gusttavodiias@gmail.com
lgdias@id.uff.br

Instituto de Computação

PAPER REVIEW: A POSTERIORI METADATA FROM AUTOMATED PROVENANCE TRACKING: INTEGRATION OF AIIDA AND TCOD

JOURNAL	AUTORES	LINK - SCOPUS	COMPILADO POR
Journal of Cheminformatics	Andrius Merkys, Nicolas Mounet, Andrea Cepellotti, Nicola Marzari, Saulius Grazulis, Giovanni Pizzi	https://goo.gl/D2EdxZ	Luiz Gustavo Dias - UFF

RESUMO

Para tornar os resultados de pesquisas acessíveis, e reusáveis é necessário agregar metadados a eles. Entretanto existem inúmeros desafios técnicos e práticos, que tornam esse processo difícil. É apesentado no trabalho um protocolo para etiquetar estruturas e suas propriedades, sem a necessidade de intervenção humana para manipular os dados. Para isso o protocolo utiliza recursos do AiiDA – uma plataforma para gerenciar e automatizar *workflows* científicos, e TCOD – um banco de dados que armazena propriedades de materiais computados por meio de uma ontologia, com isso o procedimento deposita nos dados na base de forma automática, juntamente com os metadados relevantes extraídos de AiiDA. O protocolo permite também a reprodutibilidade de dados no campo da ciência dos materiais. A interface AiiDA – TCOD é utilizada para depositar 170 estruturas juntamente com suas propriedades computadas e gráficos de provisão.

VISÃO GERAL

A modelagem e simulação são indispensáveis na compreensão científica, em particular na ciência dos materiais, devido aos avanços significativos na aproximação das teorias usadas para simular materiais (e seus códigos), e segundo pela viabilidade dos cálculos (anteriormente muito caros). Como consequência um grande numero de propriedades podem ser calculadas para grandes famílias de materiais.

Várias bases de dados emergem dessa área. Entretanto é necessário esforço para consolidar resultados com metadados sob uma ontologia, que preserve a proveniência completa dos dados computados, permitindo a reprodutibilidade dos resultados. Atualmente existem varias tentativas de definir uma ontologia teórica no campo da ciência dos materiais, como European Theoretical Spectroscopy Facility (ETSF), NOMAD, OPTiMaDe e Theoretical Crystallography Open Database (TCOD), onde essa ultima tem o objetivo de coletar resultados de vários tipos de cálculos relacionados a estruturas cristalinas e armazena-los.

Outra questão importante e a preservação da procedência para replicação dos resultados. Atualmente a maioria das publicações cientificas fornece apenas um subconjunto de todos os parâmetros de controle, entradas e interrelações de cálculos necessários para reproduzir os resultados. Este problema pode ser resolvido usando estruturas que rastreiam proveniência, como AiiDA, que e uma infraestrutura de alto rendimento que fornece o ambiente de pesquisa de alto nível para automatizar a execução de cálculos, armazenam sistematicamente as entradas e saídas bem como seus relacionamentos em um banco de dados gráfico, adaptado para acompanhar a proveniência dos dados completos e compartilhar resultados. A pesquisa consta em integrar o banco de dados TCOD com AiiDA usando um dicionário especifico. Tal integração permite o armazenamento e reprodutibilidade dos resultados, pois salva os resultados das simulações juntamente com seus metadados, de forma automática.

WORKFLOW REPRESENTATION

A reprodutibilidade de cálculos científicos e um principio fundamental da pesquisa cientifica, um sistema que permita sua replicação é de extrema importância para atingir esse objetivo. Um pré-requisito importante é que os dados sejam separados da representação. No campo da ciência dos materiais existe AiiDA, que é uma estrutura baseada em Python para simulações, onde a proveniência e armazenada automaticamente enquanto as simulações são executadas. O objetivo do trabalho e representar a proveniência total do *workflow* representado por AiiDA em forma de um grafo acíclico direcionado (DAG), em um arquivo CIF (um formato unificado para relatar e armazenar os resultados de experimentos relacionados a estruturas cristalinas, que possui como principal vantagem, a existência de dicionários CIF destinado a definir ontologias especificas de domínio). Para atingir esse objetivo, workflow é representado por uma lista ordenada de processos para execução, onde tal sequencia leva a geração de resultados.

INPUT DATA

Etapas iniciais de um *workflow* geralmente transformam dados de entrada – geralmente de fontes externas ou bases de dado – para as estruturas internas. Para preserva-los, é de extrema importância fazer referencia aos dados originais. Esse processo e relativamente simples principalmente se o recurso está disponível na internet, atribuindo URI por exemplo. Uma das partes da pesquisa foi implementar importadores de bases de dados externos como parte do AiiDA, possibilitando que os usuários buscassem e importassem determinados dados diretamente para AiiDA, onde era certificado que a fonte estava sendo gravada. Quando o grafo AiiDA era posteriormente exportado para o formato CIF e depositado no TCOD, a fonte dos dados era registrada, bem como os itens de dados.

INCLUSÃO DE CONTEUDO NO ARQUIVO CIF E CODIFICAÇÃO

Foi definido que todo o arquivo do *workflow* seria armazenado nos CIFs do TCOD, para obter as propriedades de um determinado material. No entanto devido a restrições do formato CIF, não era possível armazenar todos os dados sem que os mesmos fossem modificados em um arquivo CIF. Desta forma foi implementado um protocolo para conversão de dados compatíveis com o arquivo CIF.

ONTOLOGIAS

O dicionário CIF desenvolvido pelo conselho consultivo do TCOD, fornece itens para descrição de conjuntos de bases. Para acomodar parâmetros de entrada e resultados exportados de AiiDA, o dicionário foi complementado com outros itens.

IMPLEMENTAÇÃO: EXPORTANDO OS DADOS PARA O TCOD

O principal resultado do trabalho foi a definição e implementação de procedimentos para exportar os resultados De cálculos teóricos gerenciados com o AiiDA em CIF e depositá-los no banco de dados do TCOD. Para conseguir isso, foi implementado um conversor que, a partir de uma estrutura especificada pelo usuário dentro do AiiDA, é capaz de criar um arquivo de formato CIF. O conversor permitiu a marcação automática a posteriori completa de estruturas com seus metadados. Isso foi possível analisando a proveniência completa (armazenada no AiiDA DAG) da estrutura cristalina final, extraindo /converter todas as informações relevantes e armazená-las no campos CIF apropriados definidos nos dicionários do TCOD. Os passos para isso foram definidos como:

1. Conversão da estrutura periódica do AiiDA: Existem dois tipos de estruturas periódicas no AiiDA: estrutura e trajetória. Uma estrutura pode ser representada diretamente no CIF, enquanto etapas de uma trajetória podem ser convertidas em estruturas.
2. Detecção da simetria e redução da unidade: No AiiDA os materiais modelados são representados como células unitárias não-reduzidas de um cristal. Dessa forma, tais estruturas precisavam ser reduzidas a uma

unidade assimétrica, deixando de fora os átomos simétricos. Para isso foi aproveitado o algoritmo de Grosse-Kunstleve e Adams, umplementado em spglib.
3. Adição de propriedades de estrutura (energias totais, força residual, etc)
4. Adição de metadados para reprodução dos resultados.
5. Armazenamento do arquivo CIF resultante dos passos anteriores para o TCOD utilizando um protocolo http.

DISCUSSÃO

Desde outubro de 2017, o número de registros do TCOD cresceu mais de 2600 e mais de 170 estruturas teóricas depositadas juntamente com sua proveniência foram criadas e depositadas utilizando a interface proposta neste estudo, constituindo 7% dos registros totais no TCOD. Para garantir a integralidade dos arquivos CIF depositados, são realizadas verificações automáticas antes de aceita-las. De falo, já que os dicionários CIF contem descrições formais de itens de dados e seus valores, eles podem ser usados para validação de arquivos.

CONCLUSÃO

Foi apresentada no estudo, a integração da plataforma AiiDA (para executar e gerenciar de forma automática, *workflows* mantendo a proveniência dos dados computados) e o TCOD (que armazena dados associados a estruturas cristalinas usando uma ontologia, dentro de um banco de dados aberto que facilita sua disseminação). A integração permite padronização automática de estruturas cristalinas com metadados, propriedades calculadas e sua proveniência completa (códigos utilizados, entradas, etc). Primeiramente foi estendido os dicionários CIF do TOC para incluir informações de proveniência. Logo após foi definido um protocolo de conversão para contornar as limitações do formato CIF. O principal resultado do trabalho foi a combinação de todos os esforços e da implementação de um conversor para analisar automaticamente a proveniência dos dados armazenados no AiiDA após a execução do *workflow*. dados do banco de dados do TCOD juntamente com a sua proveniência completa pode ser facilmente recuperada e importados de volta para o AiiDA como entrada para mais cálculos e análises. Os autores esperam também que no futuro mais pesquisadores adotem os métodos e ferramentas descritas aqui para tornar os dados públicos (como atualmente exigido por muitas agências de financiamento) com esforço exigido.

ANÁLISE SEGUNDO O LEITOR

1. Qual tipo de proveniência abordada no trabalho?
   Essa informação não é explicita no texto, entretanto os autores citam que a proveniência capturada é relacionada a dependências para re-execução, fonte de dados, e etc. Desta forma assume-se que a proveniência abordada é a proveniência retrospectiva.

2. Qual tipo de ontologia utilizada no trabalho?
   O tipo de ontologia utilizado não é explicita no trabalho, entretanto assume-se que é utilizada uma ontologia de domínio (uma vez que se trata de um trabalho voltado ao domínio de estruturas de materiais), onde a mesma é utilizada como um dicionário de domínio, que classifica itens.

3. Qual a principal vantagem em se utilizar ontologia no contexto da pesquisa?
   A principal vantagem do uso de ontologias percebida no artigo é a disseminação e padronização da informação, possibilitando geração de conhecimento, uma vez que os termos do domínio que compõem a ontologia, são definidos de forma formal.

4. Questões de granularidade são abordadas no artigo?
   Não. A granularidade não é abordada no artigo.