

DESCRIÇÃO E PESQUISA: METADADOS COMO INFRA-ESTRUTURA

Michael K. Buckland
Universidade da Califórnia

RESUMO

O uso original de metadados é para descrever documentos. XML, Dublin Core e os registros de catálogo de biblioteca MARC são exemplos. O nome “metadados” (além de ou com dados) e a definição popular “dados sobre dados” têm base nesse uso. Um segundo uso de metadados é para formar estruturas de organização por meios nos quais os documentos possam ser organizados. Essas estruturas podem ser utilizadas tanto para pesquisar documentos individuais como para identificar padrões dentro de uma população de documentos. O segundo papel dos metadados envolve uma inversão da relação entre documento e metadados. Essas estruturas podem ser consideradas infra-estrutura.

Palavras-Chave: Metadados; XML; Dublin Core; MARC.

O PRIMEIRO PROPÓSITO DOS METADADOS: DESCRIÇÃO

O termo “metadados” é utilizado para denotar “dados sobre dados” e seu propósito original é descrever documentos. (Aqui não distinguimos entre *dados* e *documentos*.) Há tipos diferentes de metadados descritivos: técnico (para descrever formato, padrões de codificação, etc.); administrativo (para descrever direitos de propriedade intelectual, condições de acesso, etc.); e conteúdo (temática, escopo, autoria, etc.). Essas descrições caracterizam e explicam os dados. Metadados ajudam a compreender o que são os dados e como utilizá-los (CAPLAN, 2003; HAYNES, 2004).

Os metadados possuem dois componentes: Um formato e um conjunto de valores. XML, Dublin Core e os registros de catálogo de biblioteca MARC são formatos conhecidos e são associados a padrões específicos para especificar os tipos de descrições que podem ser utilizados com eles. A descrição pode ser muito útil, mesmo se terminologia idiossincrática for utilizada. Quase

qualquer descrição é melhor do que nenhuma. Entretanto, é sempre fortemente recomendado que metadados descritivos sigam formas padronizadas, por exemplo, utilizar um formato padrão e terminologia amplamente utilizada. O uso de formatos padronizados para armazenamento e exibição torna o uso de metadados mais fácil. O uso de vocabulários padrão possui a vantagem da consistência e auxilia na compreensão.

Toda a descrição é uma atividade de linguagem mesmo se uma notação artificial, como o sistema de Classificação Decimal de Dewey (*Dewey Decimal Classification*), for utilizada. A descrição é sempre e necessariamente de base cultural, pois descrições são baseadas em conceitos, definições e compreensões que têm desenvolvido em uma comunidade.

Quando você pesquisa documentos, especialmente documentos digitais, você está propenso a examinar metadados descritivos para compreender que tipo de documento ele é, sobre o que é e como utilizá-lo. Este processo é semelhante à forma com que uma pessoa olha para a capa de um livro para ajudar a avaliar o texto contido nele.

INFRA-ESTRUTURA

“Infra-estrutura” é um termo coletivo para as partes subordinadas de um empreendimento. A palavra foi inicialmente utilizada para referir-se a recursos fixos utilizados para transporte e operações militares e tem sido gradualmente estendida para incluir serviços subordinados para, ou em apoio a, o desempenho de uma tarefa central. Minimamente, viajar de trem requer trilhos, uma locomotiva e vagões, mas um serviço ferroviário eficaz e confiável depende de outros recursos auxiliares: sistemas de sinalização, compra e venda de bilhetes e comunicação entre as estações, depósitos de combustível, uma hierarquia administrativa, a publicação de tabelas de horários e assim por diante. O nome coletivo para esses recursos auxiliares é “infra-estrutura”.

Entretanto, tanto o termo “infra-estrutura” quanto a variante moderna “cyber infra-estrutura” (NATIONAL Science Foundation, 2003) são de certa forma problemáticos. Infra-estrutura é sempre algum tipo de estrutura, mas quais estruturas devem ser consideradas *infra-estrutura* não é sempre claro e é situacional. Um banco moderno depende de serviços de processamento de dados para fornecer seus serviços bancários e esse suporte a informática é considerado parte da infra-estrutura do banco. Para a indústria de serviços de informática, recursos auxiliares -- a infra-estrutura – inclui serviços bancários confiáveis para lidar com pagamentos. Portanto, serviços bancários são, por sua vez, parte da infra-estrutura da indústria de serviços de informática.

Padrões e protocolos consolidam uma forma intangível de infra-estrutura com conseqüências muito tangíveis e, já que infra-estrutura é considerada o ambiente de suporte que capacita e autoriza, é possível discutir-se quais convenções e mentalidades sociais (as estruturas de pensamento discutidas em *A Ordem das Coisas* de Michel Foucault (1970)) podem ser consideradas uma forma de infra-estrutura (DAY, 2006).

O SEGUNDO USO DE METADADOS

Pensar em metadados como um meio de descrever documentos individuais reflete apenas um dos dois papéis dos metadados. O segundo uso de metadados é diferente: ele ocorre quando você *começa* com os metadados ao invés dos dados, com a descrição ao invés do documento. Isso ocorre quando você pesquisa em um catálogo ou navega em qualquer índice.

Este segundo uso dos metadados é para pesquisa e seleção. Em um ambiente digital, ou seja, até agora, composto principalmente de recursos de texto, é comum e conveniente utilizar consultas textuais e pesquisar a ocorrência de fragmentos de texto nos documentos disponíveis, como em mecanismos de pesquisa na *Web*. Portanto, uma pesquisa para o tema “*Mouse*” é expressa como a seqüência de caracteres “m o u s e” e todos os documentos contendo aquela

seqüência de caracteres serão tidos como “relevantes” sejam eles realmente sobre esse tipo de mamífero (camundongo) ou estejam referindo-se ao dispositivo do computador, ou surja em outros usos figurativos da palavra. A técnica de pesquisa por seqüências de caracteres de texto funciona muito bem, mas nem sempre e não perfeitamente, porque recursos de texto não são inteiramente homogêneos. Algumas palavras possuem vários significados (polissemia, por exemplo, *mouse*); às vezes palavras diferentes utilizam a mesma seqüência de caracteres, mas com outros significados (homógrafos, por exemplo, *pane* significa painel de vidro em inglês, mas não em português); e palavras diferentes podem ser utilizadas com o mesmo significado (sinônimos, por exemplo, *câncer* e *neoplasma*).

Simple pesquisas de texto fragmentam-se em ambientes multilíngües e quando outros tipos de recursos são incluídos, como imagens, sons e conjuntos de dados numéricos. Imagens podem ser comparadas umas às outras e som pode ser comparado a outros sons, mas não, diretamente, a outras formas de mídia. Não é possível utilizar uma consulta composta de alguns pixels ou, um som, como uma consulta em um arquivo de texto.

ÍNDICES COMO ESTRUTURA

Para estabelecer uma conexão significativa entre diferentes documentos, duas ações são necessárias: Primeiro, faça uma conexão entre eles e, então, expresse a natureza da relação entre eles. Por exemplo, pode-se especificar a mesma descrição do tema, como:

- a um texto é atribuído um cabeçalho de assunto e
- a uma imagem é atribuído o mesmo cabeçalho de assunto.

O próximo passo é criar uma forma invertida dessa relação para que seja possível ir do cabeçalho de assunto tanto para o texto quanto para a imagem. Isso permite uma pesquisa unificada dos textos e das imagens relacionados ao mesmo tema. Também permite uma pesquisa transversa a partir de um texto através

de um cabeçalho de assunto para imagens atribuídas ao mesmo cabeçalho de assunto ou, igualmente, a partir de uma imagem para textos:

Texto ----- cabeçalho de assunto ----- imagens

Dessa forma, dois ou mais documentos sobre o mesmo tema, apesar de diversos em seu formato ou conteúdo, podem ser relacionados um ao outro. Esse processo depende de possuir um vocabulário descritivo comum para descrever temas ou, pelo menos, vocabulários interoperáveis.

Essa última manobra inverte a estrutura original. Ao invés de descrições sendo anexadas a documentos, documentos são anexados a descrições. O vocabulário das descrições torna-se central e os documentos tornam-se periféricos. Essa inversão é claramente vista em índices de citação. Quando você examina livros e artigos, as referências são periféricas, em rodapés ou no final, e estão freqüentemente em fonte menor. Mas um índice de citação inverte essa relação. As próprias citações e as relações entre elas tornam-se primárias. Apenas quando uma *citação* de interesse foi selecionada é um *documento*, na periferia, consultado.

ESTRUTURAS SINDÉTICAS

As relações (“estrutura sindética”) entre diferentes cabeçalhos em um vocabulário descritivo (termos preferidos, termos não preferidos, mais amplo, mais limitado e outros termos relacionados) são bem compreendidas na ciência da informação. Mas um problema muito maior tem sido negligenciado. Em uma situação fora da rede (*non-network*) você precisa de referências cruzadas *dentro* de um vocabulário. O propósito de uma rede é permitir acesso a vários recursos diferentes, e em um ambiente de rede você precisa de referências cruzadas *entre* vocabulários. Você precisa saber, por exemplo, que para, digamos, *Automóveis* são classificados como *TL 205* na *Library of Congress Classification* (Biblioteca de Classificação de

Congresso), mas como 180/280 na *US Patent Classification* (Classificação de Patentes dos EUA), como 3711 na *Standard Industrial Classification* (Classificação Industrial Padrão), e como *PASS MOT VEH* na série estatística de importação e exportação dos EUA.

Esse problema não é desconhecido e a terminologia existe: uma ligação entre índices diferentes, porém semelhantes, é chamada de faixa de cruzamento (*crosswalk*) e ligações entre cabeçalhos comparáveis em dois ou mais índices são chamadas de mapeamento (*mapping*). Um exemplo considerável é o *Unified Medical Language Systems* (UMLS, Sistemas de Linguagem Médica Unificada) da *National Library of Medicine* (Biblioteca Nacional de Medicina) (2006), mas, caso contrário, tais mapeamentos são raramente fornecidos e, na verdade, não são discutidos com frequência. Os mapeamentos feitos com habilidade do sistema UMLS são dispendiosos e, certamente, obsoletos, mas em muitas circunstâncias técnicas de associação estatística podem ser utilizadas para gerar mapeamentos úteis através dos vocabulários (BUCKLAND; GEY; LARSON, 2002 e BUCKLAND; CHEN; GEY; LARSON, em breve). A negligência geral do mapeamento por cruzamento de vocabulário sugere que as implicações de um ambiente de rede não foram totalmente compreendidas.

NOMES PRÓPRIOS

Nomes próprios são importantes para autoria e textos biográficos. A necessidade de distinguir entre pessoas diferentes com o mesmo nome e agregar nomes diferentes para a mesma pessoa é bem compreendida em arquivos, bibliotecas, museus e outros locais. Entretanto, as técnicas para lidar com relacionamentos interpessoais parecem ter sido bastante negligenciadas. Genealogistas possuem experiência com a codificação de relacionamentos familiares (pai-filho, cônjuge, etc.), mas as pessoas podem ser relacionadas umas às outras de outras maneiras importantes (ex. professor-aluno, sócio comercial) para os quais as técnicas e a terminologia precisam de maior desenvolvimento.

ÁREAS GEOGRÁFICAS: LOCAL E ESPAÇO

A pesquisa em um ambiente de texto é dominada por palavras-chave tópicas ou palavras-chave não diferenciadas, possivelmente incluindo os nomes de pessoas, lugares e instituições. Entretanto, para pesquisar em algumas fontes, como uma série de dados sócio-econômicos e fotografias, torna-se importante especificar a localização geográfica de forma confiável e exata. “Local” é uma construção cultural e isso é refletido em nomes de lugares, os quais, assim como nomes de temas, são freqüentemente múltiplos (ex. Lisboa, Lisbon, Lisbona, Lisbonne, Lissabon), ambíguos (Galícia, Polônia; Galícia, Espanha) e instáveis (ex. São Petersburgo tornou-se Leningrado e em seguida São Petersburgo novamente).

O espaço, em contraste, é definido em termos físicos de latitude e longitude, os quais fornecem descrições que não são nem ambíguas nem instáveis. Uma grande vantagem das coordenadas espaciais é que elas permitem que os locais sejam mostrados em um mapa. Há, portanto, para áreas geográficas, um sistema de nomenclatura dual de local e espaço: nomes de lugares e coordenadas espaciais. Um dicionário geográfico por nomes de locais pode ser considerado um tipo de dicionário bilíngüe entre lugares e espaços. Um dicionário geográfico possibilita que os nomes dos locais deixem de ser ambíguos e os locais possam ser localizados em um mapa. Um dicionário geográfico bem projetado irá indicar quando um nome de local estava em uso, apoiando assim mapas temporariamente dinâmicos (ZERNEKE; BUCKLAND; CARL, 2006). (Para uma discussão sobre o uso de dicionários geográficos e interfaces de mapa para aprimorar a pesquisa consulte BUCKLAND; GEY; LARSON, 2004 e BUCKLAND; CHEN; GEY; LARSON; MOSTERN; PETRAS, em breve).

EVENTOS E TEMPO

Eventos e tempo tendem a ser mutuamente explicativos. Tempo é ajustado por eventos físicos e períodos culturais por eventos culturais. Mas eventos físicos e períodos culturais são ajustados também pelo tempo no calendário. No discurso e na escrita, comumente marcamos o tempo por referência a eventos, como em “depois que me graduei” ou “antes da Segunda Guerra Mundial”. Essa dualidade de eventos e de tempo lembra a dualidade de local e espaço e convida a uma abordagem semelhante: o uso de um diretório relacionando eventos nomeados ao tempo do calendário. Associar eventos a datas dá suporte à construção de linhas do tempo e cronologias, da mesma forma que um dicionário geográfico relaciona nomes de lugares a coordenadas espaciais e exibições em mapas (PETRAS; LARSON; BUCKLAND, 2006).

BIOGRAFIA

Embora, como já observado, cientistas da informação possuam métodos eficazes para lidar com *nomes* de pessoas, métodos para lidar com *eventos* em suas vidas são muito menos desenvolvidos (TEXT ENCODING Initiative Consortium, 2006). Uma abordagem sendo investigada pela *Electronic Cultural Atlas Initiative* (Iniciativa de Atlas Cultural Eletrônico) é categorizar cada evento biográfico ou atividade de vida como um conjunto de variáveis de quatro facetas com qual tipo de atividade (aspecto tópico), quando (aspecto temporal), onde (aspecto geográfico) e com quem (aspecto biográfico) (BRINGING, 2006). Uma atração dessa abordagem é que eventos de vida podem ser codificados com a terminologia e os métodos já estabelecidos, ou sendo desenvolvidos, para a indexação do assunto, períodos de tempo, nomes de locais e dicionários biográficos.

RELAÇÕES INFRA-ESTRUTURAIS ENTRE TIPOS DE ÍNDICES

Até agora, falamos sobre índices para tema, local, tempo e pessoas como se os índices para essas facetas fossem separados e independentes, mas na prática eles não são, exceto em exemplos primitivos. Em um índice de temas desenvolvido assim como o sistema *Library of Congress Subject Headings* (Biblioteca de Cabeçalhos de Assunto de Congresso), o cabeçalho do tema estará comumente combinado a qualificadores geográficos e cronológicos, ex. *Arquitetura – Japão – período Meiji, 1868-1912*. Em outras palavras, cabeçalhos de assunto podem ter componentes geográficos e temporais assim como tópicos.

Um dicionário geográfico por nome de local comumente indica o tipo de local (“tipo de característica” geográfica) que é: castelo, igreja, cidade, lago, etc. Uma característica física não é o mesmo que um tema, mas qualquer tipo de característica pode ser tratado como um tema. Um determinado castelo é um exemplo da categoria *castelos*. Documentos sobre castelos geralmente podem ser úteis assim como qualquer documento relacionado a esse castelo em particular. E uma discussão sobre o tema *castelos* pode ser enriquecida mudando-se do cabeçalho de assunto para os códigos de tipo de característica geográfica no dicionário geográfico para identificar e localizar *exemplos* de castelos em qualquer região, portanto um mapeamento entre tipos de características e cabeçalhos de assunto pode ser útil. Já que um dicionário geográfico bem desenvolvido também terá uma indicação de *quando* aquele nome estava em uso, entradas em dicionários geográficos, como cabeçalhos de assunto, podem possuir também aspectos temporais e tópicos e aspectos geográficos.

O diretório por período de tempo, o qual modelamos em desenhos do dicionário geográfico, possui uma codificação por tipo de evento ou período. Portanto, como ocorre com entradas do dicionário geográfico, um evento específico (ex. um terremoto) pode ser vinculado a cabeçalhos de assunto tanto por nome próprio (ex. *Terremoto de Lisboa 1755*) quanto pela literatura sobre aquela classe de eventos (ex. *Terremotos*). Eventos são específicos para áreas geográficas e, portanto, um dicionário apropriado por período de tempo terá codificações

geográficas e deveria ser possível vincular cada evento tanto para cabeçalhos de assunto geográfico quanto para entradas de dicionário geográfico.

Os textos de entradas em dicionários biográficos são muito ricos em menções de (i) tipos de atividades, que poderiam ser vinculadas a cabeçalhos de assunto para aquele tipo de atividade; (ii) para locais que pudessem ser vinculados a entradas de dicionários geográficos e para cabeçalhos de assuntos geográficos; (iii) para períodos de tempo que pudessem ser vinculados a outros, eventos contemporâneos via diretório por período de tempo, linhas do tempo e cronologias; e (iv) outras pessoas com as quais o sujeito da biografia interagisse e para quais informações biográficas pudessem ser encontradas em dicionários biográficos e enciclopédias.

Índices de assunto, dicionários geográficos por nome e local, diretórios por período de tempo e dicionários biográficos são gêneros bastante diferentes para aspectos muito diferentes da realidade, mas encontramos conexões geográficas, ligações cronológicas e afinidades tópicas nos quatro. Há uma grande e útil pauta em descobrir maneiras de construir uma infra-estrutura de conexões eficaz entre esses gêneros, pois a compreensão requer um conhecimento do contexto.

CONCLUSÃO E PAUTA

O papel inicial dos metadados é como descrição, mas descrições podem ser manipuladas para fornecer apoio para pesquisa e seleção. Uma pauta central da ciência da informação tem sido a criação de descrições e de índices. Agora, com um ambiente de rede, há uma nova oportunidade de expandir essa pauta para o vínculo não apenas entre índices diferentes do mesmo tipo, mas também entre índices de diferentes tipos. Precisamos desenvolver “práticas melhores” e padrões para vincular entradas em dicionários de sinônimos, dicionários geográficos por nome e local, diretórios por período de tempo, dicionários biográficos e, especialmente, entre esses gêneros diferentes. Precisamos construir uma melhor infra-estrutura de metadados.

AGRADECIMENTOS

Este ensaio inspira-se expressivamente em uma série de estudos da *Electronic Cultural Atlas Initiative* (BUCKLAND; LANCASTER, 2004, 2006) sob a liderança de Michael Buckland, Fredric C. Gey e Ray R. Larson, parcialmente apoiados pelo *U.S. Federal Institute for Museum and Library Services*: “*Seamless Searching of Numeric and Textual Resources*” (NLG 178); “*Going Places in the Catalog: Improved Geographic Access*” (LG-02-02-0035-02) e “*Support for the Learner: What, Where, When, and Who*” (LG-02-04-0041-04).

REFERÊNCIAS

BRINGING lives to light: biography in context. 2006. Available in:
<<http://ecai.org/imls2006>>.

BUCKLAND, Michael; AITAO, Chen; GEY, Fredric C.; LARSON, Ray R. Search across different media: numeric data sets and text files. **Information Technology and Libraries**. Available in:
<<http://www.lita.org/ala/lita/litapublications/ital/italinformation.htm>>.

BUCKLAND, Michael; AITAO, Chen; GEY, Fredric C.; LARSON, Ray R.; MOSTERNS, Ruth; PETRAS, Vivien. Geographic search: catalogs, gazetteers, and maps. **College & Research Libraries**. Available in:
<<http://www.ala.org/ala/acrl/acrlpubs/crljournal/collegeresearch.htm>>.

BUCKLAND, Michael; GEY, Fredric C.; LARSON, Ray R. **Seamless searching of numeric and textual resources**: final report on Institute of Museum and Library Services National Leadership Grant No. 178. Berkeley, CA: University of California, School of Information Management and Systems, 2002. Available in:
<<http://metadata.sims.berkeley.edu/papers/SeamlessSearchFinalReport.pdf>>.
Accessed in: 4 November 2006.

BUCKLAND, Michael; GEY, Fredric C.; LARSON, Ray R. **Going places in the catalog**: improved geographic access - final report. 2004. Available in:
<http://ecai.org/imls2002/imls2002-final_report.pdf>. Accessed in: 4 November 2006.

BUCKLAND, Michael; LANCASTER, Lewis. Combining time, place, and topic: the electronic cultural atlas initiative. **D-Lib Magazine**, v.10, n. 5, May 2004. Available in:

<<http://www.dlib.org/dlib/may04/buckland/05buckland.html>>. Accessed in: 2 November 2006.

BUCKLAND, Michael; LANCASTER, Lewis R. Advances in discovery: the electronic cultural atlas initiative experience. **First Monday**, August 2006. Available in: <http://www.firstmonday.org/issues/issue11_8/buckland/index.html>. Available in: Accessed in: 2 November 2006.

CAPLAN, Priscilla. **Metadata fundamentals for all librarians**. Chicago: American Library Association, 2003.

DAY, Ron. **Notes on infrastructure and development**. 2006. Available in: <<http://ella.slis.indiana.edu/~roday/infrastructure.html>>. Accessed in: 2 November 2006.

FOUCAULT, Michel. **The order of things: an archaeology of the Human Sciences**. New York: Pantheon Books, 1970.

HAYNES, David. **Metadata for information management and retrieval**. London: Facet Publishing, 2004.

NATIONAL Library of Medicine. **Unified medical language system**. Factsheet [Webpage], 2006. Available in: <<http://www.nlm.nih.gov/pubs/factsheets/umls.html>>. Accessed in: 29 October 2006.

NATIONAL Science Foundation. **Revolutionizing science and engineering through cyberinfrastructure**: report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure - "Atkins Report". 2003. Available in: <<http://www.nsf.gov/cise/sci/reports/atkins.pdf>>. Accessed in: 2 November 2006.

PETRAS, Vivien; LARSON, Ray R.; BUCKLAND, Michael. Time period directories: a metadata infrastructure for placing events in temporal and geographic context. In: **Opening information horizons**: Joint Conference on Digital Libraries (JCDL), Chapel Hill, NC, June 11-15, 2006. Available in: <<http://metadata.sims.berkeley.edu/tpdJCDL06.pdf>>. Accessed in: 4 November 2006.

TEXT ENCODING Initiative Consortium. 2006. Report on XML Mark-up of Biographical and Prosopographical Data. Available in: <<http://www.tei-c.org/Activities/PERS/persw02.xml>>. Accessed in: 2 November 2006.

ZERNEKE, Jeannette L.; BUCKLAND, Michael K.; CARL, Kim. Temporally dynamic maps: the electronic cultural atlas initiative experience. **Human IT**, v.8, n.3, p.83-94, 2006. Available in: <<http://www.hb.se/bhs/ith/3-8/jzmbkc.pdf>>. Accessed in: 4 November 2006.

Michael K. Buckland

School of Information
University of California
Berkeley, USA 94720-4600

Artigo recebido em: 1, 11, 2006

Artigo aceito em: 1, 11, 2006