

XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT-2 – Organização e Representação do Conhecimento

A QUESTÃO DA INTEROPERABILIDADE EM REPOSITÓRIOS INSTITUCIONAIS E SISTEMAS DE INFORMAÇÃO DE PESQUISAS CORRENTES (CRIS): UMA ABORDAGEM PRELIMINAR

Nilson Theobald Barbosa - Universidade Federal Fluminense (UFF)

Maria Luiza de Almeida Campos - Universidade Federal Fluminense (UFF)

THE ISSUE OF INTEROPERABILITY ON INSTITUTIONAL REPOSITORIES AND CURRENT RESEARCH INFORMATION SYSTEMS (CRIS): A PRELIMINARY APPROACH

Modalidade da Apresentação: Pôster

Resumo: O acesso aberto às publicações, métodos e resultados das pesquisas científicas depende da garantia do efetivo compartilhamento de dados entre os pesquisadores e a sociedade. Nesse sentido surgem os repositórios institucionais e a proposta de criação dos *Current Research Information Systems*. Para que o efetivo compartilhamento e abertura das informações ocorram, se coloca na ordem do dia a questão da interoperabilidade entre repositórios e bancos de dados de pesquisa. O presente trabalho procura mostrar algumas iniciativas neste campo, apresentando também as bases de funcionamento dos *Current Research Information Systems* e do *Common European Research Information Format*. Mostra-se a seguir alguns caminhos que podem ser trilhados pela Ciência da Informação como contribuição à solução do problema da interoperabilidade, tal como a aplicação das bases formuladas por Dahlberg para a Teoria do Conceito. Ao se contrapor dois tipos possíveis de interoperabilidade, seja a interoperabilidade baseada em nomes e formas verbais e a interoperabilidade baseada em conceitos, mostra-se que a interoperabilidade semântica se torna possível a partir do alinhamento no significado dos conceitos. O trabalho conclui afirmando o ainda incipiente estágio da busca pela interoperabilidade, em que pesem os esforços já de décadas neste sentido. Mostra, ainda, o importante papel a ser desempenhado pela Ciência da Informação neste campo.

Palavras-Chave: Interoperabilidade; Repositórios; *Current Research Information Systems*; Teoria do Conceito.

Abstract: Open access to publications, methods and results of scientific research depends on ensuring the effective sharing of data between researchers and society. In this sense, the institutional repositories and the proposed creation of *Current Research Information Systems* appear. In order for the effective sharing and opening of information to occur, the question of interoperability between repositories and research databases is put on the agenda. The present work tries to show some initiatives in this field, also presenting the operating bases of *Current Research Information Systems* and the *Common European Research Information Format*. Following are some paths that can be traced by Information Science as a contribution to the solution of the problem of interoperability, such as the

application of the bases formulated by Dahlberg for the Theory of Concept. By contrasting two possible types of interoperability, be it verbal-based name and form interoperability and concept-based interoperability, it is shown that semantic interoperability becomes possible from alignment on the meaning of concepts. The paper concludes by affirming the still incipient stage of the search for interoperability, which weighs the efforts already of decades in this sense. It also shows the important role to be played by Information Science in this field.

Keywords: Interoperability; Repositories; Current Research Information Systems; Concept Theory.

1 INTRODUÇÃO

Podemos perceber um crescente interesse em escala mundial no acesso aberto e na ciência aberta em todos os níveis, desde as publicações científicas, até os dados científicos de pesquisa, passando por todas as etapas e processos envolvidos com a pesquisa científica, tais como financiamentos públicos e privados, recursos utilizados, prestação de contas para a sociedade, recursos humanos utilizados, e assemelhados. Este é um cenário em que se vislumbra um avanço em todos os aspectos do processo de pesquisa científica: compartilhamento e reuso de informações entre os pesquisadores, divulgação científica dos métodos e recursos utilizados, acompanhamento e transparência por parte das agências financiadoras e da sociedade.

Mas, por outro lado, o aumento exponencial da quantidade de informações cria uma grande dificuldade para encontrar informações relevantes, dificultado, ainda, pela enorme diversidade de áreas do conhecimento e instituições, uma grande diversidade de formatos de dados e diversidade de metadados utilizados.

É nesse ambiente que os repositórios institucionais de dados de pesquisa têm a oportunidade de cumprirem um papel mais efetivo no processo global das pesquisas, surgindo como ferramentas e soluções indispensáveis para a implantação e efetivação do acesso aberto e do efetivo compartilhamento de dados entre os pesquisadores, para os gestores universitários e governamentais e para a sociedade em geral.

Dentro desta discussão sobre repositórios surge também a proposta do ambiente CRIS (*Current Research Information System*), que tem a missão de gerenciar todas as informações de pesquisa relevantes para uma determinada instituição ou grupo de instituições ao tentar promover cooperação e compartilhamento de conhecimento entre a comunidade científica e implantar interoperabilidade de informações de pesquisas. As áreas de interesse deste ambiente cobrem bancos de dados de pesquisa, *datasets* de dados científicos de pesquisa,

repositórios institucionais de acesso aberto, assim como regulam o próprio acesso aos dados e mecanismos de intercâmbio (EUROCRIS, 2017).

Mas estes sistemas, em que pese seu avanço qualitativo e quantitativo, não resolvem a grande tarefa de fornecer aos pesquisadores informações úteis, completas, ligadas e com pouco ruído. É muito comum que as informações requeridas não estejam armazenadas em apenas um ambiente ou instituição, normalmente podem estar espalhadas entre sistemas diferentes bastante heterogêneos, sendo necessário um grande esforço e tempo para compreender e juntar as informações relevantes, levando a que estes pesquisadores não se sintam propensos e estimulados a publicar os resultados completos de sua pesquisa em sistemas de informações de pesquisas, por não terem como vislumbrar a potencialidade do sistema.

Portanto, uma tarefa que temos hoje é criação desta rede de informações que consiga, apesar da diversidade e heterogeneidade, dar sentido e ligar informações espalhadas por diversas instituições e diversas áreas do conhecimento.

É neste ambiente que se coloca a questão da interoperabilidade, como uma condição para que sistemas de informação heterogêneos possam interligar informações semanticamente semelhantes.

Desta forma, este trabalho pretende trazer para uma discussão ainda preliminar as bases de funcionamento e as questões que envolvem a interoperabilidade em ambientes heterogêneos e ambientes CRIS, usando como metodologia uma pesquisa qualitativa de natureza exploratória.

2 INICIATIVAS EXISTENTES

Temos hoje algumas iniciativas importantes para lidar com a heterogeneidade e diversidade dos dados relativos à pesquisa científica e sua utilização em repositórios. Uma dessas iniciativas é a desenvolvida pelo *Open Archives Initiative*, que define sua missão como “desenvolver e promover padrões de interoperabilidade que objetivem facilitar a disseminação eficiente de conteúdo, tendo suas raízes no acesso aberto e nos movimentos pelos repositórios institucionais” (OAI, 2017).

Uma das soluções propostas pela OAI é o ‘Protocolo para Colheita de Metadados’¹ cuja função é criar mecanismos para extração de registros contendo metadados de repositórios,

¹ Protocol for metadata harvesting (OAI-PMH). Especificações em <https://www.openarchives.org/pmh>.

dando assim uma opção técnica simples para que os provedores de dados coloquem seus metadados disponíveis para outros serviços. Este protocolo se baseia nos padrões abertos implementados pelos HTTP e pelo XML.

No ambiente dos Current Research Information Systems (CRIS) a proposta apresentada pelo euroCRIS é o ‘Formato Comum Europeu para Informações de Pesquisas’². Segundo suas especificações, o CERIF é um conceito, uma descrição e uma formalização, respectivamente nos níveis conceitual, lógico e físico, de entidades de pesquisa e seus relacionamentos, propondo um padrão internacional de modelo de dados relacionais para armazenamento e interoperabilidade de informações de pesquisa, sendo uma recomendação da UE a todos seus estados membros.

O princípio básico do CERIF sustenta que a maioria das características (atributos) de um objeto (entidade) não é armazenada junto com a entidade, mas expressada através das entidades de ligação, permitindo que múltiplos papéis/características possam ser expressadas para o mesmo aspecto ou objeto. As entidades de ligação são usadas para expressar o tipo e o significado de uma relação entre dois objetos no campo da pesquisa científica, classificar objetos segundo um esquema de classificação, por exemplo, um vocabulário controlado, e ainda para mapear diferentes esquemas de classificação um para o outro, por exemplo, mapear uma lista de palavras chave da bioquímica usada na Medicina para outra lista usada por biólogos (EUROCRIS, 2017). A Figura 1 mostra um caso real exemplificando a utilização das regras do CERIF.



Fonte: <http://www.eurocris.org/cerif/main-features-cerif>.

² Common European Research Information Format (CERIF). Especificações em <http://www.eurocris.org/cerif/main-features-cerif>.

Diversos outros esforços, atuais e no passado, já foram feitos no sentido de elaborar metadados e ontologias que atendessem ao ambiente universitário e de pesquisas, apenas a título de exemplo, podemos citar o SHOE (*Simple HTML Ontology Extensions*) que propôs a *University Ontology*³, que define elementos para descrever universidades e as atividades que ocorrem nelas, tais como atividades de pesquisa e publicações científicas.

3 INTEROPERABILIDADE E A CIÊNCIA DA INFORMAÇÃO

Para Payette et al. (1999), a partir de trabalho apresentado na Universidade de Cornell, a interoperabilidade é definida como sendo a habilidade dos componentes ou serviços serem funcionalmente e logicamente intercambiáveis pelo fato de terem sido implementados em acordo com um conjunto de interfaces bem definidas e publicamente conhecidas.

Mas o acesso a sistemas heterogêneos pode envolver documentos (ou *datasets*) de diferentes grupos de pesquisa e instituições e até diferentes países que tem suas informações indexadas de forma bastante diferentes, seja linguisticamente, estruturalmente ou em variados esquemas conceituais. Em um ambiente diversificado em alguns ou todos destes aspectos, pode não ser possível a utilização de mesmas interfaces e nomenclaturas e os processos de recuperação podem levar a resultados insatisfatórios. Nestes casos os usuários podem ser levados a ter que realizar diversas consultas para obter resultados relevantes ou, às vezes, sequer conseguir obter algum resultado. Para o estabelecimento de interoperabilidade entre diferentes sistemas ou linguagens de indexação, alguns métodos, em diferentes níveis, podem ser adotados.

O aumento considerável das tecnologias que promovem interfaces abertas e a criação de padrões de metadados abrangentes, que sejam capazes de representar os ambientes de pesquisa acadêmica são, sem dúvida, grandes passos no sentido de avançar o processo de interoperabilidade. Dentro do ambiente da *web* semântica o estudo e a criação de ontologias tem aspecto relevante na solução do problema. Conforme Campos (2007),

Ontologia é um conjunto de conceitos padronizados onde termos e definições devem ser aceitos por uma comunidade no âmbito de um domínio, e tem por finalidade permitir que múltiplos agentes compartilhem conhecimento. Uma ontologia consiste em termos, definições e axiomas relativos a eles. As ontologias constituem um meio poderoso de inter-relacionar sistemas. São elaboradas, principalmente, visando à estruturação de bases de conhecimento

³ <http://www.cs.umd.edu/projects/plus/SHOE/onts/univ1.0.html>. Acesso em 15 jul. 2017.

ou para serem utilizadas como ferramentas semânticas no suporte à interoperabilidade entre sistemas de informação (CAMPOS, 2007).

Nesse sentido, é de grande importância as instituições em qualquer âmbito criarem as suas ontologias ou vocabulários controlados e especializados, mas a enorme heterogeneidade semântica existente não é capaz de permitir por si só a interoperabilidade semântica, uma vez que não seria nunca viável a criação de uma ontologia global onde todos simplesmente aderissem. Portanto, no mundo real é necessário criar mecanismos que permitam esta interoperabilidade semântica entre sistemas heterogêneos.

A interoperação entre dois sistemas parece somente ser possível quando podemos superar as diferenças linguísticas, terminológicas e verbais que são definidas pelas condições idiossincráticas de cada ambiente, para podermos comparar e ligar as definições e conceitos associados a cada objeto.

Dahlberg (1976), ao tratar da questão do conceito, mostra a sua importância para a representação e as questões atuais de interoperabilidade. Ao definir que cada conceito tem um referente e que podem ser feitas afirmações verificáveis sobre este conceito, a partir daí poderão ser sumarizadas por um termo que represente este conceito nos diversos sistemas heterogêneos de que ele participe.

A proposta do triângulo conceitual (Figura 2) definido por Dahlberg (1978) explicita seu modelo para construção de conceitos.

Figura 2: Triângulo conceitual de Dahlberg.



Fonte: Dahlberg (1978).

Onde o vértice A aponta o REFERENTE, o vértice B aponta as CARACTERÍSTICAS e o vértice C aponta a FORMA VERBAL associada.

XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017
23 a 27 de outubro de 2017 – Marília – SP

O conceito, então, é definido como “sínteses rotuladas de afirmações verdadeiras feitas a respeito de objetos do pensamento, com as afirmações conduzindo à identificação ou à separação dos conceitos que também podem ser vistas como elementos dos conceitos” (DAHLBERG, 1976).

Para que, então, diferentes ontologias possam ser consideradas compatíveis devemos trabalhar no plano de suas definições, pois apenas estas descrevem o conteúdo semântico do signo linguístico, conforme Campos (2007).

Temos ainda, que

[...] esta descrição possibilita que agentes inteligentes possam entender o significado de um termo e estabelecer inferências sobre esses significados, pois a definição é composta de características de conceitos, que são também conceitos que se relacionam formando o entendimento semântico dos termos em questão (CAMPOS, 2007).

Uma vez que um conceito, na visão da Teoria do Conceito de Dahlberg, é composto pelas suas três partes distintas, a identificação completa de suas características e sua compreensão permite explicitar suas relações dentro de um sistema e determinar o grau de relacionamento entre eles, sem que as diferenciações ocasionados pelas diferentes formas verbais adotadas impeçam ou induzam a erro nestas relacionamentos. Ou seja, os referentes devem ser identificados pelos elementos que os definem.

Podemos identificar, então, dois tipos de interoperabilidade que podem ser utilizados, cada um com características específicas com relação às estratégias utilizadas e com as implicações que tem com relação a características específicas de recuperação, a interoperabilidade baseada em palavras ou a interoperabilidade baseada em conceitos (HUBRICH, 2010).

O primeiro tipo, interoperabilidade baseada em palavras, se pauta basicamente pela comparação de cadeias de caracteres. A similaridade de cadeias de caracteres pode ser analisada e determinada por um tipo de alinhamento léxico, responsável por permitir afirmar que duas cadeias de caracteres se equivalem. O uso de dicionários e vocabulários controlados pode aumentar a efetividade deste ambiente. Mas, devido à sua alta imprecisão, este tipo de relação entre sistemas baseada em palavras não podem atender adequadamente aqueles que procuram explorar conhecimento entre diferentes bases de dados. Já no caso da interoperabilidade baseada em conceitos, a interoperabilidade semântica é definida e obtida pela identidade ou por um forte alinhamento no significado dos conceitos. Desta forma,

diferentemente da estratégia baseada em palavras e cadeias de caracteres podemos obter uma alta precisão nas relações intersistemas predefinidas. É importante lembrar que neste modelo temos uma alta demanda na experiência e conhecimento das pessoas responsáveis na criação dos vínculos.

4 CONSIDERAÇÕES FINAIS

Paepcke et al (1998) mostram que a interoperabilidade vem sendo um problema crítico desde os anos 1990 e continuará a ser num futuro previsível, uma vez que o número de sistemas computacionais, repositórios de informação, aplicações e usuários se multiplicam em uma taxa explosiva. Pelo fato da interoperabilidade ser, por natureza, um problema extremamente complexo e em evolução, e pesquisadores tenham trabalhado neste sentido desde os anos 1970, muitas vezes não está claro que princípios foram estabelecidos ou quais resultados chave foram obtidos.

Em que pese o grande avanço computacional, a grande diversidade de novas tecnologias disponíveis e os esforços para a criação de repositórios interligados semanticamente, notadamente e citados aqui, dos ambientes CRIS e euroCRIS, podemos ver que a questão da interoperabilidade é um problema ainda em aberto cuja solução passa por uma atuação conjunta da ciência da informação, onde podemos citar preliminarmente a teoria do conceito, teoria da terminologia, a teoria da classificação e a compatibilização de linguagens, e da ciência da computação.

REFERÊNCIAS

CAMPOS, M.L.A., Integração de Ontologias: o domínio da Bioinformática. **RECIIS – R. Eletr. de Com. Inf. Inov. Saúde**. Rio de Janeiro, v. 1, n. 1, p. 117-121, jan.-jun., 2007.

DAHLBERG, I. Classification theory, yesterday and today. **International Classification**, v.3. 1976.

DAHLBERG, I. **Ontical Structures and universal classification**. 1978.

EUROCRIS. Current Research Information Systems. **The International Organisation for Research Information**. Página web. Disponível em: <<http://www.eurocris.org/>>. 2017. Acesso em 15 jul. 2017.

HUBRICH, J. Intersystem relations: characteristics and functionalities. **Concepts in Context**. Proceedings of the Cologne Conference on interoperability and semantics in knowledge organizations. p. 69-90. 2010.

XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017
23 a 27 de outubro de 2017 – Marília – SP

OAI. Open Archives Initiative. **Standards for Web Content Interoperability**. Página web. Disponível em: <<https://www.openarchives.org/>>. 2017. Acesso em 15 jul. 2017.

PAEPCKE, A. et al., Interoperability for Digital Libraries Worldwide, **Communications of the ACM**, Volume 41, 4, April 1998.

PAYETTE, S. et al. Interoperability for digital objects and repositories: the Cornell/CNRI Experiments. **D-Lib Magazine**, v.5, n.5, 1999. Disponível em: <<http://dlib.org/dlib/may99/payette/05payette.html>>. Acesso em: 1 jul 2017.