é grande, especialmente quando a coleta da proveniência respectiva é granular.

A disseminação da proveniência é classificada por (Simmhan et al., 2005) de três formas: consultas (*queries*), grafo visual (*visual graph*) e API de serviços. A distribuição dos projetos nessa classificação é respectivamente quatro, quatro e um. Simmhan et al. (2005) conclui que o projeto PASOA (Moreau & Ibbotson, 2006) está na direção certa ao buscar a definição de uma API de serviços para proveniência, mas ressalta que precisa de maiores refinamentos para elucidar como a proveniência é representada e recuperada. Acrescenta ainda que qualquer padrão proposto para proveniência deve satisfazer às necessidades de múltiplos domínios, cujos requisitos necessitam ser identificados. Conclui apontando que um deles é seguramente a padronização da semântica dos termos.

Simmhan et al. (2005) apresenta algumas questões ainda em aberto relativas à proveniência e os desafios que precisam ser superados para que o tema proveniência se torne permeável a diferentes comunidades. O compartilhamento de dados entre organizações é crescente, por isso Simmhan et al. (2005) destaca que é essencial que a proveniência também o seja. Acrescenta que a maioria dos projetos pesquisados possui protocolos proprietários para gerenciar a proveniência e que não fundamentam a sua coleta, representação, armazenamento e consulta em padrões abertos, o que dificulta a interoperabilidade.

Com o objetivo de alcançar essa padronização, buscamos identificar fragmentos de ontologias de alto nível (seção 2.3) e apoiar a construção deste trabalho de dissertação em normas internacionais e projetos de boa reputação (seção 2.4) que permitam a representação da proveniência. As contabilizações realizadas nesta seção serão uma base para o direcionamento da pesquisa.

2.3.
Ontologias de Alto Nível

2.3.1.
Preliminares

Entendemos as ontologias de alto nível (do inglês *Upper Level Ontologies*) como fontes interessantes de padrões de modelagem porque a construção da hierarquia de classes e as propriedades de cada classe presentes nessas ontologias estão muito bem definidas.

Esse rigor nos conceitos seria uma das principais vantagens do uso de ontologias de alto nível como fontes de consulta e referência para construção das classes de novos modelos conceituais. Essas consultas e referências oferecem a possibilidade de uma integração *a priori* (Bellatreche et al., 2006). De fato, o esforço de realizar uma boa modelagem minimiza a necessidade de uma integração *a posteriori* que, em geral, é sempre mais custosa e fortemente dependente do alinhamento de conceitos entre esquemas a partir do processamento e análise de suas respectivas instâncias (Brauner et al., 2006).

Se as fontes de dados fossem em sua grande maioria modeladas a partir de padrões não ambíguos, estaríamos muito provavelmente minimizando o custo de integração quando uma determinada aplicação necessitasse realizar uma consulta a essas fontes, uma vez que esses conceitos teriam uma definição precisa.

Mas, ainda assim, há inúmeras ontologias de alto nível e não é foco deste trabalho a análise de todas as existentes. Portanto, o primeiro passo foi selecionar três fontes que satisfizessem os critérios:

- Licenciamento de código aberto;
- Conter em suas definições os conceitos: evento, agente e ação;
- Existência de estudos que apresentem a avaliação em conjunto ou individual das ontologias selecionadas;
- Disponibilidade de consulta a arquivos em formato OWL ou qualquer outro compatível com visualizadores de código aberto.

As ontologias que atenderam de forma mais adequada a esses critérios foram: a DOLCE (do inglês *Descriptive Ontology for Linguistic and Cognitive Engineering*) (Masolo et al., 2003), produzida a partir da metodologia (Guarino & Welty, 2002); a SUMO (do inglês *Suggested Upper Merged Ontology*) (Niles & Pease, 2001) modelada pelo IEEE Standard Upper Ontology Working Group; e a

OpenCyc[6], que é considerada a maior e mais antiga ontologia de alto nível disponível. Essas três ontologias de alto nível foram objeto de análise preliminar em (Semy et al., 2004), que qualificou a ontologia DOLCE como padrão candidato a atender ao domínio de governo americano.

A opção deste texto foi traduzir o mínimo necessário para garantir a sua qualidade e compreensão. Com isso, ao longo deste texto, ocorrerão termos em inglês sempre quando optamos por não traduzí-los, evitando assim a introdução de outros termos em português. Limitamo-nos apenas a descrever o termo em inglês com um texto em português para garantir a fluidez da leitura. Há ainda termos que foram deixados em seu formato original (inglês) propositalmente.

A consulta às ontologias de alto nível foi feita utilizando os conceitos evento, agente e ação, a partir do mapeamento da noção de história (seção 2.1.3). Usamos ainda aproximações dos três primeiros com o propósito de realizar a identificação de invariantes (seções 2.3.2, 2.3.3 e 2.3.4). Para exemplificar como a estrutura conceitual varia de uma ontologia de alto nível para outra, ilustraremos uma parte da hierarquia de especialização que inclui apenas a classe que representa a noção de evento. Por fim, apresentamos o alinhamento dos conceitos de proveniência entre as três diferentes ontologias selecionadas (seção 2.3.5).

2.3.2. DOLCE

A versão estudada é a 3.9 da DOLCE-Lite, acrescida de algumas extensões básicas, chamada de DOLCE-Lite-Plus ou DLP, de 28 de março de 2006. A DOLCE-Lite incorpora o resultado do relatório final D18 do projeto *WonderWeb*, que corresponde à documentação final e completa da DOLCE (*Descriptive Ontology for Linguistic and Cognitive Engineering*).

A DOLCE baseia-se nos princípios da metodologia OntoClean (Guarino & Welty, 2002). A principal divisão da taxonomia é entre *Perdurant* (Perdurável) e *Endurant* (Contínuo). Os primeiros são entidades que se estendem no tempo enquanto que os segundos são entidades presentes continua e independentemente do tempo.

---

[6] http://www.opencyc.org

A Figura 7 ilustra essa divisão e destaca quais instâncias da classe *Endurant* podem ter a propriedade espaço-temporal (*PhysicalEndurant*) ou não (*NonphysicalEndurant*). Por outro lado, especializações da classe *Perdurant* são instâncias da classe *Event* ou da classe *Stative*, dependendo de suas características temporais (Masolo et al., 2003).

Na DOLCE, a classe Evento (*Event*) está no mesmo nível da classe DeEstado (*Stative*). A definição da classe Evento afirma que, em geral, eventos diferem de situações porque não tem uma descrição da qual dependam, ou seja, não tem um estado. Eventos podem ter uma relação sequencial dada por algum fluxo, mas não requerem uma descrição como critério de unificação.

Por outro lado, é possível que se crie uma descrição que determina valores para as restrições que caracterizam um tipo de evento e, nesse caso específico, o evento seria na verdade uma situação.

Outra noção de evento também investigada pela DOLCE está relacionada à idéia de mudança. Com isso, eventos poderiam ser considerados como aspectos ou partes de processos (Masolo et al., 2003).

Por exemplo, o processo “recuperação da mata ciliar do rio Macacú” pode ser conceituado como: uma realização (o processo de recuperação como o resultado de ações prévias), um estado (se colapsarmos o intervalo de recuperação em um único ponto, contabilizando o número de árvores de mata nativa replantadas e totalizadas nesse ponto) ou como um evento (quais mudanças ocorreram que provocaram alterações em um trecho do rio levando-o de um estado para outro).

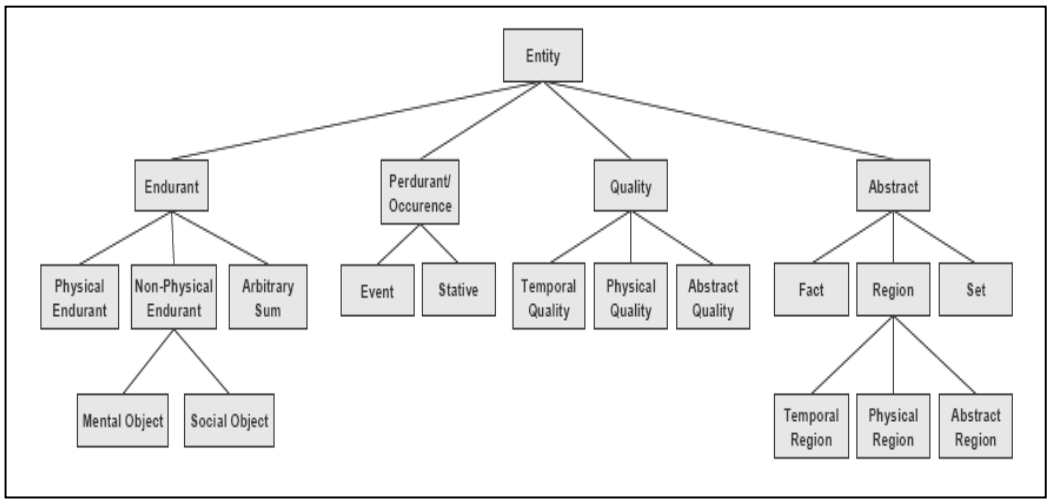

Figura 7: Hierarquia de especialização parcial da ontologia DOLCE (Semy et al., 2004)

Esses diferentes aspectos podem ser mais ou menos adequados dependendo dos objetivos de representação do conceito evento: foco na causa, foco no efeito, sumarização, identificação de transições entre outros. (Masolo et al., 2003)

Nem sempre é tão fácil compreender as descrições utilizadas em uma ontologia de alto nível para todos os conceitos apresentados. A compreensão de definições consome um tempo considerável da busca e análise de padrões.

A classe Agente (*Agent*) é uma dessas classes que muitas vezes confundem ao contrário de esclarecer. Um agente identifica que algo atuou com certa ação, ou há um executor inicial ou ainda um papel foi desempenhado em uma ação. Agentes podem representar internamente descrições, planos, objetivos ou possíveis ações, mas que não necessariamente atuam. (Masolo et al., 2003)

Instâncias da classe Ação (*Action*) exemplificam a intenção de um agente. Ações podem ser interrompidas, incompletas, abortadas e ainda assim permanecerem como uma (potencial) realização. Ter um resultado depende de um método, então uma ação ainda é uma ação mesmo se o resultado produzido é incompleto. (Masolo et al., 2003)

2.3.3. SUMO

A SUMO (*Suggested Upper Merged Ontology*) foi criada pela empresa Teknowledge com contribuições diretas da lista de distribuição SUO (*Suggested Upper Ontology*). Sua primeira versão data de 2001 e foi proposta pelo *SUO Working Group* para criar uma única, abrangente e concisa estrutura a partir da consolidação de outras ontologias e conteúdos de domínio público, entre eles: servidor Ontolingua[7], a ontologia de topo proposta por John F. Sowa[8], ontologias desenvolvidas pelo ITBM-CNR[9] (*National Research Council, Institute. of Biomedical Technologies*) e teorias mereológicas[10] (Niles & Pease, 2001). A versão 1.73 é a mais atual, mas optamos por estudar nesta seção, a versão que

[7] http://ontolingua.stanford.edu

[8] http://www.jfsowa.com/

[9] http://www.itb.cnr.it

[10] Mereologia é o estudo lógico-matemático das relações entre as partes e o todo. Fonte: Tradução da definição do Webster Online.

é considerada oficial: *SUMO starter document* que corresponde à versão 1.52 de 25 de Abril de 2003.

Não há a classe Evento (*Event*) na ontologia SUMO. Destacamos então a classe *Process* que é a classe de fenômenos que acontecem e tem partes ou estágios temporais. A Figura 8 apresenta a hierarquia parcial de especialização da ontologia SUMO, onde podemos identificar que há uma divisão dual entre Concreto (*Physical*) e Abstrato (*Abstract*). Repare que a classe *Process* está no ramo de entidades concretas. Originalmente na ontologia de topo de John F. Sowa, utilizada para a construção da SUMO, a classe Evento é uma especialização da classe Processo Discreto (*DiscreteProcess*) que por sua vez é subclasse de *Process*, esta sim, presente da Figura 8. Alguns exemplos de instâncias de *Process* são: eventos estendidos como uma partida de futebol ou corrida de carros, ações como ler ou escrever ou processos biológicos. A definição formal é algo que dura um período de tempo, mas não é um objeto. Vale notar que um processo pode ter participantes que são objetos, como os jogadores da partida de futebol ou os pilotos da corrida. (Niles & Pease, 2001)

Instâncias da classe Agente (*Agent*) são algo ou alguém que podem atuar por conta própria e produzir mudanças no mundo. Agentes têm direitos, mas podem ou não ter responsabilidades e racionalidade. Se o agente tem habilidade de racionalizar então a instância é também atribuída à classe AgenteCognitivo (*CognitiveAgent*). Animais são exemplos de instâncias da classe *SentientAgent* que é disjunta da classe *CognitiveAgent*. (Niles & Pease, 2001)

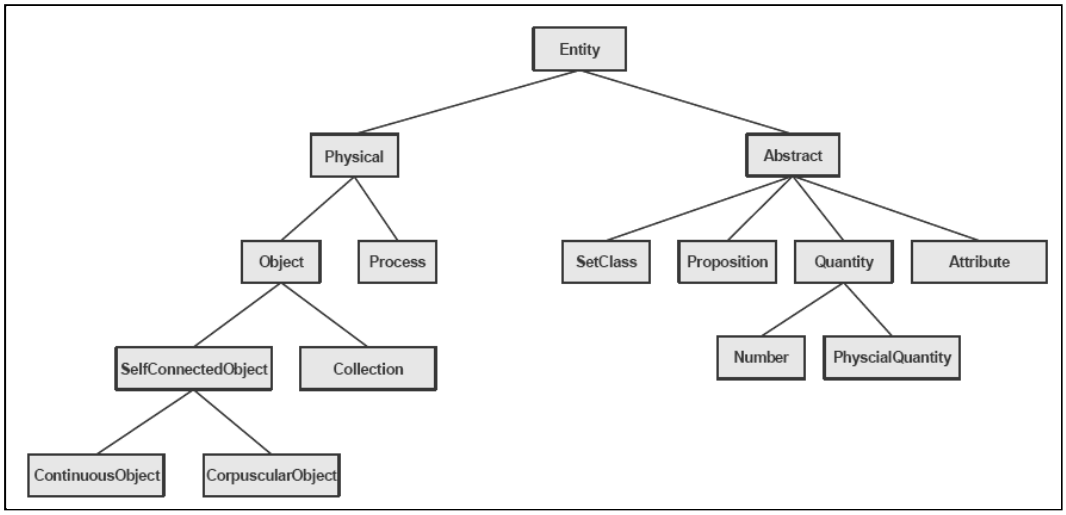

Figura 8: Hierarquia de especialização parcial da ontologia SUMO (Semy et al., 2004)

Não existe a classe Ação (*Action*) nessa ontologia. A classe mais próxima em nossa avaliação é *IntentionalProcess*, que pode ser definido como um processo que tem um propósito específico para um agente cognitivo (instância da classe *CognitiveAgent*) que o executa. (Niles & Pease, 2001)

2.3.4. OPENCYC

A ontologia de alto nível OpenCyc é a versão de licenciamento aberto da ontologia de alto nível Cyc comercial. A Cyc foi inicialmente desenvolvida na MCC (*Microelectronics and Computer Technology Corporation*) no início de 1984, construída como uma base de conhecimento do senso comum para o processamento de linguagem natural.

Essa ontologia possui centenas de milhares de termos atômicos, milhares de conceitos e uma quantidade de axiomas pelo menos uma ordem grandeza maior que a quantidade de conceitos. A versão estudada é a 1.0.2 de 14 de julho de 2006. O manuseio da ontologia, pelo seu tamanho e quantidade de classes, não é muito prático. Portanto, todas as definições desta seção são traduções dos respectivos conceitos, provenientes de consultas estruturadas realizadas à fonte www.opencyc.org.

Nesse imenso universo de termos e conceitos, estudamos apenas as definições para os três principais conceitos que elegemos como fragmento que captura parcialmente a noção de proveniência: Evento, Agente e Ação.

A Figura 9 apresenta a hierarquia de especialização parcial da ontologia OpenCyc. A classe Evento (*Event*) é uma especialização da classe *Situation* que é por sua vez, uma especialização da classe *IntangibleIndividual*. Cada instância da classe Evento é algo que alguém diz que acontece.

Eventos são intangíveis porque são mudanças por si só. Portanto, não são objetos tangíveis que sofrem mudanças. Alguns exemplos de especializações da classe Evento incluem: *LocalizedEvent, PhysicalEvent, Action* e *Transfer*. Eventos não devem ser confundidos com intervalos de tempo (classe *TimeIntervals*). A delimitação temporal de eventos, esta sim, é dada por um intervalo de tempo.

Não há uma classe com o nome Agente, mas sim Agente-Genérico (*Agent-Generic*) que é uma especialização de algo que existe de forma estável no tempo. Cada instância da classe Agente-Genérico é um ser que tem desejos ou intenções e habilidade de agir a partir deles. Essas instâncias podem ser

indivíduos ou podem consistir de vários agentes genéricos operando em conjunto como um grupo.

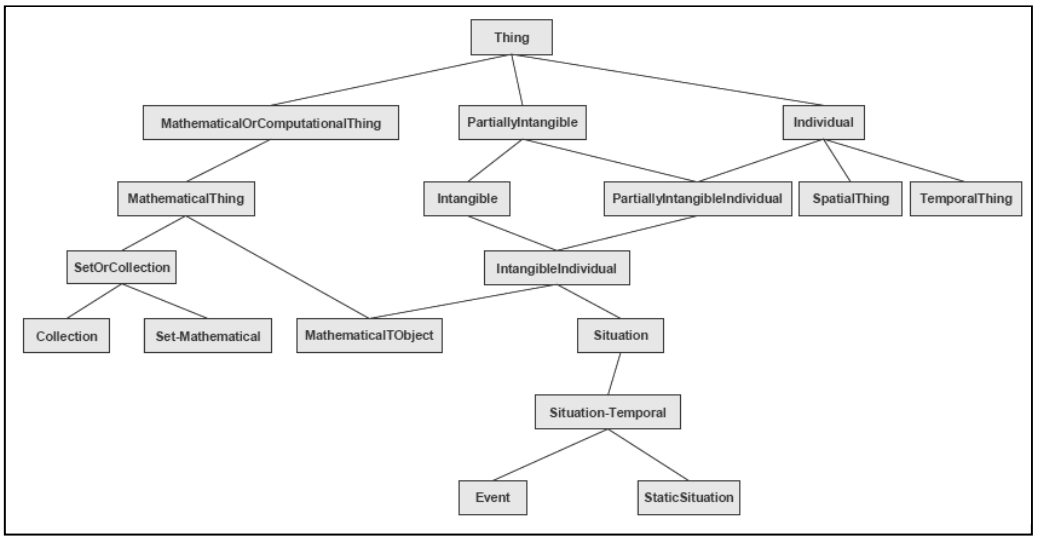

Figura 9: Hierarquia de especialização parcial da ontologia OpenCyc (Semy et al., 2004)

A classe Ação (*Action*) é uma coleção de eventos que são executados por uma instância da classe Executor (*Doer*) que descreve atores que executam eventos. Instâncias da classe *Action* incluem qualquer evento no qual um ou mais atores produzem uma mudança no estado do mundo, tipicamente ao despender um esforço ou energia. Não é necessário que nenhum objeto tangível seja movido, modificado, produzido ou destruído para que uma ação ocorra.

Os efeitos de uma ação podem ser intangíveis (por exemplo, a intimidação de um subordinado). O executor de uma ação, tipicamente uma instância de alguma classe que é uma especialização da classe *Agent-Generic* não necessariamente é um ser vivo. Dependendo do contexto um executor de uma ação pode também ser um animal ou um objeto inanimado (por exemplo, uma jaca que ameaça o teto de um carro estacionado em frente ao RDC).

2.3.5.
COSMO e o Alinhamento entre Ontologias de Alto Nível

A ontologia de alto nível COSMO[11] (*Common Semantic Model*) é uma pesquisa em desenvolvimento, conduzida pelo grupo de trabalho COSMO-WG (*COSMO Working Group*). Esse grupo é uma ramificação do grupo de trabalho ONTACWG (*Ontology and Taxonomy Coordinating Working Group*) que teve seu início em 05 de outubro de 2005. As definições e demais informações resumidas nesta seção são um resumo das consultas realizadas ao Wiki colab.cim3.net/cgi-bin/wiki.pl?CosmoWG.

A COSMO propõe consolidar diferentes ontologias de alto nível como OpenCyc, SUMO, DOLCE entre outras, com o objetivo de acomodar em uma única ontologia diversos conceitos muitas vezes logicamente incompatíveis.

Atualmente o COSMO-WG disponibiliza o que é chamado de Hierarquia de Topo (*Top Level Hierarchy*) onde é possível identificar um alinhamento parcial de conceitos elementares, entre eles, aqueles que capturam parcialmente o conceito de proveniência. Este não é um alinhamento trivial porque existem diferenças estruturais por características intrínsecas a construção de cada ontologia de alto nível, como exemplificado sucintamente pelas Figura 7, Figura 8 e Figura 9.

A Tabela 2 ilustra as consultas as classes Evento (*Event*), Agente-Genérico (*Agent*) e Ação (*Action*) da ontologia COSMO mapeadas nos conceitos abstratos de proveniência. Essas classes são definidas pelo alinhamento entre os conceitos das ontologias de alto nível (seções 2.3.2, 2.3.3 e 2.3.4). Iniciamos neste momento o mapeamento de conceitos de proveniência nos conceitos de ontologias de alto nível.

Tabela 2: Alinhamento parcial de Ontologias de Alto Nível

Conceito abstrato	COSMO	DOLCE	SUMO	OpenCyc
What	*Event*	*Event*	*Process*	*Event*
Who	*Agent-Generic*	*Agent*	*Agent*	*Agent-Generic*
How	*Action*	*Action*	*Intentional Process*	*Action*

[11] http://colab.cim3.net/cgi-bin/wiki.pl?OntologyTaxonomyCoordinatingWG/DefiningVocabulary

Enriqueceremos (seção 2.4) o estudo desenvolvido até aqui com os resultados obtidos a partir da análise de padrões internacionais e, também, de projetos das áreas de museologia, biblioteconomia e comércio eletrônico que possuem classes que capturam o conceito de proveniência. Por hora, exibimos uma segunda prévia (seção 2.3.6) de nossa ontologia parcial para proveniência que inclui os resultados apurados até aqui.

2.3.6. Ontologia Parcial para Proveniência (prévia 2)

A ferramenta cognitiva (seção 1.3) utilizada para o alinhamento (seção 2.3.5) é adotada também na representação da ontologia parcial para proveniência apresentada na Figura 10.

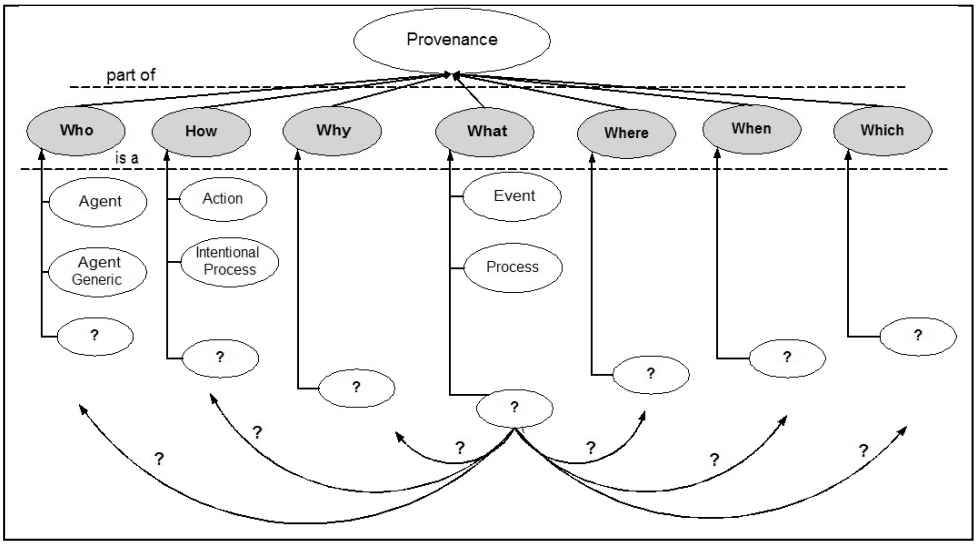

Figura 10: Ontologia parcial para Proveniência (prévia 2)

A partir do alinhamento entre ontologias, associamos o conceito *Agent* ao conceito abstrato *Who*, uma vez que todas as fontes consultadas apresentaram essa invariante de alguma forma. Manteremos o conceito Evento associado à *What*, mas ressalvamos que a noção de processo aparece como relevante nas ontologias de alto nível, como algo que tem uma duração e tem partes (eventos).

Por hora, o conceito *Event* apareceu quase unanimemente, mas deixamos indicado que *Process* poderia estar também associado à noção abstrata *What.* A Figura 10 apresenta então os resultados acumulados. Há dúvidas ainda quanto

ao conceito de processo intencional (*IntentionalProcess*), mas temporariamente o associamos ao conceito abstrato *How*, assim como o fazemos para *Action*.

Lembramos que a ontologia parcial apresentada nesta prévia ou em qualquer outra tem o único e somente objetivo didático e nenhum outro além deste.

2.3.7. Considerações Finais

Semy et al. (2004) cita outras ontologias de alto nível: BFO (*Basic Formal Ontology*), BWW (Bunge-Wand-Weber), GFO/GOL[12] (*General Formalized Ontology/General Ontology Language*), OCHRE (*Object-Centered High Level Reference Ontology*), e a ontologia de topo (*Top Level Ontology*) proposta por John F. Sowa. Adicionalmente, mencionamos também, a ontologia PUO (*Proposed Upper Ontology*) (Cassidy, 2003) e observamos que as ontologias BFO e OCHRE são resultantes do projeto WonderWeb[13], assim como a ontologia DOLCE.

Mapear um domínio específico em uma ontologia de alto nível ou mesmo estendê-la para atendê-lo não é uma tarefa simples. Não há um padrão reconhecido amplamente e também há raras implementações que possam ser provadas (Semy et al., 2004). Essencialmente essa dificuldade é inerente a abordagens teóricas diferentes para a estrutura de construção das ontologias de alto nível. Além disso, há ausência de consenso para eleger uma referência que pudesse ser considerada a absoluta ontologia de alto nível.

De todo modo, ainda é válido considerar as ontologias de alto nível porque seus conceitos são fruto de um estudo aprofundado e não ambíguo. Isto confere alguma garantia que um determinado conceito ancorado em uma ontologia de alto nível pode ser um bom começo para uma definição comprometida com o senso comum.

As ontologias de alto nível oferecem uma fundamentação teórica e são boas fontes de consulta para definições de conceitos. Durante a concepção de novos modelos, essas ontologias devem ser avaliadas mesmo que um

[12] http://www.onto-med.de/

[13] http://wonderweb.semanticweb.org/

mapeamento direto, ideal, não seja adotado. No mínimo, essa avaliação reduzirá o potencial de definição de uma semântica nova para termos com a mesma sintaxe. O resultado natural desse comprometimento é uma semântica mais rica para o modelo, pelo simples fato de referenciar o senso comum que está lastreado por definições precisas pré-existentes em ontologias de alto nível.

2.4. Padrões e Projetos cobrindo o conceito de proveniência

2.4.1. Preliminares

O conceito de evento está diretamente relacionado ao conceito de proveniência porque é o ponto de partida para capturar evidências para registro de episódios ou períodos históricos, conforme ilustrado na Figura 3. Estudamos esse conceito, e classes relacionadas, em dois padrões e três projetos de diferentes áreas com o objetivo de identificar possíveis padrões de modelagem que poderiam nos ajudar a representar o conceito de proveniência e reforçar a direção de integração a priori.

Optamos por dois padrões internacionais (seção 2.4.2) que possuem os conceitos que representam proveniência. Complementando nosso estudo descrevemos (seção 2.4.3) os projetos que apresentam estruturas e conceitos que capturam o conceito de proveniência, mas que não são declarados com o propósito de estudo da proveniência. Esses projetos guardam entre si uma relação de pesquisa interessante que potencializa a capacidade de integração a priori porque suas principais classes se alinham. Além disso, o conceito de evento em todos esses projetos assume um papel central em seus modelos.

Há ainda vários outros projetos (Simmhan et al., 2005) e (Bose & Frew, 2005) relacionados ao estudo do tema proveniência, mas que não serão detalhados neste trabalho. Interessa-nos aqui apenas os projetos que adotam o conceito de evento como central porque consideramos essa característica como sendo representativa na construção da nossa solução.

2.4.2.
Padrões

Os padrões ISO[14] (*International Organization for Standarization*) são reconhecidos em dezenas de países, Brasil, Estados Unidos, Japão e outros, dentre os quais uma parte significativa optou por adotá-los como normas nacionais. Na prática são os órgãos governamentais e empresas desses países que passam a exigir a conformidade com os padrões. O conteúdo desses padrões não está disponível gratuitamente, mas podem ser adquiridos e utilizados para fins comerciais.

Descrevemos as classes dedicadas à proveniência contidas no padrão ISO 14721:2003 (seção 2.4.2.1), sendo essa característica a principal razão de sua seleção na ausência de um padrão dedicado ao tema proveniência. Apresentamos as classes que capturam o conceito de proveniência (seção 2.4.2.2) presentes no padrão ISO 21127:2006 que foi selecionado por simbolizar uma área fundamental na representação de conhecimento histórico (museologia) que tem a preocupação com o contexto da herança cultural. Finalmente, alinhamos conceitos-chave (seção 2.4.2.3) entre esses padrões e as ontologias de alto nível estudadas (seção 2.3).

2.4.2.1.
ISO 14721:2003 (OAIS)

Este padrão internacional, também conhecido como padrão OAIS (*Open Archival Information System*), foi recentemente traduzido e homologado, em 20 de abril de 2007, como a norma brasileira NBR 15472. Adotaremos nesta seção os termos traduzidos da norma brasileira para sistemas espaciais de dados e informações. A norma apresenta um modelo de referência para sistemas abertos de arquivamento de informação (SAAI).

Os sistemas de arquivamento têm como principal objetivo preservar os documentos em seus componentes individuais (conteúdo, estrutura e contexto), bem como sua relação dentro de uma série histórica de eventos. Assim é possível reconstruir fatos, eventos ou transições com a confiança necessária.

[14] http://www.iso.org

Essas são características importantes para a preservação da proveniência e do valor de prova (Thomaz, 2004).

Garrett & Waters (1996) explicam que proveniência se tornou um dos conceitos de organização central da ciência moderna de arquivamento. Acrescentam que a integridade de um objeto de informação é parcialmente garantida ao rastrear de onde ele vem. Para preservar a integridade de um objeto informacional, arquivos digitais devem obrigatoriamente preservar o registro de suas origens e a sua respectiva cadeia de custódia.

Atualmente, no Brasil, o Arquivo Nacional caminha na direção de adotar a NBR 15472:2007 para o desenvolvimento de seus sistemas, que têm, entre outros objetivos, preservar documentos confiáveis, autênticos e acessíveis.

A CTDE (Câmara Técnica de Documentos Eletrônicos) do Conarq (Conselho Nacional de Arquivos) é um órgão nacional, que conta com colaboradores do Arquivo Nacional e define diretrizes para arquivos públicos e privados. A Carta para a Preservação do Patrimônio Arquivístico Digital[15], que alerta para a perda de patrimônio nacional e mundial, é um exemplo. A proveniência é parte importante dessa realidade e tem, na norma brasileira, descrições específicas que podem apoiar o desenvolvimento de políticas e procedimentos nessa direção.

Para compreendermos a orientação da NBR 15472:2007 em relação à proveniência é necessária uma breve menção a alguns conceitos antes de focarmos nos detalhes de interesse desta pesquisa. Destacamos alguns conceitos da norma que aparentemente utiliza o conceito de informação como dado. Com isso, sugerimos ao leitor que interprete informação como dado ao longo desta seção. Por exemplo, informação de proveniência propomos que seja interpretado como dado de proveniência.

De acordo com a ABNT (NBR 15472), informação de proveniência (*provenance information*) documenta o histórico de um conteúdo. Essa informação relata a origem ou a fonte da informação de conteúdo, sua custódia e mudanças desde a sua produção. Por exemplo, o pesquisador principal que registrou os dados e a informação sobre seu arquivamento, manuseio e migração.

[15] Publicada pelo Conarq e UNESCO em 2005

Segundo a ABNT (NBR 15472), a informação de conteúdo é o conjunto de informações-alvo original da preservação. É um objeto de informação composto por seu objeto de dados de conteúdo e sua informação de representação. Por exemplo, uma simples planilha de previsão de vendas, representada, e entendida como, cotas de vendas, mas que não inclua a documentação que explica seu histórico e origem, seus relacionamentos com outros objetos (a informação de descrição de preservação).

A ABNT (NBR 15472) acrescenta que a informação de descrição de preservação é composta por informação de referência, de proveniência, de contexto e de fixidez explicados mais adiante. A Tabela 3 exemplifica essas classificações, considerando um "pacote de software" como a respectiva informação de conteúdo.

Tabela 3: Exemplo de um pacote de informação (informação de conteúdo + informação de descrição de preservação) (NBR 15472:2007)

Tipo de informação de conteúdo	Referência	Proveniência	Contexto	Fixidez
Pacote de software	- Nome - Autor/Produtor - Número da versão - Número de série	- Histórico de revisão - Proprietário da licença - Registro - Copyright	- Arquivo de ajuda - Guia de usuário - *Software* relacionado - Linguagem	- Certificado - Soma de fechamento (checksum) - Criptografia - CRC

Referência é a identificação da informação de conteúdo que inclui uma forma unívoca, não ambígua de identificação, por exemplo, o número de série. Fixidez - traduzida na NBR 15472:2007 para português do termo *Fixidity* da ISO 14721:2003 - pode ser entendida como um mecanismo de identificação para que o objeto de informação de conteúdo não seja alterado de forma não documentada. Em maior detalhe, destacamos nas Figura 11 e Figura 12 os conceitos relacionados às duas informações de descrição de preservação restantes, contexto e proveniência. Apresentamos também as associações com os conceitos abstratos de proveniência.

Note que Evento é um conceito central e que a representação do conceito abstrato *Where*, apesar de não estar presente na figura, poderia estar associado ao conceito Nota (*note*) relacionado a um Evento.

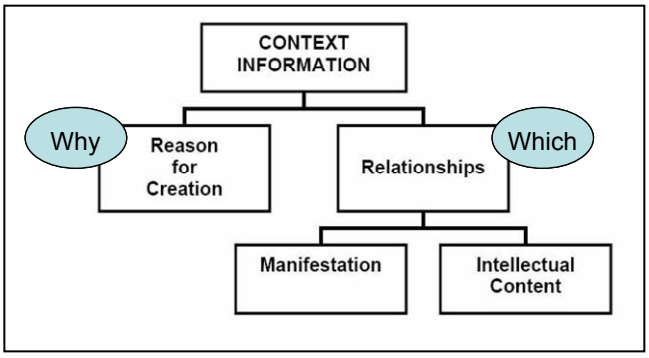

Figura 11: Detalhamento de informação de contexto (Lavoie et al., 2002)

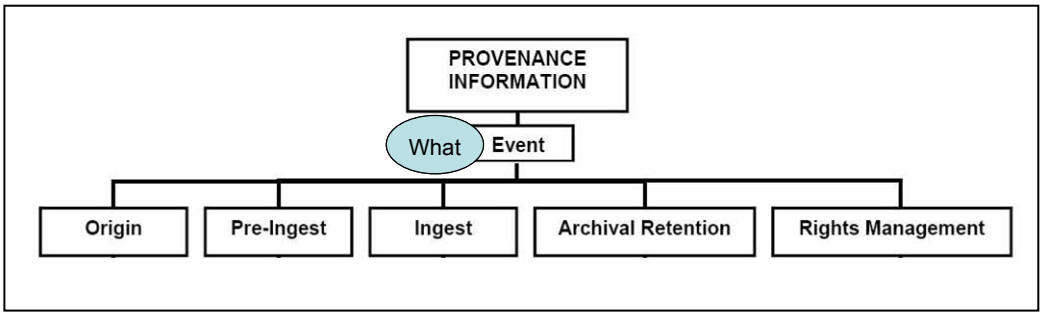

Figura 12: Detalhamento de informação de proveniência (Lavoie et al., 2002)

Lavoie et al. (2002) destaca que *Origin, Pré-Ingest, Ingest, Archival Retention e Rights Management* são tipos de eventos. Portanto, o evento é um elemento central para a captura da informação de proveniência. Destacamos que entre outros atributos de um evento estão *Procedure, Responsible Agency, Date e Note*, que respectivamente poderiam ser mapeados para os conceitos abstratos *How, Who, When e Where*. A informação de proveniência tem um aspecto temporal (*When*) intrínseco compatível com a necessidade de registrar todos os eventos relacionados a um objeto de informação de conteúdo, desde sua criação, se estendendo até o seu estado corrente. Então, a informação de proveniência descreve a informação de conteúdo como uma entidade dinâmica. Garrett & Waters (1996) afirmam que o objeto de conteúdo pode ser visto como resultante de um processo evolutivo, onde o período que se refere à fase de arquivamento pode corresponder apenas a uma pequena parte de seu histórico. Acrescentam ainda que a informação de proveniência pode ser considerada um metadado baseado em eventos (*What*), que deve registrar particularidades dos eventos, bem como respectivos impactos no objeto de informação de conteúdo.

No contexto de preservação digital, a motivação para registrar a informação de proveniência está na necessidade ou requisito de documentar os procedimentos e resultados de processos arquivísticos, ou seja, registrar as ações (*How*). Posicionar tais processos no contexto do ciclo de vida do objeto de informação significa idealmente considerar a história prévia desse objeto, registrando a informação de proveniência correspondente ao período anterior à submissão à entidade arquivística.

Portanto, qualquer entidade, de um modo geral, deve se preocupar com o registro da proveniência para a preservação da integridade da informação. Documentar o que acontece com o objeto de informação está intimamente relacionado a documentar o contexto (Garrett & Waters, 1996) como forma de integridade. O contexto inclui as dimensões: técnica (software, hardware, métodos, normas etc), relacional (relacionamentos com outros objetos) (*Which*), espacial (*Where*), racional (*Why*) entre outras.

Garrett & Waters (1996) ressaltam que a história prévia à admissão do objeto de informação pela entidade arquivística pode ser contada sob três perspectivas: pessoal, corporativa e acadêmica.

Considere que uma mensagem de correio eletrônico sob essas três perspectivas. No contexto estritamente pessoal, pode conter uma piada ou anexo humorístico. Um indivíduo produz e relaciona objetos de informação que dizem respeito à sua vida privada, a deveres e responsabilidades assumidos, ou mesmo ao seu modo de vida. A responsabilidade da entidade arquivística centra-se na capacidade de estabelecer a proveniência de forma concisa, que identifique univocamente o indivíduo como a fonte dos objetos (*Who*) e que rastreie a cadeia de custódia até ao momento prévio à admissão do objeto (Garrett & Waters, 1996).

Por outro lado, uma mensagem eletrônica pode ser o veículo para o contexto da comunicação formal e expressa, por exemplo, na troca de documentos entre chefe e subordinado, e vice-versa. O desafio nesse caso é estabelecer a proveniência dos objetos de informação de forma a preservar a interpretação de políticas e processos, bem como de responsabilidades representadas pelos sistemas de informação, produtos e serviços comercializados.

Por fim, na perspectiva acadêmica, o conteúdo pode representar o resultado de um fluxo de trabalho de um experimento em desenvolvimento. Rastrear os objetos de informação científicos significa identificar os elementos

que instrumentalizaram sua produção e descrever as características de projeto das ferramentas utilizadas.

Garrett & Waters (1996) destacam que para representar o contexto, diferentes eventos devem ser descritos e que múltiplas ocorrências do mesmo evento seriam aquelas que apresentam as mesmas entradas e saídas. A extensão de um evento poderia incluir as informações de equipamentos, software e especificações, entre outras, que são utilizadas durante sua ocorrência. Acrescentam ainda que o resultado de um evento pode ser a geração de um novo objeto de informação, que deve ser igualmente preservado pela entidade arquivística.

Uma forma de preservar é pensar os metadados de proveniência como "transcendentes" a representações para manifestações físicas de um objeto de informação, aplicando a idéia de obra, que é um conceito abstrato para representar uma coleção de objetos que guardam entre si algo em comum. Conceitos abstratos que poderiam ser considerados adequados para estender a norma são os conceitos de *Conceptual Object* ou *Work* apresentados mais adiante (respectivamente nas seções 2.4.2.2 e 2.4.3.1) e harmonizados logo em seguida (seção 2.4.3.4.1).

2.4.2.2. ISO 21127:2006 (CIDOC CRM)

Ainda não há uma iniciativa nacional voltada à tradução do padrão ISO 21127:2006 como uma norma brasileira, analogamente a ISO 14721:2003 (seção 2.4.1.1). A ISO 21127 foi homologada em setembro de 2006, mas sua origem data de 1996.

Em 1996, nascia o projeto de um modelo orientado a objeto, desenvolvido como Modelo de Referência Conceitual (CRM - *Conceptual Reference Model*) pelo Grupo de Padrões de Documentação (*Documentation Standards Group*) do Comitê Internacional de Documentação (CIDOC - *International Committee of Documentation*) do Conselho Internacional de Museus (ICOM - *International Council of Museums*). Esse modelo passou a ser conhecido abreviadamente como CIDOC CRM.

A motivação do CIDOC CRM (atualmente na versão 4.2.2) - e principal papel do padrão equivalente ISO 21127:2006 - é servir de base para mediação de informações de herança cultural. O modelo essencialmente provê uma semântica comum, necessária para que as inúmeras fontes locais e distribuídas

de registro cultural documentem seu acervo de forma adequada, preservando-o como um valioso bem global.

A ISO 21127:2006 pode ser interpretada como uma ontologia onde os conceitos e relacionamentos são aqueles relevantes para a documentação de heranças culturais. É um modelo extenso, com dezenas de classes e centenas de relacionamentos. (Doerr et al., 2006)

A Figura 13 representa uma visão muito simplificada de algumas classes desse padrão. Os retângulos representam classes e as setas representam relacionamentos. As interseções entre os retângulos e algumas elipses - conceitos abstratos de proveniência - identificam os possíveis alinhamentos.

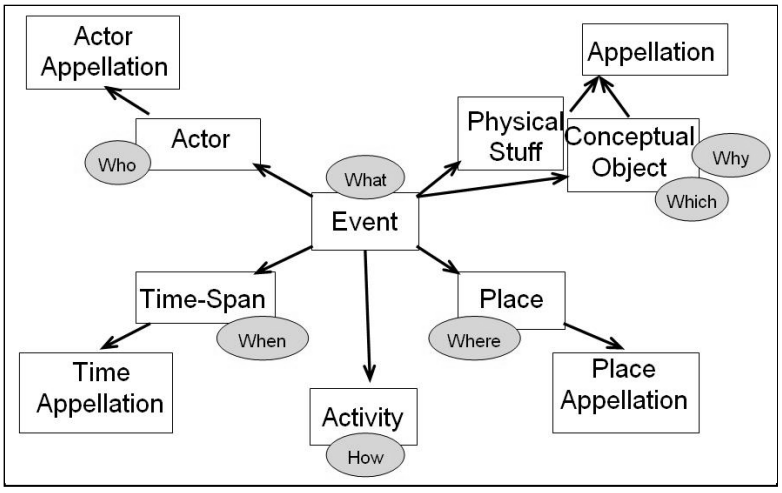

Figura 13: Algumas classes do CIDOC CRM e destaque para o evento no centro

Na Figura 13 destacam-se quatro tipos diferentes de nomeação: *Time Appellation, Actor Appellation, Place Appellation e Appellation*. As três primeiras são especializações da última. Qualquer classe do modelo pode estar relacionada a uma instância da classe *Appellation* ou de sua especialização, e podem ser nomes próprios, frases ou códigos, significativos ou não.

O conceito de proveniência Como (*How*) está inicialmente associado à classe Atividade (*Activity*), que é uma especialização da classe *Event* (Evento) presente na Figura 13.

Crofts et al. (2007) adota o critério de letras maiúsculas seguidas imediatamente de números para representar os identificadores únicos de classes e propriedades. A Figura 14 ilustra a taxonomia para a classe E2 Temporal Entity da qual fazem parte a classe E5 Event e a classe E7 Activity, apresentando

sempre ao lado do nome da classe o seu identificador global exatamente como o fizemos neste parágrafo. As setas indicam a direção de especialização.

A Tabela 4 destaca as propriedades de classes *Event, Actor* e *Activity* do modelo CIDOC CRM respectivamente relacionadas aos conceitos de proveniência Evento, Agente e Ação. A cobertura das propriedades dessas classes equivale a 21% do total de propriedades existentes no modelo. Optamos por deixar a Figura 14 em uma página dedicada, apresentada a seguir, para oferecer ao leitor uma figura mais visível.

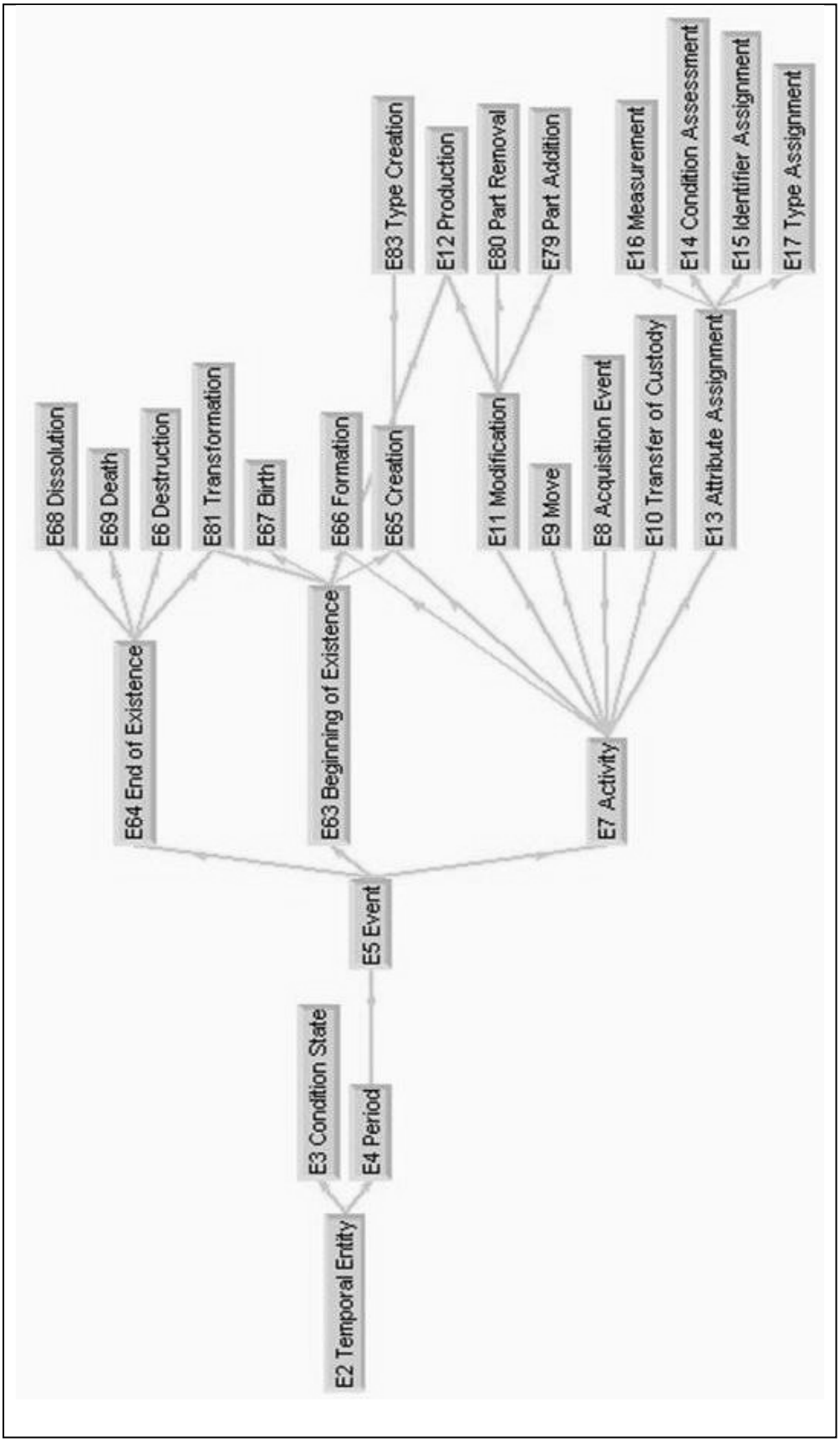

Figura 14: Taxonomia para E2 Temporal Entity (Doerr, 2005)

Tabela 4: Propriedades que apresentam as classes Evento, Agente e Ação como classe-domínio ou classe-imagem (ISO 21127 2006).

Id	**Nome da Propriedade**	**Classe-Domínio**	**Classe-Imagem**
P12	occurred in the presence of (was present at)	E5 Event	E77 Persistent Item
P16	- used specific object (was used for)	E7 Activity	E70 Thing
P11	- had participant (participated in)	E5 Event	E39 Actor
P14	- - carried out by (performed)	E7 Activity	E39 Actor
P22	- - - transferred title to (acquired title through)	E8 Acquisition	E39 Actor
P23	- - - transferred title from (surrendered title through)	E8 Acquisition	E39 Actor
P28	- - - custody surrendered by (surrendered custody through)	E10 Transfer of Custody	E39 Actor
P29	- - - custody received by (received custody through)	E10 Transfer of Custody	E39 Actor
P1	is identified by (identifies)	E1 CRM Entity	E41 Appellation
P131	- is identified by (identifies)	E39 Actor	E82 Actor Appellation
P49	has former or current keeper (is former or current keeper of)	E18 Physical Thing	E39 Actor
P50	- has current keeper (is current keeper of)	E18 Physical Thing	E39 Actor
P51	has former or current owner (is former or current owner of)	E18 Physical Thing	E39 Actor
P52	- has current owner (is current owner of)	E18 Physical Thing	E39 Actor
P74	has current or former residence (is current or former residence of)	E39 Actor	E53 Place
P75	possesses (is possessed by)	E39 Actor	E30 Right
P76	has contact point (provides access to)	E39 Actor	E51 Contact Point
P105	right held by (has right on)	E72 Legal Object	E39 Actor
P107	has current or former member (is current or former member of)	E74 Group	E39 Actor
P109	has current or former curator (is current or former curator of)	E78 Collection	E39 Actor
P15	**was influenced by (influenced)**	**E7 Activity**	**E1 CRM Entity**
P16	- used specific object (was used for)	E7 Activity	E70 Thing
P17	**- was motivated by (motivated)**	**E7 Activity**	**E1 CRM Entity**
P134	- continued (was continued by)	E7 Activity	E7 Activity
P19	**Was intended use of (was made for)**	**E7 Activity**	**E71 Man-Made Thing**
P20	Had specific purpose (was purpose of)	E7 Activity	E7 Activity
P21	had general purpose (was purpose of)	E7 Activity	E55 Type
P125	used object of type (was type of object used in)	E7 Activity	E55 Type

Não identificamos nenhuma classe específica que pudéssemos destacar como representante do conceito abstrato Por que (*Why*). Entretanto há algumas propriedades que poderiam representar esse conceito, com seu identificador marcado em negrito na Tabela 4.

A classe E2 Temporal Entity é disjunta da classe E77 Persistent Item. Instâncias da classe E77 Persistent Item têm uma existência limitada no tempo, mas preservam sua identidade entre eventos, como são as intâncias da classe E39 Actor que é uma especialização da classe E77 Persistent Item.

A classe Atividade (E7 Activity) compreende ações intencionais executadas por instâncias da classe Ator (E39 Actor) que resultam em mudanças de estado documental de sistemas culturais, sociais ou físicos. A noção inclui ações complexas, compostas, de curta ou de longa duração que vão do simples fechamento de uma porta a toda uma guerra. (Crofts et al., 2007)

Alguns exemplos de instâncias da classe E7 Activity são: a Batalha de Estalingrado, a Conferência de Yalta de 1945; a celebração do meu aniversário 12 de junho; a conclusão da escrita (E65) da dissertação “Modelos Conceituais para Proveniência”; a formação (E66) do grupo de pesquisa do projeto de um Centro de Informações.

Destacamos na Figura 15 um exemplo de classes do padrão ISO 21127:2006 expressas com termos do Dublin Core (seção 2.1.2.1).

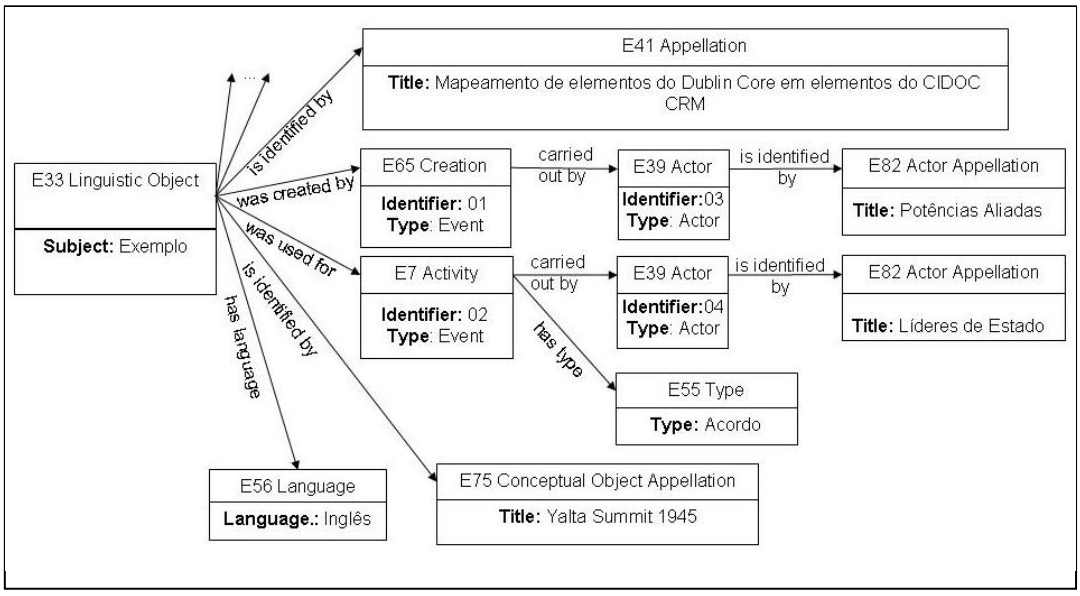

Figura 15: Algumas classes da ISO 21127:2006 expressas com elementos do Dublin Core baseado em (Doerr, 2005)

A representação gráfica da Figura 15 utiliza dois retângulos, um acima do outro. O retângulo de cima contém o identificador único e o nome da classe no padrão ISO 21127:2006. O retângulo debaixo destaca em negrito os nomes dos elementos do Dublin Core e atribui valores de exemplo a cada elemento, logo após o caracter ":" (que é apenas um separador). As setas unidirecionais com ou sem descrições representam relacionamentos do modelo ISO 21127:2006.

A ISO 21127 (2006) e Doerr et al. (2007).contém todas as propriedades do modelo CIDOC CRM distribuídas nas seguintes colunas: Identificador Único da propriedade (Property id), Nome da Propriedade (*Property Name*), Entidade - Domínio (*Entity – Domain*) e Entidade - Imagem (*Entity – Range*).

O conjunto básico de elementos do Dublin Core é limitado (Martin, 1998). Quando o objetivo é representar, de forma mais precisa, a noção de evento e seus relacionamentos, o Dublin Core não oferece uma estrutura de termos suficiente para capturar relacionamentos mais sofisticados e descrever resenhas de fatos históricos. Por outro lado, a ISO 21127:2006 (CIDOC CRM) exige um grande esforço para representação por conter uma grande quantidade de classes e relacionamentos.

Entre esses dois extremos existe um subconjunto da ISO 21127:2006, conhecido como CIDOC CRM Core, ou apenas CRM Core. Sinclair et al. (2006b) argumenta que o CRM Core é um subconjunto condensado de elementos de metadados da ISO 211127:2006 que captura os relacionamentos fundamentais que conectam fenômenos (*things*), conceitos (*concepts*), pessoas (*people*), tempo (*time*) e lugar (*place*), e modela precisamente informações (herança cultural) baseadas em eventos. Por ser condensado, o CRM Core é mais fácil de ser aplicado do que a ISO 211127:2006, preservando a compatibilidade com essa norma. CRM Core não é apenas um formato para elementos de metadados para descoberta de recursos, mas também um esquema simples de resenha de fatos históricos (Sinclair et al., 2006b). Assim como expressamos algumas classes da ISO 21127:2006 como termos do Dublin Core, é possível fazer o mesmo para todo o conjunto de classes do CRM Core, pois é um subconjunto de classes bem mais conciso.

Sinclair (2006) e Doerr (2005) destacam que o CRM Core permite explorar o fato de que os metadados sobre a criação, uso ou descoberta constituem fatos históricos comparáveis às informações encontradas nos respectivos documentos propriamente ditos. A Figura 16 é uma representação gráfica para o CIDOC CRM Core, onde se identificam as classes que representam os principais relacionamentos:

- Identificação (*Identification, Description*): permite estabelecer a relação entre identificadores únicos e um ou mais nomes (*appellations)* ou títulos para esses identificadores;
- Classificação (*Classification e todas as classes terminadas em _Type*): classifica os recursos em tipos a partir de tesauros ou vocabulários controlados. Por exemplo, o tesauro *Getty Art and Architecture Thesaurus*[16] (AAT);
- Participação (*Participation, Event, Role_in_Event, Date, Place*): identifica a participação de pessoas e objetos em eventos. Relaciona itens persistentes a entidades temporais e cria uma noção de História;
- Decomposição de partes (*has_part*, *part_of*): permite a representação de relacionamentos do tipo todo-parte (*whole-part*). Por exemplo, representar uma coleção que consista de um número de objetos, ou um evento que é parte de um evento maior;
- Referências (*Event_Related, refers_to, is_referred_to_by, Relation_To*): são referências entre objetos de informação e qualquer item do mundo real. Por exemplo, um evento referencia outro evento com uma ordem específica que não estritamente a cronológica;
- Similaridade (*show_features_of*): permite representar similaridade ou influências entre objetos e, atividades (*activities*) ou produtos (*products*) e vice-versa. Por exemplo, uma porcelana Coreana contemporânea de uma porcelana Chinesa, pode mostrar traços e formas similares à Chinesa.

Optamos por deixar a Figura 16 em uma página dedicada, apresentada a seguir, para oferecer ao leitor uma figura mais visível.

[16] http://www.getty.edu/research/conducting_research/vocabularies/aat/

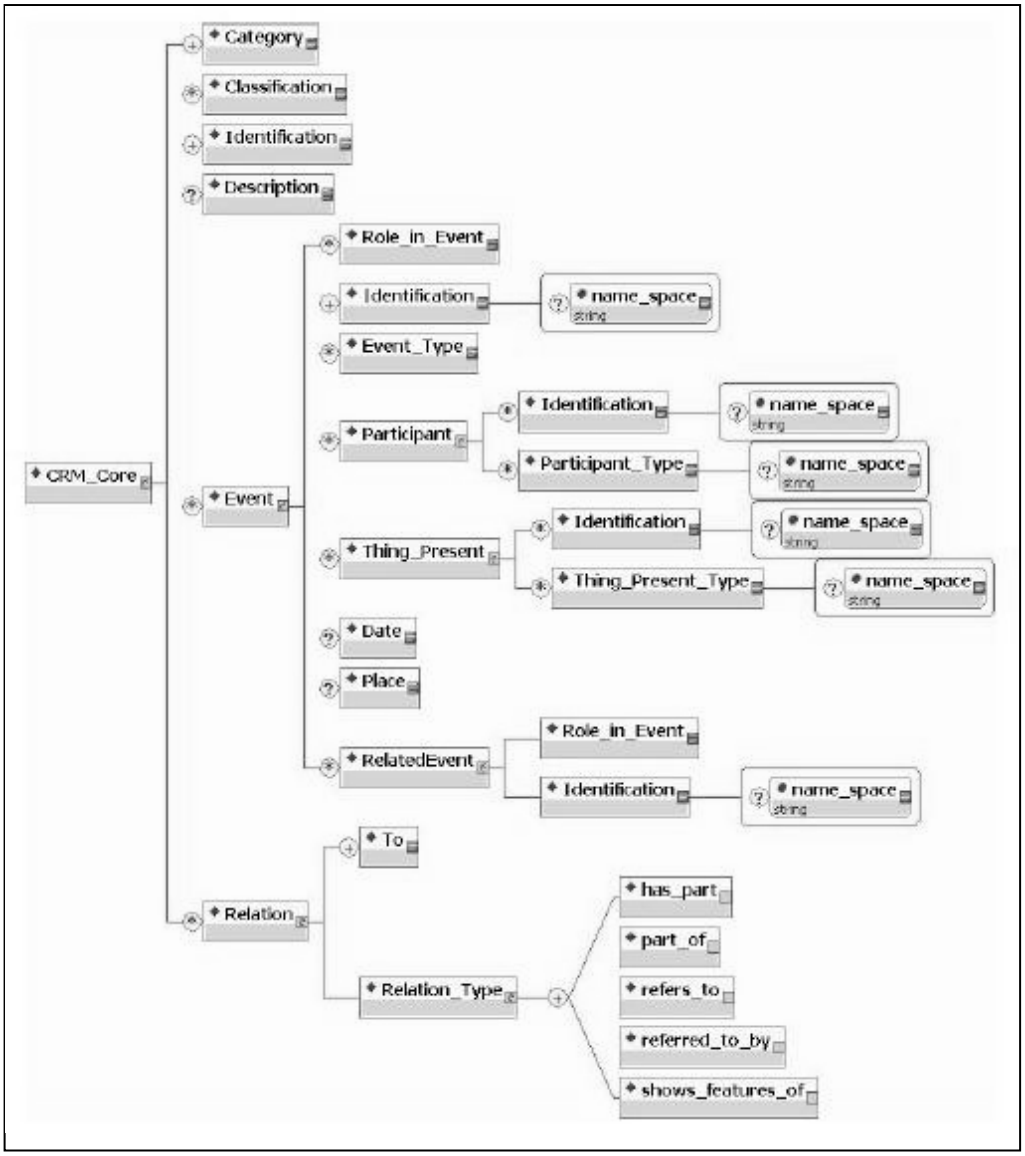

Figura 16: Representação Gráfica do CIDOC CRM Core DTD[17] (Sinclair et al., 2006a)

Optamos por deixar a Figura 17 em uma página dedicada, apresentada a seguir, para oferecer ao leitor uma figura mais visível.

[17] ***Document Type Definition***

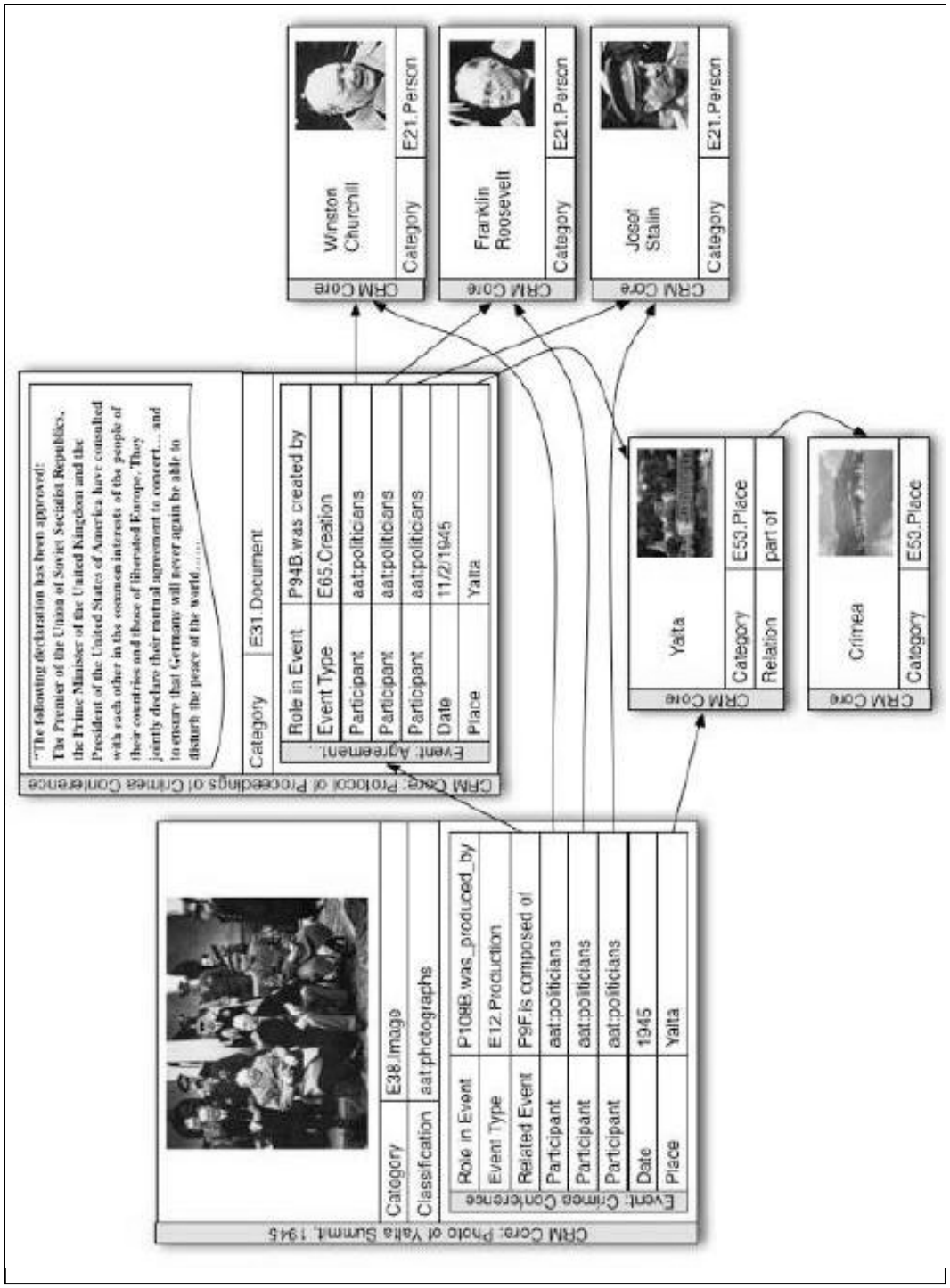

Figura 17: Exemplo utilizando o CRM Core (Sinclair et al., 2006a)

A Figura 17 é um exemplo de representação, utilizando o CIDOC CRM Core, de documentos históricos que tem a si e entre si eventos relacionados. Sem o relacionamento entre os eventos, não seria possível compreender completamente a história apenas com os documentos. Portanto, associados os documentos tem mais valor que meramente agrupados isoladamente. Nessa figura, há seis fotos e um documento texto, totalizando sete objetos de informação representados com classes e relacionamentos do CRM Core.

Nesse exemplo, cada objeto se refere a um documento (imagem ou texto) e contém uma lista de classes, cada uma com um valor de instanciação. A foto mais à esquerda refere-se ao evento “Crimea Conference”, que representa a produção de uma conferência. O documento texto, identificado por seu conteúdo parcial, mais ao topo e centralizado, se refere a outro evento, “Agreement”, que representa a criação do documento texto. O primeiro evento tem uma relação (composto por) com o segundo e permite interpretar que o documento texto foi criado como resultado da conferência identificada na foto.

Há ainda, na Figura 17 três participantes-chave, nos eventos, que estão destacados como as fotos de líderes de estado dos respectivos países organizadores e associados a ambos os eventos. Há também duas fotos de dois locais, Yalta e Crimea, respectivamente a cidade e a república da Ucrânia. A cidade está associada a ambos os eventos e representada como parte da república.

Observamos neste ponto que um conjunto de objetos e relacionamentos, modelado segundo o padrão ISO 21127:2006, pode ser representado em XML. A Figura 18 exemplifica a descrição do objeto informacional mais à esquerda da Figura 17.

A representação da Figura 18 corresponde ao modelo CIDOC CRM Core e descreve a foto[18] da Figura 17, que registra o evento da conferência. Note que vários dos relacionamentos possíveis do CIDOC CRM Core estão presentes na Figura 18. Por exemplo, um evento possui participantes, data e local, que podem ser identificados pelo aninhamento dos respectivos elementos XML <Participant>, <Date> e <Place> dentro do elemento <Event>.

[18] http://en.wikipedia.org/wiki/Image:Yalta_Conference.jpg

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CRM_Core SYSTEM "CRM_Core.dtd">
<CRM_Core>
    <Category>E38 Image</Category>
    <Classification name_space="http://www.getty.edu/research/conducting_research/vocabularies/aat/"
        >photographs</Classification>
    <Identification>http://en.wikipedia.org/wiki/Image:Yalta_Conference.jpg</Identification>
    <Description>Yalta summit in 1945 with Winston Churchill, Franklin Roosevelt and Josef Stalin</Description>
    <Event>
        <Role_in_Event>P138 represents</Role_in_Event>
        <Identification name_space="http://cidoc.ics.forth.gr/crm_core/demo">Crimea Conference</Identification>
        <Event_Type name_space="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs">E7_Activity</Event_Type>
        <Participant>
            <Identification>http://en.wikipedia.org/wiki/Churchill</Identification>
            <Participant_Type name_space="http://www.getty.edu/research/conducting_research/vocabularies/aat/"
                >politicians</Participant_Type>
        </Participant>
        <Participant>
            <Identification>http://en.wikipedia.org/wiki/Franklin_Delano_Roosevelt</Identification>
            <Participant_Type name_space="http://www.getty.edu/research/conducting_research/vocabularies/aat/"
                >politicians</Participant_Type>
        </Participant>
        <Participant>
            <Identification>http://en.wikipedia.org/wiki/Stalin</Identification>
            <Participant_Type name_space="http://www.getty.edu/research/conducting_research/vocabularies/aat/"
                >politicians</Participant_Type>
        </Participant>
        <Date>1945</Date>
        <Place name_space="http://www.getty.edu/research/conducting_research/vocabularies/tgn/">Yalta (inhabited place)</Place>
        <RelatedEvent>
            <Role_in_Event>P9 is composed of</Role_in_Event>
            <Identification name_space="http://cidoc.ics.forth.gr/crm_core/demo">Creating Yalta Agreement</Identification>
        </RelatedEvent>
    </Event>
</CRM_Core>
```

Figura 18: Exemplo XML do CIDOC CRM Core (Sinclair et al., 2006a)

As classes da ISO 21127:2006, ou seu subconjunto CIDOC CRM Core, permitem a criação de uma rede global e interligada de conhecimento, a partir da representação de fatos históricos, tendo como elementos fundamentais eventos e relacionamentos. Algumas dessas classes estão associadas com os conceitos abstratos de proveniência.

Apresentamos o alinhamento dos conceitos (seção 2.4.2.3) de proveniência Evento, Agente e Ação entre ontologias de alto nível (seção 2.3) e os padrões internacionais (seção 2.4.2).

2.4.2.3. Alinhamento entre Padrões e Ontologias de Alto Nível

Casanova et at. (2007) propõe que a modelagem de um banco de dados federado pode utilizar a estratégia a priori de alinhamento de esquemas e sugere o registro da passagem das obras de arte por seus proprietários.

Destaca que o projeto do modelo pode começar pela análise de diferentes ontologias de alto nível (seções 2.3.2, 2.3.3, 2.3.4 e 2.3.5), que incluem conceitos de proveniência como evento, agente e ação. De fato, a proveniência de obras de arte pode ser modelada como uma sequência de eventos resultante de ações de compra e venda em leilões, executadas por agentes como artistas, colecionadores ou museus. A partir desses conceitos, um esquema comum pode

ser adotado pelos bancos de dados da federação para publicar dados que descrevam a proveniência das obras de arte. (Casanova et at., 2007)

Outra alternativa para a modelagem do esquema é a simples adoção de um fragmento de um padrão internacional como a ISO 14721:2003 (seção 2.4.2.1), que oferece classes específicas para o conceito de proveniência, ou como a ISO 21127:2006 (seção 2.4.2.2), adequada para representação de fatos históricos.

Tabela 5: Alinhamento parcial sugerido entre ontologias de alto nível e padrões ISO adaptado de (Casanova et al., 2007)

Conceito abstrato	ISO 21127 (2006)	ISO 14721 (2003)	COSMO	DOLCE	SUMO	OpenCyc
What	*Event*	*Event*	*Event*	*Event*	*Process*	*Event*
Who	*Actor*	*Responsible Agency*	*Agent-Generic*	*Agent*	*Agent*	*Agent-Generic*
How	*Activity*	*Procedure*	*Action*	*Action*	*Intentional Process*	*Action*

2.4.3. Projetos

Os projetos estudados e apresentados nesta seção possuem duas propriedades:

- O conceito Evento assume um papel de destaque em relação às demais classes;
- Há referências cruzadas entre os projetos ou alguma influência direta ou indireta da ISO 21127:2006;

Para cada projeto apresentado, iniciamos sua respectiva seção com uma descrição como introdução, seguida de destaques com visões gráficas ou tabelas dos conceitos mais representativos. Identificamos e analisamos também quais são os conceitos abstratos de proveniência presentes e como estão associados às classes em evidência.

O modelo do projeto CIDOC CRM (seção 2.4.2.2) serviu como base para pesquisas em diferentes domínios. A Figura 19 sinaliza a influência direta ou indireta da ISO 21127:2006 (seção 2.4.2.2) nos projetos estudados (seções 2.4.3.1, 2.4.3.2 e 2.4.3.3).

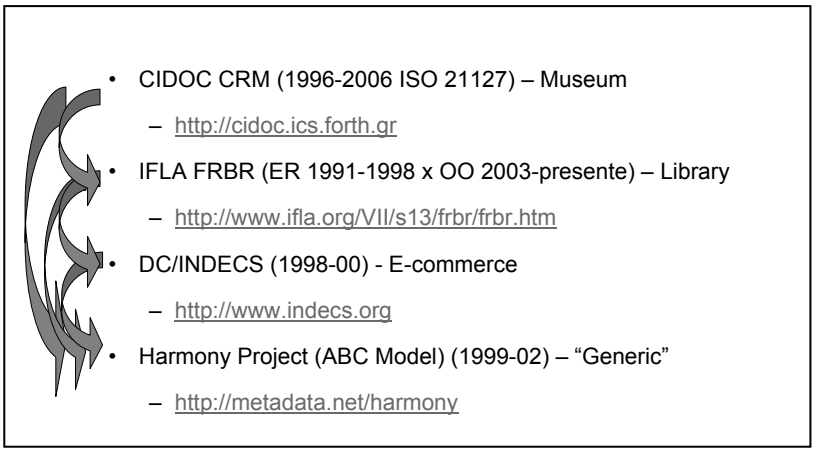

Figura 19: Relação entre projetos estudados e a ISO 21127:2006

2.4.3.1. FRBRoo

O grupo de pesquisa de requisitos funcionais para registros bibliográficos (FRBR - *Functional Requirements for Bibliographic Records*) da IFLA (*International Federation of Library Associations and Institutions*) atuou em duas fases durante o desenvolvimento do projeto do modelo de referência.

A primeira ocorreu entre 1991 e 1995 com o objetivo de pesquisar possíveis codificações para conceitos que capturassem uma visão generalizada do universo de catalogação e pesquisa para bibliotecas (locais públicos ou particulares onde se instalam grandes coleções de livros e outros objetos bibliográficos, para uso público ou particular).

Entre 1996 e 1998, o projeto, em sua segunda fase, dedicou-se à produção de um modelo ER[19] (Entidade e Relacionamento). Desenvolvido com objetivo de se tornar um modelo de referência, é abreviadamente conhecido como FRBR. (FRBR Review Group, 1998).

Carlyle (2006) destaca que o modelo FRBR é uma continuação e uma extensão natural dos modelos usados na catalogação bibliográfica. A Figura 20 apresenta uma visão geral do FRBR com seus três grupos de entidades e

[19] CHEN, Peter P. The Entity-Relationship Model:Toward a unified of data. *ACM transactions on Database Systems*, n. 1, p. 9-36, 1976.

principais relacionamentos. As interseções entre os retângulos e algumas elipses identificam os alinhamentos entre os conceitos abstratos de proveniência e as entidades do modelo FRBR (ER).

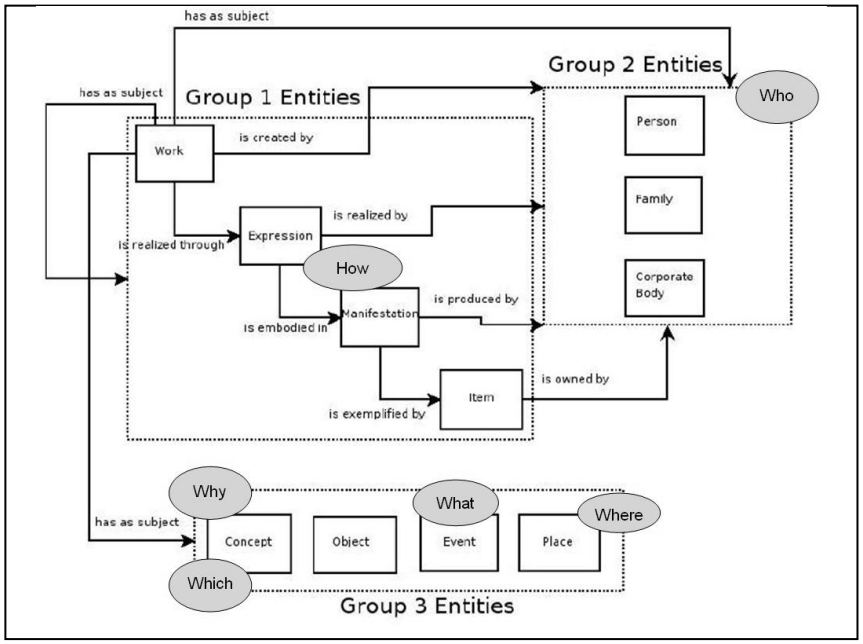

Figura 20: Visão Geral do Modelo FRBR baseado em (FRBR Review Group, 1998)

Silveira (2007) descreve os 3 grupos da Figura 20:

- Grupo 1: são entidades bibliográficas como produtos de trabalho intelectual ou artístico: Obra (*Work*), Expressão (*Expression*), Manifestação (*Manifestation*) e Item (*Item*).
- Grupo 2: são entidades responsáveis pelo conteúdo, produção, disseminação e guarda das entidades do primeiro grupo: Pessoa (*Person*) e Entidade Coletiva (*Corporate Group*).
- Grupo 3: servem como assuntos (*Subject*) de uma obra: Conceito (*Concept*), Objeto (*Object*), Evento (*Event*) e Lugar (*Place*). Um assunto por ser também qualquer entidade do Grupo 1 ou do Grupo 2.

A Figura 21 – traduzida de (Tillett, 2003) - ilustra um exemplo para instâncias do Grupo 1 que captura o núcleo de representação de um objeto bibliográfico e que se divide em imaterial (*Work/Expression*) e material

(*Manifestation*/*Item*). Uma Obra (Work) é uma entidade abstrata e não se refere a nenhum material físico propriamente dito, tampouco a uma linguagem de representação. A realização de uma Obra é representada por uma Expressão, que é na verdade a forma artística ou intelectual, a linguagem adotada para expressar a obra. Quando uma Expressão é fisicamente materializada existe uma Manifestação em algum meio: áudio, vídeo, papel, tela, ambiente etc. Por fim, um Item é definido como uma entidade concreta e se refere a uma ou mais cópias de um mesmo objeto físico.

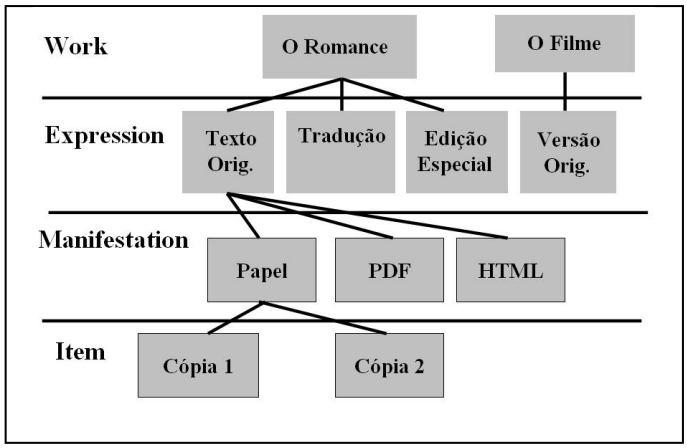

Figura 21: Instanciações para o Grupo 1 baseado em (Tillett, 2003)

O conceito de proveniência Quando (*When*) está presente apenas como o atributo data (*date*) em algumas entidades do modelo, como *Work, Expression, Manifestation, Person e Corporate Group*.

Do Grupo 2 a entidade *Person* possui os atributos nomes, datas e títulos e, a entidade *Corporate Group* possui nome, número, lugar e data. As entidades do Grupo 3 possuem cada uma apenas um único atributo chamado termo. Um termo é uma palavra, frase ou grupo de caracteres usado para designar a entidade. Uma entidade do Grupo 3 pode ser designada por mais de um termo, por exemplo, o termo de um evento pode ser Conferência de Yalta ou 2ª Guerra Mundial.

Por outro lado, é possível identificar atributos, em algumas entidades, que estão diretamente ligados à idéia de proveniência. A entidade Obra (*Work*), por exemplo, possui o atributo “Contexto da Obra” que representa o contexto histórico, social, intelectual ou qualquer outro no qual a obra foi concebida e que pode ajudar a responder o porquê (*Why*) de sua produção. Já à entidade Item

(*Item*) pertence o atributo “Proveniência do Item”, que corresponde ao seu registro histórico de propriedade ou custódia.

Doerr et al. (2007) afirma que a representação temporal oferecida pelo modelo FRBR ER é insuficiente. De fato, não há atributos ou relacionamentos suficientes para capturar o contexto bibliográfico dentro de uma visão histórica mais representativa. Além disso, o modelo é centrado no item (*Item*) bibliográfico, entidade do Grupo 1, e não no evento (*Event*) (entidade do Grupo 3 que pode ser definida apenas como o assunto de uma Obra).

Entidade temporal, ou *perdurant* em ontologias de alto nível (seção 2.3.2), é um conceito central para o modelo CIDOC CRM (seção 2.4.2.2) porque é através de instâncias de classes do tipo dessa entidade que o modelo permite relacionar (Tabela 4) objetos físicos ou não, eventos, intervalos de tempo, ações e agentes.

Em 2003, um grupo para conduzir a harmonização dos modelos FRBR e CIDOC CRM (*International Working Group on* FRBR/CIDOC CRM *Harmonization*) foi constituído com representantes de ambas as comunidades com os objetivos (Doerr et al., 2007):

- Expressar o modelo FRBR com conceitos, ferramentas, mecanismos e convenções oferecidos pelo modelo CIDOC CRM;
- Alinhar ou mesclar os modelos para construir uma solução para a interoperabilidade semântica entre as estruturas de representação para bibliotecas e museus;

Se satisfeitos os dois objetivos, toda informação sobre herança cultural que tenha alguma equivalência deve poder ser recuperada a partir de noções comuns, de forma independente de sua distribuição por distintas fontes de dados.

Em 2006 foi publicada então a primeira versão, v.0.6.7, do modelo FRBR orientado a objeto, abreviadamente FRBRoo, como resultado de uma rígida interpretação lógica das entidades, atributos e relacionamentos expressos no modelo FRBR original. (Doerr et al., 2007).

As entidades temporais (*Temporal Entities*) foram introduzidas no modelo FRBRoo (Doerr et al., 2007) através da declaração de algumas de suas classes como subclasses das seguintes classes do modelo CIDOC CRM: E65 Creation, E12 Production e E13 Attribute Assignment. Com isso, o modelo FRBRoo passa a ter uma representatividade para a pergunta Quando (*When*) antes existente apenas como atributos que não capturavam intervalos de tempo.

Apresentaremos os principais conceitos do projeto DC/INDECS (seção 2.4.3.2) e destacaremos como o projeto FRBR o influenciou.

2.4.3.2. INDECS

O projeto INDECS (*Interoperability of Data in E-Commerce Systems*) (Rust & Bide, 2000) durou 17 meses, entre 1998 e 2000, e teve como objetivo integrar várias iniciativas de diferentes setores da economia. Bearman et al. (1999) destaca que algumas das iniciativas pesquisadas incluem projetos da indústria fonográfica IFPI[20] (*International Federation of the Phonographic Industry*), como o ISRC[21] (*International Standard Recording Code*), iniciativas da área audiovisual, como o ISAN[22] (*International Standard Audiovisual Number*), e da área de produção textual, como ISBN[23] e ISSN[24], trabalhos da Organização Internacional de Propriedade Intectual WIPO[25] (*World Intellectual Property Organization*) e, por fim, o projeto DOI[26] (*Digital Object Identifier*). Foram pesquisados nessas iniciativas, diversos padrões e ferramentas comuns que habilitassem a interoperabilidade de identificadores e metadados.

Como resultado final, foi proposto um *framework* de metadados que representasse a propriedade intelectual no contexto de comércio eletrônico. Esse *framework* está fundamentado na dicotomia conceito/percepção. Divisões do mundo em dois grandes grupos datam da época de Platão que propunha a divisão entre essência e matéria. As ontologias de alto nível (seção 2.3) mais recentemente formalizam tais dualidades, criando diferentes divisões: material/imaterial, concreto/abstrato, durável/perdurável, entre outras possíveis, e que podem considerar a representatividade do tempo, adotando uma representação no espaço 4D, ou não, adotando uma representação no espaço 3D.

---

[20] http://www.ifpi.org

[21] http://www.ifpi.org/content/section_resources/isrc.html

[22] http://www.isan.org/

[23] http://www.isbn.org/

[24] International Standard Serial Number disponível em http://www.issn.org/

[25] http://www.wipo.int

[26] http://www.doi.org/

O modelo INDECS possui três entidades elementares: Conceito (*Concept*), Percepção (*Percept*) e Relação (*Relation*). Conceito é algo concebido no cérebro. Percepção é algo captado por pelo menos um dos cinco sentidos (Rust, 2005). A classe *Relation* é responsável por estabelecer uma ligação entre quaisquer duas ou mais entidades. Esse exemplo nos parece um erro de modelagem porque sugere que todos os relacionamentos binários estariam abaixo de uma única classe do modelo. A documentação disponível para consulta não fornece detalhes suficientes para que essa decisão de projeto seja compreendida, portanto, discutida mais aprofundadamente.

A Figura 22 destaca elementos – *Being, Thing, Time e Place* - do modelo INDECS que se relacionam a partir do Evento. Seres (instâncias da classe Being) e fenômenos (instâncias da classe Thing) são percebíveis. *Time* é a noção de tempo e *Place,* a de lugar. As elipses identificam os conceitos abstratos de proveniência e não fazem parte da figura original. Os eventos são como uma cola entre todas as demais entidades, por isso, desempenham um papel central. Ou seja, aparentemente, eventos são modelados como relacionamentos quaternários, reificados para instâncias de uma classe.

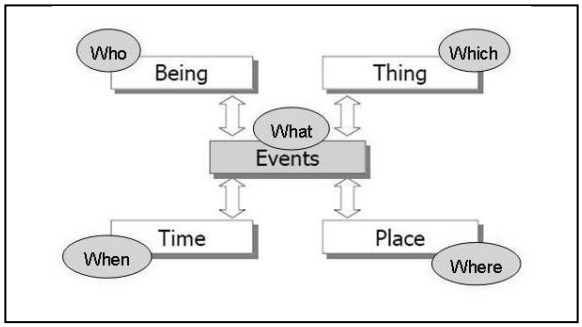

Figura 22: Principais classes e relacionamentos INDECS adaptado de (Bearman et al., 1999) [27]

Rust (2005) chama atenção para a influência – representação de entidades bibliográficas e seus respectivos relacionamentos - do modelo FRBR (seção 2.4.3.1) no modelo INDECS. Acrescenta que manifestações artísticas ou

[27] As imagens desta fonte estão protegidas por Copyright. As utilizamos aqui amparados por http://en.wikipedia.org/wiki/Fair_use

intelectuais e seus respectivos objetos físicos (itens) podem ser generalizados através de abstrações.

No modelo INDECS, o conceito *Manifestation* representa algo percebível, que pode ser transiente (*Performance*) ou persistente (*Fixation*). Rust (2005) destaca que essa interpretação é uma atualização para a versão apresentada na conclusão do projeto INDECS em 2000, onde o termo *manifestatio*n era originalmente usado apenas para referenciar manifestações “fixas”.

Os conceitos *Performance* e *Fixation* foram adotados no padrão MPEG-21 RDD (*Rights Data Dictonary*), DOI e Mi3P (*Music Industry Integrated Identifiers Project*). Por isso, o modelo INDECS foi atualizado para incluir esses conceitos. A nova leitura permite alinhar, na Tabela 6, os respectivos conceitos presentes no modelo FRBR. Originalmente, no modelo FRBR (seção 2.4.3.1), uma manifestação é tida como algo material. Entretanto, estudos mais recentes (Doerr et al., 2003) esclarecem que a interpretação mais adequada é que uma manifestação deve ser algo abstrato.

Tabela 6: Alinhamento parcial entre modelo INDECS e FRBR baseado em (Rust, 2005)

Interpretação	**Conceito do modelo INDECS**	**Conceito do modelo FRBR**
Criações Conceituais (classes)	*Abstraction*	*Work* *Expression* *Manifestaion*
Criações Percebíveis (indivíduos)	*Performance* *Fixation*	*Item*

As criações conceituais e percebíveis são instâncias da classe *Thing* (Figura 22). A compreensão dos conceitos *Abstraction, Performance e Fixation* pode ser alcançada pelo exemplo destacado na Tabela 7, que ilustra em suas últimas 2 linhas uma cadeia de manifestações que levam ao registro de respectivas abstrações.

Tabela 7: Detalhamento da interpretação de abstrações e manifestações do modelo INDECS baseado em (Rust, 2005)

	Abstração (*Abstraction*)	**Manifestação (*Manifestation*)**	
		Performance	*Fixation*
Entidade	Conceito	Percepção	Percepção
Temporalidade	Atemporal	Transiente	Persistente
Estrutura	Pensamentos "Eu concebi"	Ações "Eu executei"	Átomos/Bits "Eu produzi"
Exemplo de Performance, Abstração e "Re-Performance"	(3) O nome temporário da canção	(1) Anotar as cifras de uma canção	(2) O papel com as anotações
	(4) O título oficial	(5) Executar a canção. Gravar a canção.	(6) O CD com a canção gravada

Os números (1) a (6) nas últimas linhas da Tabela 7 ilustram a sequência cronológica típica de como uma performance, que se desdobra em uma abstração, pode por sua vez desencadear outras performances.

A Figura 23 é um detalhamento da Figura 22, apresentando de forma consolidada as três visões do *framework* INDECS. Essas visões não estão detalhadas nesta seção porque essencialmente representam as entidades (conceito, percepção e relação) sob três aspectos diferentes (visão geral, comercial e de propriedade intelectual) (Rust & Bide, 2000). A Figura 23 realça a importância dos eventos e identifica os conceitos abstratos de proveniência. Ressaltamos que a Figura 23 é de 2000, portanto, anterior a discussão que apresentamos nesta seção. Na definição original, o conceito *manifestatio*n era usado apenas para referenciar manifestações "fixas". Fazendo uma leitura mais atual (Rust, 2005) da Figura 21 interprete *expression* como *performance*, *manifestation* troque por *fixation*.

Com a ajuda da Tabela 7 é possível identificar outros conceitos de proveniência que não estavam evidentes (*How* e *Why*). De forma implícita, *Performance* é uma subclasse de *Event*, assim como acontece no modelo CIDOC CRM, e guarda a estrutura de ações (*How*). Já as abstrações, por representarem pensamentos, poderiam ser uma boa fonte para explicar (*Why*) as manifestações transientes e persistentes do modelo INDECS.

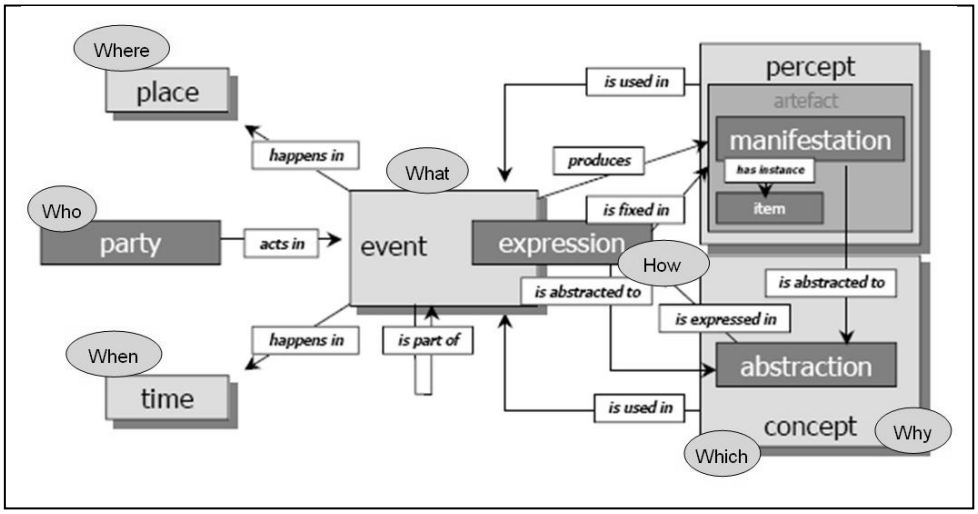

Figura 23: Visão centralizada da classe Evento no modelo INDECS adaptada de (Rust & Bide, 2000) [28]

Tabela 8: Conceitos de proveniência do modelo INDECS

Conceito abstrato	Conceito do modelo INDECS
What	*Event*
Who	*Party*
How	*Performance*
When	*Time*
Where	*Place*
Which	*Abstraction*
Why	

A Tabela 8 propõe o alinhamento entre conceitos abstratos e conceitos que capturam a proveniência do modelo INDECS.

[28] As imagens desta fonte estão protegidas por Copyright. As utilizamos aqui amparados por http://en.wikipedia.org/wiki/Fair_use

2.4.3.3. Harmony

O projeto Harmony é fruto de uma parceria internacional entre a Universidade de Cornell, o DSTC (*Australian Distributed Systems Technology Centre*) e o ILRT (*Institute for Learning and Research Technology*) da Universidade de Bristol. O projeto durou 3 anos, entre meados de 1999 e 2002, com um propósito genérico e, foi influenciado pela norma ISO 21127:2006 (seção 2.4.2.2), pelo projeto FRBR (seção 2.4.3.1) e pelo projeto INDECS (seção 2.4.3.2).

Wiseman et al. (1999) destaca que o objetivo do projeto Harmony foi investigar os aspectos relevantes para descrever recursos multimídia complexos, armazenados em bibliotecas digitais, através da colaboração entre comunidades que estudam padrões de metadados e da pesquisa de um modelo conceitual para representar estruturas complexas e seus respectivos relacionamentos. O modelo conceitual de uma biblioteca digital deve ser projetado para armazenar e recuperar recursos multimídia complexos, que combinam componentes de texto, imagem, áudio e vídeo.

Investigar princípios e conceitos comuns a diferentes projetos é a base para compreender e analisar vocabulários e modelos de metadados. Lagoze & Hunter (2001) afirmam que os resultados dessas investigações, são entre outros, a construção de um guia conceitual para comunidades que projetam novos modelos e de uma metodologia para mapeamento entre diferentes esquemas de banco de dados.

Lagoze & Hunter (2001) elucidam que os projetistas de metadados frequentemente necessitam oferecer consultas que incluam atributos de múltiplas entidades e que cubram perguntas sobre “quem é responsável pelo que, quando e onde”. Com isso, um modelo de metadados deve prover uma fundamentação lógica para raciocínio temporal e relacionar consistentemente agentes, ações e transições de objetos ao longo do tempo.

A recuperação desse tipo de informação envolve noções que capturam a proveniência. A Figura 24 apresenta a taxonomia do modelo conceitual ABC, criado a partir do projeto Harmony, lado-a-lado com os conceitos abstratos de proveniência, representados com elipses.

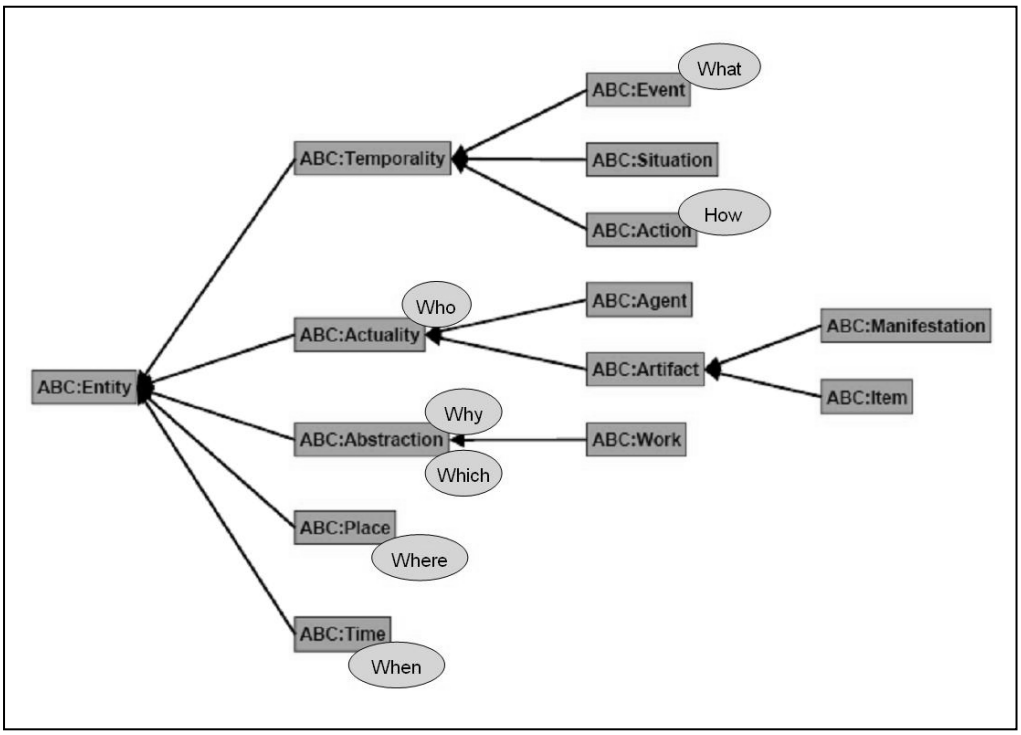

Figura 24: Taxonomia para Modelo ABC baseado em (Doerr et al., 2003)

Observando a Figura 24 note que o modelo ABC também representa os conceitos *Work*, *Manifestation* e *Item*. Esses conceitos são importados do modelo FRBR, que os amadureceu e consolidou ao longo de vários anos de pesquisa (seção 2.4.3.1). Um item pode representar alguma ferramenta ou objeto que foi utilizado durante um evento. Na Figura 24 a classe *Manifestation* está representada como subclasse de *Artifact* e a classe *Work* como uma subclasse de *Abstraction.*

Lagoze & Hunter (2001) destacam que tempo e transições não são frequentemente representadas em modelos centrados em recursos, como são os modelos da área de biblioteconomia tradicional (seção 2.4.3.1). Lagoze (2000) aponta que a modelagem centrada em recursos (*resource-centric*) é inadequada em muitos contextos:

- Em museologia, transições temporais de um objeto (descoberta, classificação, história) são essenciais;
- Em arquivologia, onde proveniência de um objeto é crucial para estabelecer integridade;
- Em propriedade intelectual, onde perguntas como “quem fez o que, onde e quando” são essenciais para registro de propriedade.

Lagoze & Hunter (2001) ressaltam que os diversos modelos centrados em recursos são inadequados para expressar entidades como pessoas, lugares, idéias e especialmente transições temporais. O modelo ABC inclui as noções de Evento (*Event*) e Situação (*Situation*), que capturam respectivamente transições e propriedades existenciais. Lagoze & Hunter (2001) acrescentam que ambas foram adicionadas ao modelo porque estão fundamentadas em modelagem de processos (*Petri Nets*) e extensões temporais para lógica de primeira ordem (*Situational Calculus*).

Os conceitos abstratos de proveniência *What, Who, How* estão associados (Figura 24) respectivamente às classes *Event, Actuality* (superclasse de *Agent*) e *Action*. No modelo ABC, eventos marcam a transição de uma situação para outra e sempre possuem a propriedade temporal. Já as instâncias da classe *Situation* têm sua duração definida implicitamente pelos eventos que a precede e sucede. Agentes são instâncias da classe *Agent* que estão presentes durante um evento ou são executores de alguma ação, podendo ser pessoas, instrumentos, organizações etc. Por fim, ações permitem modelar o conhecimento de envolvimento e responsabilidade de agentes em eventos e denotam um verbo no contexto de um evento. A propriedade *hasAction* tem classe-domínio *Event* e classe-imagem *Action*.

O modelo ABC então é capaz de representar períodos de tempo onde essas e outras entidades se relacionam. As entidades podem apresentar algumas propriedades estáticas e os eventos são transições que alteram propriedades de algumas entidades. Há duas facetas para entidades do modelo ABC, a universal e a existencial, que equivalem respectivamente aos conjuntos de propriedades globais e transientes.

A Figura 25 é uma ilustração gráfica da representação do evento (EV0) correspondente ao nascimento de uma pessoa, onde estão envolvidos três agentes (pai, mãe e o obstetra). O evento precede o fato universal que a criança é do sexo feminino e ao existencial que identifica seu peso. Imagine ainda outros eventos, por exemplo, um evento (EV1) que sucederia a situação (ST0) e que representaria a saída dos pais e da criança do local onde a mãe deu a luz.

Ressaltamos que a influência do projeto INDECS no modelo ABC pode ser evidenciada em sua capacidade de extensão. Hunter (2002) avaliou a importação de classes dos padrões MPEG-7 (descrição de conteúdo multimídia) e MPEG-21 (INDECS/RDD, veja seção 2.4.3.2) para estender o modelo ABC e

torná-lo capaz de representar e recuperar objetos multimídia e respectivos dados sobre propriedade (direito autoral e intelectual).

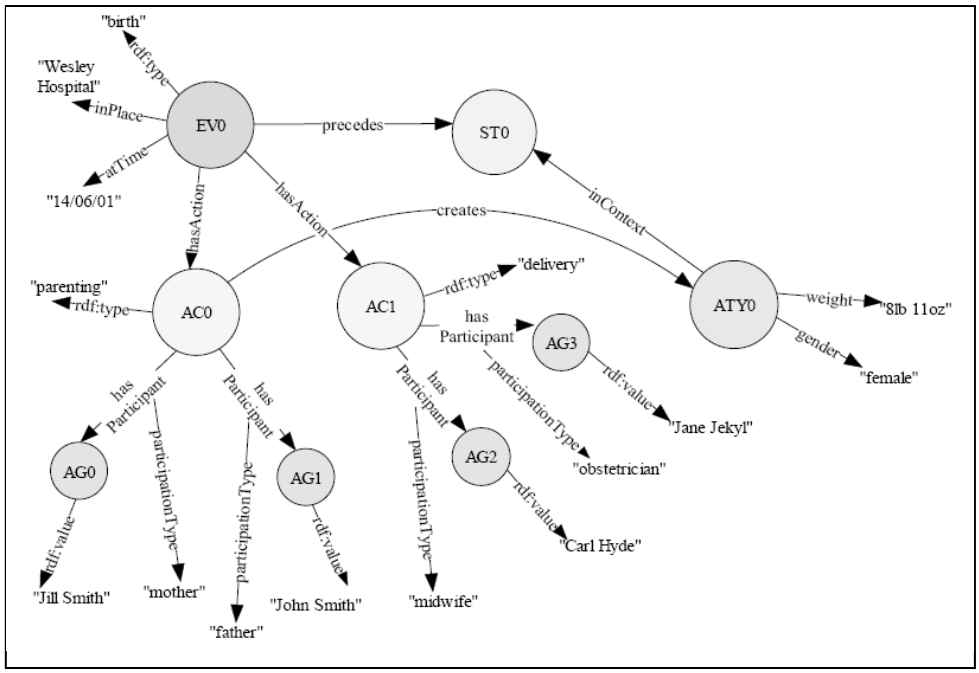

Figura 25: Exemplo de Instanciação do modelo ABC (Lagoze & Hunter, 2001)

A preocupação com interoperabilidade também é uma característica do modelo ABC, pois durante seu desenvolvimento foram exploradas técnicas para alavancar o alinhamento e a representação de seus metadados, a partir de metadados de outros modelos.

Tabela 9: Alinhamento parcial entre os modelos ABC e Dublin Core adaptado de (Hunter, 2000)

Conceitos do ABC	**Elementos do Dublin Core**
Event.Title	*Title*
Event.Description	*Description*
Event.Type	*Type*
Event.Identifier	*Identifier*
Time	*Date*
Place	*Coverage*
Action/Agent	*Creator*
	Publisher
	Contributor
Situation	*Source*
Artifact/Situation	*Relation*

A tradução direta de um esquema de banco de dados em outro quase nunca é trivial e esbarra em desafios sintáticos e semânticos, bem como de expressividade. A Tabela 9 é um exemplo de um possível alinhamento parcial entre o modelo ABC e quase todos os conceitos elementares do Dublin Core. Para este exemplo, considere que a classe *Event* tenha os seguintes atributos: *Title, Description, Type* e *Identifier*.

Em seguida, destacamos o alinhamento (seção 2.4.3.4.1) entre os modelos do projeto FRBRoo (seção 2.4.3.1) e da ISO 21127:2006 (seção 2.4.2.2). Por fim apresentamos alguns resultados do esforço de colaboração que se estabeleceu entre o projeto Harmony e a ISO 21127:2006 (seção 2.4.2.2), destacando o alinhamento de suas classes e propriedades (seção 2.4.3.4.2).

2.4.3.4. Alinhamento entre Projetos e o Padrão ISO 21127

O modelo orientado a objeto FRBR reutiliza muitas classes da ISO 21127:2006 (CIDOC CRM), bem como nomenclaturas dessa norma. Por exemplo, suas classes E4 Period e F11 Event alinham-se diretamente. No projeto INDECS[29] (*INteroperability of Data in E-Commerce Systems*), as cinco principais classes se relacionam umas com as outras a partir da classe Evento, que é central ao modelo.

Todos esses três projetos analisam o processo de como os fenômenos (*things*) se transformam e não se limitam à sua interpretação e representação. Acima de tudo, desejam expressar as transformações ao longo do tempo. Por isso, todos definem prioritariamente a descrição de relacionamentos como a base para seus metadados.

O projeto Harmony[30] - modelo ABC - é influenciado por todos esses três projetos. O modelo ABC é motivado por bibliotecas digitais, mas propõe-se a ser genérico.

Destacamos os alinhamentos dos conceitos (seções 2.4.3.4.1 e 2.4.3.4.2) oriundos dos projetos que foram influenciados diretamente pela ISO 21127:2006, não nos limitando ao conceito Evento, mas incluindo também outros conceitos associados aos conceitos abstratos de proveniência.

[29] http://www.indecs.org

2.4.3.4.1. FRBRoo / CIDOC CRM

Doerr et al. (2007) argumenta que os modelo FRBR ER e OO continuam a coexistir: o primeiro com propósitos pedagógicos, enquanto que o segundo está voltado ao desenvolvimento de aplicações governadas por ontologias.

O modelo FRBRoo está em desenvolvimento e, atualmente, está disponível na versão 0.8.1c. Doerr et al. (2007) apresenta 41 classes (pág. 17) e 56 propriedades (pág. 20 e 21), que mapeiam o modelo de entidade-relacionamento original. Além dessas, o modelo orientado a objeto re-usa 44 classes (pág. 87) e 45 propriedades (pág. 88) do modelo CIDOC CRM (Doerr et al., 2007).

O padrão de identificação das classes do modelo FRBRoo é semelhante ao adotado no modelo CIDOC CRM. As classes definidas têm seu identificador começado pela letra “F” e as propriedades têm sua identificação iniciada com a letra “R”. Essas letras correspondem respectivamente às letras “E” e “P” da convenção adotada pelo CIDOC CRM. (Doerr et al., 2007)

O modelo FRBRoo é um pouco menor que o modelo CIDOC CRM, mas ainda assim exige um grande esforço cognitivo pois apresenta um quantidade de classes e propriedades da ordem de centenas. A Figura 26 apresenta o alinhamento entre as principais classes dos modelos FRBRoo e CIDOC CRM. Destacamos ao lado de cada classe os conceitos de proveniência associados (Bouef & Doerr, 2007).

A Tabela 10 destaca as classes do modelo FRBRoo alinhadas diretamente com as classes do modelo CIDOC CRM. As classes F1 Work, F2 Expression, F4 Manifestation Singleton e F5 Item do modelo FRBRoo (Figura 26) - correspondentes às entidades do Grupo 1 no modelo FRBR ER (seção 2.4.3.1) - não têm alinhamento direto com classes do modelo CIDOC CRM, mas são respectivamente especializações das classes importadas E28 Conceptual Object, E73 Information Object, E24 Physical Man-Made thing e E84 Information Carrier.

[30] http://metadata.net/harmony/

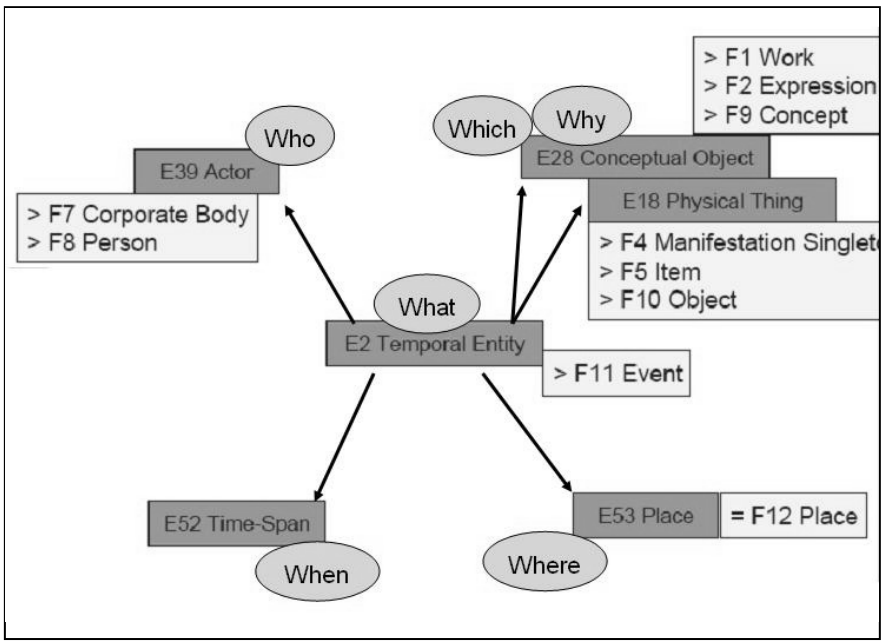

Figura 26: Alinhamento de classes dos modelos FRBRoo e CIDOC CRM (ISO 21127:2006) baseado em (Bouef & Doerr, 2007)

Tabela 10: Alinhamento parcial de classes entre modelos FRBRoo e CIDOC CRM (Doerr et al., 2007)

FRBRoo	**CIDOC CRM**
F11 Event	E4 Period
F8 Person	E21 Person
F9 Concept	E28 Conceptual Object
F7 Corporate Body	E74 Group
F13 Name	E41 Appellation
F12 Place	E53 Place
F10 Object	E18 Physical Thing

A Tabela 11 apresenta a hierarquia de especialização do modelo FRBRoo para a classe F11 Event (E4 Period). As classes em negrito são classes nativas do modelo FRBRoo e as demais foram importadas parcialmente do modelo CIDOC CRM. Parcialmente aqui quer dizer que foram mantidas apenas as

propriedades estritamentes essenciais ao domínio de bibliotecas, eliminando as demais.

Tabela 11: Hierarquia de especialização do modelo FRBRoo para a classe F11 Event

E4	Period = F11 Event
E5	Event
E7	Activity
F52	**Performance**
E11	Modification
E12	Production
F31	**Expression Creation**
F40	**Carrier Production Event**
F44	**Reproduction Event**
E13	Attribute Assignment
F33	**Identifier Assignment**
F36	**Representative Manifestation Assignment**
F37	**Representative Expression Assignment**
E65	Creation
F30	**Work Conception**
F31	**Expression Creation**
F55	**Recording Event**
F45	**Publication Event**

A partir das classes identificadas em negrito na Tabela 11, destacamos nas Tabela 12 e Tabela 13 as propriedades relacionadas, criadas no modelo FRBRoo, ratificando a importância do conceito de evento. As propriedades das Tabela 12 e Tabela 13 correspondem a 50% das propriedades novas criadas para o modelo FRBRoo.

Tabela 12: Propriedades que representam a importância da classe Evento no modelo FRBRoo

Id	Nome da Propriedade	Classe-Domínio	Classe-Imagem
R64	performed (was performed in)	F52 Performance	F50 Performance Plan
R45	created (was created by)	F31 Expression Creation	F4 Manifestation Singleton
R49	created a realisation of (was realised through)	F31 Expression Creation	F46 Individual Work
R38	produced things of type (was produced by)	F40 Carrier Production Event	F3 Manifestation Product Type
R39	followed (was followed by)	F40 Carrier Production Event	F39 Production Plan
R40	used as source material (was used by)	F40 Carrier Production Event	F41 Publication Expression
R41	produced (was produced by)	F40 Carrier Production Event	F5 Item
R59	reproduced (was reproduced by)	F44 Reproduction Event	E84 Information Carrier
R60	produced (was produced by)	F44 Reproduction Event	E84 Information Carrier
R24	assigned to (was assigned by)	F33 Identifier Assignment	E1 CRM Entity
R25	assigned (was assigned by)	F33 Identifier Assignment	F14 Identifier
R26	used constituent (was used in)	F33 Identifier Assignment	F13 Name

Tabela 13: Propriedades que representam a importância da classe Evento no modelo FRBRoo (continuação da Tabela 12)

Id	Nome da Propriedade	Classe-Domínio	Classe-Imagem
R52	used rule (was the rule used in)	F33 Identifier Assignment	F16 Identifier Rule
R53	assigned (was assigned by)	F36 Representative Manifestation Assignment	F4 Manifestation Singleton
R31	assigned to (was assigned by)	F36 Representative Manifestation Assignment	F2 Expression
R32	assigned (was assigned by)	F36 Representative Manifestation Assignment	F3 Manifestation
R33	assigned to (was assigned by)	F37 Representative Expression Assignment	F21 Complex Work
R34	assigned (was assigned by)	F37 Representative Expression Assignment	F2 Expression
R16	carried out by (performed)	F36 Representative Manifestation Assignment	F28 Bibliographic Agency
R17	carried out by (performed)	F37 Representative Expression Assignment	F28 Bibliographic Agency
R21	initiated (was initiated by)	F30 Work Conception F1 Work	
R22	created (was created by)	F31 Expression Creation	F2 Expression
R45	created (was created by)	F31 Expression Creation	F4 Manifestation Singleton
R49	created a realisation of (was realised through)	F31 Expression Creation	F46 Individual Work
66	recorded (was recorded through)	F55 Recording Event	E7 Activity
67	created (was created through)	F55 Recording Event	F56 Recording
68	realised (was realised through)	F55 Recording Event	F53 Recording Work
55	created production plan (was created by)	F45 Publication Event	F39 Production Plan

Um evento deixa de ser apenas um assunto que descreve, através de um termo, uma obra bibliográfica para ser representado pela classe F11 Event, que é uma especialização da classe E2 Temporal Entity do modelo CIDOC CRM, relacionando conceitos de Agente (E49 Actor) e Ação (E7 Activity) (seção 2.4.2.2). Estas últimas duas classes são classes importadas também para o modelo FRBRoo.

2.4.3.4.2. Harmony / CIDOC CRM

O modelo ABC do projeto Harmony pode ser entendido como uma extensão compatível com o modelo CIDOC CRM (seção 2.4.2.2), onde todos os estados devem ser explicitados. Doerr et al. (2003) comenta que o resultado das sete reuniões ao longo de um ano de trabalho entre membros de ambos os projetos é fruto de um processo conciso e comprometido ontologicamente com a harmonização dos conceitos.

Ambos os modelos se propõem a oferecer uma representação de mudanças ao longo do tempo. Entretanto, a natureza dessa mudança é entendida de forma distinta por cada um deles. O modelo ABC foca a atenção em como os objetos mudam ao longo de um período, enquanto que o modelo CIDOC CRM pretende representar as mudanças de contexto e atribuições, sem necessariamente focar atenção a mudança do objeto propriamente dita.

Outra diferença importante é que a classe E39 Actor do modelo CIDOC CRM descreve que a noção está associada a uma pessoa ou grupo durante sua existência. Já o modelo ABC define que agentes - instâncias da classe *Agent* - são fases (*phaseOf*) de uma pessoa ou máquina atuando durante um evento. As definições não são iguais, mas cada uma isoladamente parece lógica. A pergunta é: qual é a definição mais apropriada e em qual contexto?

O esforço de harmonização ou apenas de alinhamento é possível, mas árduo. Doerr et al. (2003) identifica aproximações entre os principais conceitos desses modelos e propõe equivalência entre algumas de suas classes (Tabela 14) e propriedades (Tabela 15).

Dentre os resultados do projeto Harmony que influenciaram alterações no projeto CIDOC CRM (seção 2.4.2.2), Doerr et al. (2003) salienta que as classes *Abstraction* e *Actuality* correspondem à classe *Persistent Item* (ou *endurants*) no modelo CIDOC CRM.

Tabela 14: Alinhamento parcial de classes entre modelos ABC e CIDOC CRM (Doerr et al., 2003)

ABC	**CIDOC CRM**
Entity	*Entity*
Temporality	*Temporal Entity*
Event	*Event*
Action	*Activity*
Artifact	*Man-Made Object*
Place	*Place*
Time	*Time-Span*

Além dos alinhamentos (classes consideradas equivalentes), existem outros resultados interessantes da harmonização entre esses modelos, como a ratificação da interpretação da classe *Manifestation* como conceito abstrato (conforme já destacado na seção 2.4.3.2).

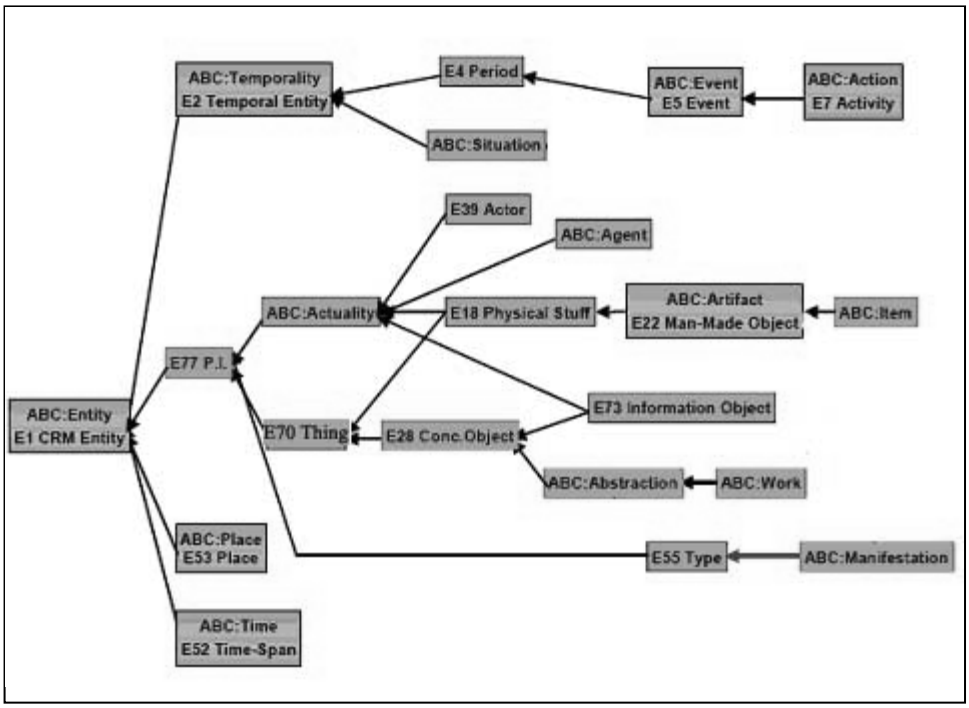

Figura 27: Alinhamento parcial das classes dos modelos ABC e CIDOC CRM (ISO 21127:2006) baseado em (Doerr et al., 2003)

A Figura 27 apresenta a taxonomia harmonizada e inclui as classes que não tem alinhamento direto, complementando os alinhamentos identificados pela

Tabela 14. As classes são representadas como retângulos e os arcos simbolizam a direção da generalização. Os nomes das classes da Figura 27 são apresentados iniciados pelo seu respectivo identificador. Para o modelo ABC o identificador é fixo e igual a “ABC:”. Por exemplo, “ABC: Entity” é a classe mais geral do projeto Harmony. Para o modelo CIDOC CRM (ISO 21127:2006) o identificador começa com a letra “E” seguida imediatamente de um número único. Por exemplo, “E1 CRM Entity” é a classe mais geral desse modelo. Quando ambos os nomes estão dentro do mesmo retângulo significa que estão alinhados diretamente. Quando estão em retângulos diferentes ligados por uma seta significa a sugestão de harmonização proposta. Por fim, a classe E77 P. I. é a abreviação da classe E77 Persistent Item.

Tabela 15: Alinhamento parcial de propriedades entre modelos ABC e CIDOC CRM (Doerr et al., 2003)

ABC	**CIDOC CRM**
hasParticipant	*hadParticipant*
hasPresence	*occurred in the presence of*
destroys	*took out of existence*
creates	*brought into existence*
usesTool	*used specific object*

A Figura 28 apresenta a harmonização das propriedades e inclui as propriedades que não tem alinhamento direto, mas que complementam as apresentadas na visão destacada pela Tabela 15. As classes são representadas como retângulos e as propriedades como ovais. Se as classes estão alinhadas diretamente, então estão dentro do mesmo retângulo. As classes adotam a mesma nomenclatura descrita para a Figura 27. Analogamente, os nomes das propriedades na Figura 28 são apresentados iniciados pelo seu respectivo identificador. Para o modelo ABC o identificador é fixo e igual a “ABC:” e as propriedades começam com letra minúscula. Por exemplo, “ABC: involves” é uma propriedade do modelo. Para o modelo CIDOC CRM (ISO 21127:2006) o identificador da propriedade começa com a letra “P” seguida imediatamente de um número único. Por exemplo, “P12 ocurred in the presence of (was present at)” é uma propriedade do modelo CIDOC CRM que na harmonização é sugerida como propriedade “mãe” da propriedade ABC: involves. As setas que ligam duas classes ou duas propriedades identificam a direção de generalização. As setas

que ligam uma classe a uma propriedade ou uma propriedade a uma classe respectivamente identificam a origem da classe-domínio e o destino da classe-imagem onde opera a propriedade.

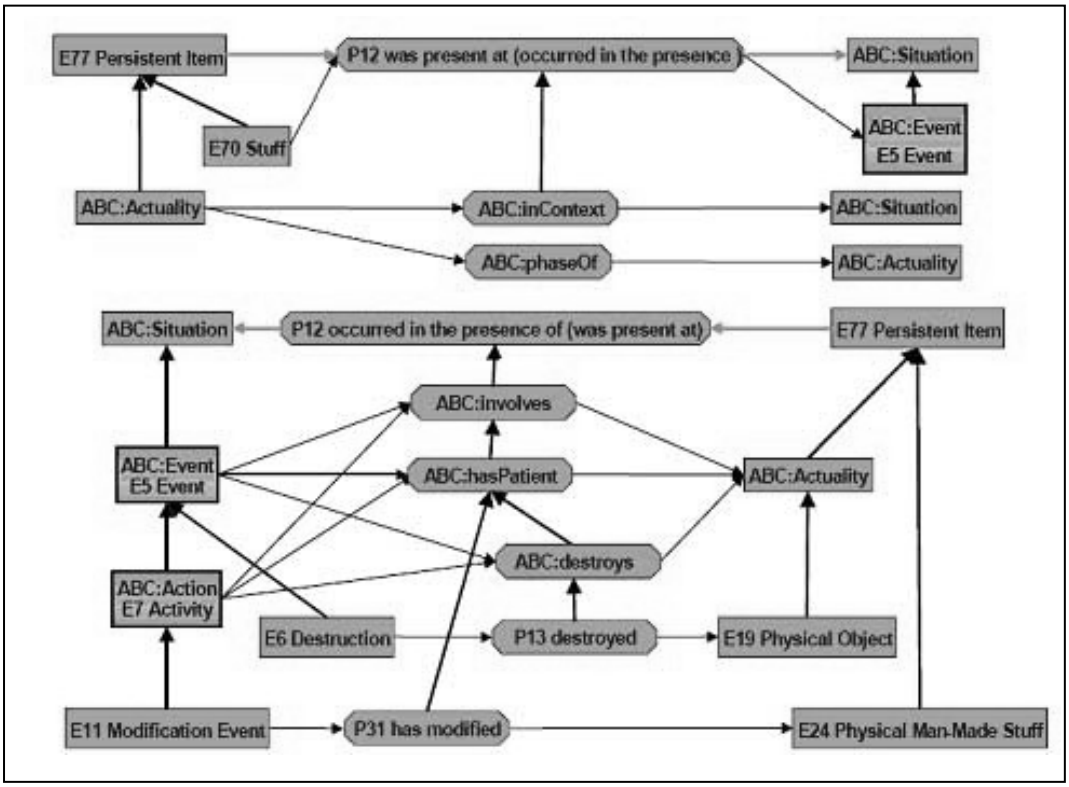

Figura 28: Alinhamento parcial de propriedades entre modelos ABC e CIDOC CRM (ISO 21127:2006) baseado em (Doerr et al., 2003)

O modelo ABC - ao contrário do modelo CIDOC CRM - inclui a possibilidade de representar períodos de tempo durante os quais alguns objetos mantêm estáticas suas propriedades. O primeiro também define que há estados e situações entre eventos, enquanto que o modelo CIDOC CRM representa apenas eventos e descreve transições. Mas o modelo CIDOC CRM oferece algo similar à descrição de um estado a partir da noção de *Conditon State*, que descreve a duração de uma fase, onde é seguro assumir que as propriedades de um objeto são estáveis.

Ambos os modelos apresentam vantagens e desvantagens (Doerr et al., 2003):

- Eventos em geral são conhecimentos primários, mas o testemunho absoluto de um estado é raro. Estados normalmente são resultantes de inferências;

- Representar estados apenas através de transições podem ocasionar o descarte de conhecimento, como a não identificação de outras transições intermediárias;
- Transições podem ser inferidas a partir de dois estados observados (por exemplo, “construído em 1944” e “destruído em 1945”).
- Estados são subjetivos e relativos ao contexto. Se alguma propriedade é alterada, mas o novo valor não é representativo, possivelmente será considerada estática;
- O conhecimento sobre um estado, ainda que inferido ou subjetivo pode ser relevante.

Por fim, exibimos a terceira e última prévia (seção 2.4.4) de nossa ontologia parcial para proveniência que inclui os resultados mais relevantes – segundo nossa visão - apurados até aqui.

2.4.4. Ontologia Parcial para Proveniência (prévia 3)

A ontologia parcial da Figura 29 é o resultado acumulado, dos conceitos que consideramos mais relevantes, dos diferentes modelos estudados.

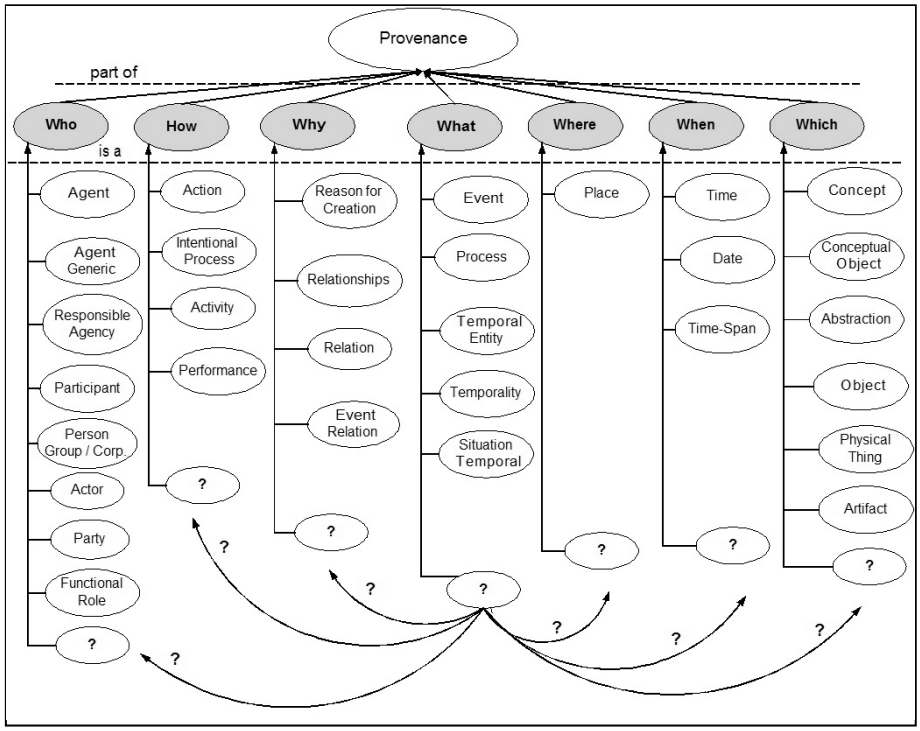

Figura 29: Ontologia parcial para Proveniência (prévia 3)

O propósito deste resultado é meramente didático para ajudar o leitor a manter uma referência dos principais conceitos explorados. As propriedades não são apresentadas na Figura 29 para evitar a sobrecarga de informação, mas lembramos que são fundamentais para a análise de fragmentos.

A partir do contexto para proveniência acumulado (capítulo 2), esclareceremos as decisões de projeto e apresentaremos o modelo conceitual genérico de proveniência em detalhes a seguir (capítulo 3).

2.4.5. Considerações Finais

Ainda não existe um padrão ISO publicado, especifico e dedicado a proveniência. Há apenas um único com status em desenvolvimento. O padrão ISO/NP 8000-120 (*Data quality - Part 120: Master data: Provenance*) alcançou o estágio 10.99 (novo projeto aprovado) em 14 de novembro de 2007. Porém, durante o curso desta pesquisa, sua documentação ainda estava indisponível para comercialização. Esse padrão é parte da ISO 8000 em desenvolvimento, como norma de qualidade da informação.

A pressão competitiva no mercado mundial - tal qual existe na área de rastreabilidade de alimentos (Bechini et al., 2005) e (Dorp, 2004) - é o principal catalisador para que um padrão dedicado a proveniência seja homologado e amplamente adotado nos próximos anos.

Enquanto não há um padrão dedicado a proveniência, uma alternativa para a modelagem do esquema é a simples adoção de um fragmento de um padrão internacional como a ISO 14721:2003 (seção 2.4.2.1), que oferece classes específicas para o conceito de proveniência, ou como a ISO 21127:2006 (seção 2.4.2.2), adequada para representação de fatos históricos.