22. Repositório Semântico de Dados

Autores:

Glauber Mauch de Carvalho

Renato Semmler

Mário Olímpio de Menezes

Instituto de Pesquisas Energéticas e Nucleares, IPEN – CNEN/SP

Repositório Semântico de Dados de Análise por Ativação Neutrônica – uma proposta para eScience

Glauber Mauch de Carvalho, Renato Semmler, Mário Olímpio de Menezes

Instituto de Pesquisas Energéticas e Nucleares, IPEN – CNEN/SP

Abstract

Scientific Data Repositories are being made available each time more often, in a search for more transparency for scientific research and its results, making, in this way, possible to validate, reproduce and reuse the data in other studies. In this work, we present the ongoing research, part of a master dissertation, being developed at IPEN-CNEN/SP, seeking to build a semantic repository for Neutron Activation Analysis research data. The study began with the ontology construction and some of the preliminar results of this phase are presented.

Resumo

Repositórios de Dados Científicos estão sendo disponibilizados cada vez com mais frequência, buscando dar maior transparência às pesquisas e resultados obtidos, que podem, assim ser validados, reproduzidos e terem seus dados utilizados em outros trabalhos. Neste artigo apresentamos o estudo em andamento que faz parte de uma dissertação de mestrado que está sendo realizado no IPEN-CNEN/SP para a construção de um repositório semântico de dados de pesquisas científicas da área de Análise por Ativação Neutrônica. Iniciamos os estudos pela construção da Ontologia, e apresentamos algumas das atividades relacionadas a esta construção.

1.1. Introdução

Um dos pilares da ciência é a possibilidade de reprodução dos resultados de pesquisas científicas por pesquisadores independentes, tornando possível a validação dos métodos, dos resultados e suas conclusões.

A disponibilização dos dados subjacentes aos artigos científicos publicados, tem sido colocada como exigência por algumas revistas inovadoras no meio científico [4]. Entretanto esta disponibilização precisa vir acompanhada de conhecimento sobre a origem dos dados; igualmente importante é saber como tais dados foram produzidos, de modo que é necessário associar aos dados, os métodos, algoritmos ou demais técnicas empregadas em sua obtenção bem como nas análises que culminaram na publicação à qual estão associados.

O entendimento da semântica dos dados publicados é, portanto, um dos pontos fundamentais para garantir a reusabilidade destes, e esta necessidade se torna ainda mais

urgente nesta era do *Big Data*, sendo uma condição necessária para a reutilização eficiente dos dados publicados.

Para fazer frente a este novo cenário, um conjunto complexo e amplo de conhecimentos, sistemas, métodos e tecnologias estão envolvidos. Modelos matemáticos, repositórios digitais e gerenciamento de dados, novos hardwares, softwares, protocolos, ferramentas e serviços, são alguns dos itens necessários para se atender as demandas deste novo paradigma da ciência [3].

Recentemente, vários repositórios de dados de propósito geral têm surgido, alguns de escala institucional outros de escala global [5]. Um problema, entretanto, tem afetado estes repositórios: a não utilização de padrões de formato de dados ou para a descrição destes dados. Como resultado, presencia-se uma mudança no ecossistema de dados, que passou de um sistema centralizado para um sistema diversificado, não integrado, que aumenta o problema de descoberta e reutilização de dados tanto para humanos como para sistemas computacionais.

Por conta destes problemas existentes nos repositórios tradicionais, as tecnologias semânticas estão ganhando cada vez mais espaço em áreas da e-Science tais como física solar-terrestre, ecologia, ciências marítimas e dos oceanos, saúde e ciências da vida, entre tantas outras. Cada vez mais são utilizados metodologias, ferramentas e outros componentes baseados em semântica (ontologias); a modelagem de conhecimento, a verificação de hipóteses baseada em lógica, a integração de dados semânticos, composição de aplicações e descoberta de conhecimento e análises de dados integrados para diferentes domínios de conhecimentos são atividades cada vez mais presentes nestes contextos [1].

Apresentamos neste artigo o trabalho em andamento da construção de um repositório semântico para os dados de Análise por Ativação Neutrônica (AAN), obtidos no Centro do Reator de Pesquisas (CRPq), do IPEN-CNEN/SP. A construção de tal repositório visa permitir que se possa organizar, classificar, selecionar, compartilhar e disponibilizar o acesso ao crescente volume de dados gerados pelas pesquisas do CRPq, afim de proporcionar que os resultados destas pesquisas possam ser reutilizados, reproduzidos, validados por pesquisadores independentes, tornando possível não só a validação dos métodos, resultados e conclusões, mas também o avanço de áreas diversas que podem se beneficiar de tais dados.

1.2. Análise Por Ativação Com Neutrôns (AAN)

O princípo da análise por ativação com nêutrons é a interação de um dado material com nêutrons, que induz uma reação nuclear em um átomo de um elemento alvo. O produto da reação é detectado e quantificado por um fóton pronto ou emissão de uma partícula ou, mais comumente, por suas propriedades de decaimento. Para a AAN, várias reações nucleares são possíveis dependendo do núcleo alvo e da energia do nêutron.

No Laboratório de Ativação Neutrônica do CRPq, o método AAN empregado corriqueiramente é o método comparativo, no qual a amostra desconhecida tem sua massa determinada e é irradiada simultaneamente com um padrão (material de concentração conhecida do elemento a ser determinado), e após um determinado período de tempo, a intensidade e energia de picos de raios gamas da amostra são medidos. A comparação

entre as atividades específicas induzidas nos padrões e amostras desconhecidas é a base para o cálculo da concentração do elemento na amostra.

Durante um experimento de AAN muitos dados são gerados, como mostrado na figura 1.1.

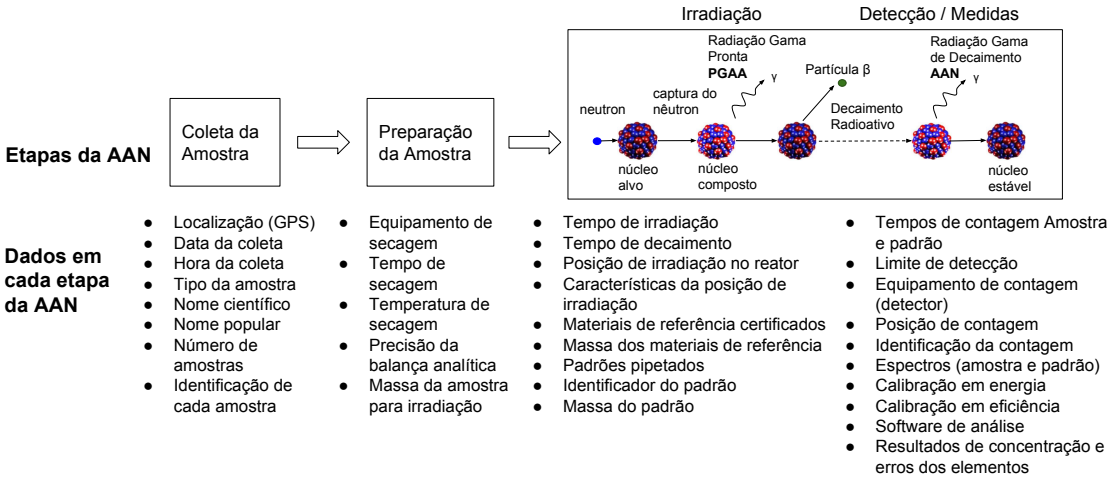

Figure 1.1. Dados gerados do processo de AAN

Este grande volume de dados gerado pelos laboratórios de Análise por Ativação com Nêutrons (AAN), resultou na iniciativa de criar um repositório semântico de dados oriundos da AAN.

Este repositório de dados, munido de capacidade semântica, integra-se a uma proposta maior, que é a infraestrutura de e-Science do IPEN, que tem, dentre outros, os seguintes componentes:

i. Repositório semântico de dados;

ii. Softwares de análise de dados;

iii. Mecanismos de Governança de Dados, incluindo anotação, curadoria e proveniência para os dados de pesquisa, de modo a garantir sua confiabilidade; e

iv. Mecanismos de Segurança da Informação.

Ontologia

Inúmeras definições a cerca do termo ontologia têm surgido; afim de ilustrar o conceito empregado neste trabalho, podemos utilizar a definição dada por Gruber [2], para quem uma área de conhecimento formalmente representada baseia-se numa conceitualização, isto é, os objetos, conceitos e outras entidades que se supõe existir em alguma área de interesse e as relações que mantêm entre eles.

Neste sentido, a conceitualização é uma visão abstrata e simplificada do mundo que se deseja representar de forma explícita ou implícita. Ela é independente do tipo

Table 1.1. Alguns dos termos da Ontologia para os dados de AAN

Termo	Descrição
Tempo de irradiação	Período de tempo que a amostra fica sendo irradiada no reator
Posição de irradiação	Local onde a amostra é irradiada no reator
Característica da posição de irradiação	Esta relacionado com a intensidade do fluxo de nêutrons térmicos e rápidos.
Limite de detecção	Atividade mínima necessária para se determinar a concentração de elemento.
Posição de contagem	Local onde a amostra é posicionada no sistema de aquisição.

de linguagem, o que facilita a interpretação por parte dos especialistas sobre o modelo proposto. Assim temos a seguinte definição:

> "Ontologia é a especificação formal explicita de uma conceitualização compartilhada"

O termo “explícita” diz respeito aos elementos e suas restrições, enquanto “conceitualização” trata de um modelo abstrato de uma área de conhecimento; o termo “compartilhada” implica um conhecimento consensual. Esta definição permite concluir que podemos fazer uso das ontologias para manipular, compartilhar e processar informações de modo preciso.

No processo de construção da ontologia, os principais termos e conceitos da área de Análise por Ativação Neutrônica são compilados e avaliados, tendo os seus significados escrutinados por especialistas, visando garantir a precisão de sua definição. A tabela 1.1 apresenta alguns destes termos da ontologia em construção.

1.3. Metodologia

A metodologia adotada para a realização deste estudo, que é a construção de um repositório semântico de dados de pesquisa de AAN, foi divida em duas etapas:

1. Construção da Ontologia de descrição dos dados de AAN;
2. Construção do protótipo do repositório semântico.

A etapa 1, Construção da Ontologia está em desenvolvimento, para a qual já foram realizadas algumas atividades, tais como: levantamento bibliográfico de variadas

dissertações e teses produzidas pelos discentes do Laboratório de Análise por Ativação Neutrônica, a fim de realizar o primeiro levantamento dos termos e conceitos de AAN. Em seguida, foi construída uma tabela contendo a identificação, contextualização e descrição dos termos com base na literatura para o presente estudo, conforme mostrado na Tabela 1.1; a mesma foi averiguada por alguns pesquisadores da área de Análise por Ativação Neutrônica, de modo que os termos relevantes tivessem sua descrição refinada e validada.

Na etapa 2, de Construção do Repositório, estão sendo desenvolvidas as seguintes atividades: o estudo da arquitetura do sistema de armazenamento (repositório) e a integração da camada semântica, ambas em fase inicial, não apresentando, desta forma, resultados expressivos até o momento. É importante ressaltar que as etapas mencionadas acima não são excludentes, dessa forma estão sendo executadas simultaneamente.

1.4. Discussões

A construção de um repositório semântico de dados se faz cada vez mais premente diante das grandes quantidades de dados, bem como a necessidade de utilização, de maneira mais eficaz, de tais dados em outras áreas de pesquisas.

A camada semântica demanda a construção de uma Ontologia para a descrição destes dados, o que está sendo feito de modo criterioso neste trabalho; espera-se que nos próximos meses a Ontologia, seja finalizada, bem como a arquitetura do sistema de armazenamento (repositório semântico); finalizadas estas atividades, espera-se dar início à construção do sistema, empregando-se as metodologias ágeis de construção de software.

A criação deste repositório semântico evidencia o esforço do IPEN em disponibilizar um repositório de dados que permita a utilização completa dos dados de pesquisas, primeiramente os dados de AAN e, porteriormente, as outras áreas de pesquisa do IPEN.

Agradecimento O Projeto e-Science do IPEN recebeu apoio financeiro da FAPESP, através de reserva técnica institucional, pelo projeto 2015/14689-6.

Referências

[1] Peter Fox and James Hendler. *The Fourth Paradigm – Data-Intensive Scientific Discovery*, chapter Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science, pages xvii–xxxi. Microsoft Research, Redmond, Washington, 2009.

[2] Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[3] Tony Hey, Steward Tansley, and Kristin Tolle. The Fourth Paradigm. *Data-Intensive Scientific Discovery. Microsoft Research*, 2009.

[4] Fabrício Marques. Ciência transparente. *Revista Fapesp*, (218):54–58, 2014.

[5] Mark D. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, mar 2016.