Ontologias e Proveniência em *Workflows* Científicos

Um mapeamento

Luiz Gustavo Dias
Instituto de Computação
Universidade Federal Fluminense
Niterói, Rio de Janeiro
lgdias@id.uff.br

RESUMO

O uso de *workflows* traz diversos benefícios para o processo científico, como a acurácia nos resultados e agilidade no processo experimental. Entretanto, cientistas encontram dificuldades em reusar ou reproduzir experimentos executados por *workflows*, pois a maioria não fornece metadados de proveniência. Paralelo a isto tem-se ontologias, que vêm sendo utilizadas em uma infinidade de situações, principalmente por seu perfil formal, sua estrutura de relacionamentos, organização do conhecimento, compartilhamento da informação, dentre outras características. Desta forma, o presente estudo teve como objetivo investigar o emprego de ontologias para questões de proveniência em *workflows* científicos a partir de um mapeamento sistemático realizado em pesquisas do repositório Scopus, publicadas nos anos de 2016 e 2017. Como resultados tem-se que 50% dos trabalhos do repositório trabalham a proveniência retrospectiva, pesquisas publicadas no último ano mostram tendência em abordar ambos os tipos de proveniência, e que questões de granularidade são pouco abordadas na documentação científica do repositório analisado.

CONCEITOS CCS

• **Information systems → Data provenance**; *Ontologies*; • **Computing methodologies → Scientific visualization**;

PALAVRAS-CHAVE

Ontologias; Proveniência; *Workflows*

1 INTRODUÇÃO

Workflows científicos são compostos por diversos módulos que manipulam dados e parâmetros de entrada, gerando novos dados que servirão de entrada para outros módulos em um fluxo. Sua utilização traz diversas vantagens para quem o emprega [8], como por exemplo a modularidade, que permite por exemplo separar as regras de negócio do suporte operacional, resultando na simplicidade no processo de mudança. Outro benefício advindo do uso de *workflows* é na gestão do conhecimento, uma vez que as informações do experimento são gravadas em arquivos digitais, que seguem fluxo padronizado, o que contribui no processo de rastreabilidade [20]. A partir dessa premissa, torna-se extremamente importante compartilhar e preservar *workflows*, especialmente pelo papel colaborativo.

Entretanto na prática percebe-se que apenas compartilhar *workflows* não garante seu sucesso [2], pois grande parcela é disponibilizada à comunidade, mas não utilizada por não acompanhar informações de apoio a seus usuários. O histórico de um conjunto de dados é conhecido como proveniência, e pode ser caracterizada quanto a especificação do *workflow* (proveniência prospectiva), e quanto a sua execução (proveniência retrospectiva)[6]. A falta de proveniência impacta de forma direta na utilização de *workflows*, influenciando diretamente na qualidade da produção científica, visto que seria resgatada maior relevância, se incluídos contextos viáveis, tipos de dados de entrada e saída e formas de execução quando os mesmos fossem compartilhados [7]. Todo o aparato adicional deve ser pensado de forma a auxiliar principalmente os processos de configuração e utilização.

Paralelo a isto tem-se modelos e métodos utilizados para a descrição de experimentos baseados em ontologias, que voltados a *workflows* apoiam a preservação e reprodução de experimentos, como por exemplo *wfdesc*, e OBI (*Ontology for Biomedical Investigation*) [3]. *wfdesc* é utilizada para descrever a estrutura do *workflow*, mapeando seus estágios destacando a relação dos mesmos, retratando de forma clara o que acontece em cada um dos estágios. OBI por sua vez, provê termos relacionados a investigações no contexto da biomedicina, trabalhando com extenso vocabulário de domínio, auxiliando pesquisadores em duas pesquisas específicas. O perfil categórico e de relacionamentos proporcionado pelo uso de ontologias, possibilita visão detalhada do domínio ou tarefa, e integração com vários recursos desenvolvidos a partir de diversas tecnologias, que variam de módulos lógicos, a linguagens de marcação e *web* semântica.

Tendo em vista o exposto, o presente trabalho objetivou analisar e relacionar os principais trabalhos disponibilizados pela plataforma Scopus[1], publicados nos anos de 2016 e 2017, que abordam a aplicabilidade de ontologias para proveniência em experimentos científicos baseados em *workflow*.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Workflow

Workflows científicos são encontrados em diferentes domínios e são definidos como automação de um processo científico expresso em sub-tarefas, estruturado com base em suas dependências e dados [25]. No contexto da ciência podem ser definidos como um conjunto de processos conectados que descreve um experimento científico, e ao contrário de *workflows* de negócio, são centrados no processo de transformação dos dados [4].

[1] https://www.scopus.com

O ciclo de vida de um *workflow* científico possui quatro estágios básicos segundo Ludäscher et al. [13]:

- Projeto: Consiste na idealização de um *workflow* para a execução de um experimento.
- Instanciação: Consiste ao processo de preparo para aplicar o *workflow* à uma tarefa específica. É importante salientar que neste estágio é importante indicar entradas e parâmetros para sua execução.
- Execução: Diz respeito a execução do *workflow* projetado e instanciado. As sub-tarefas que compõem o *workflow* são executadas, produzem dados que são reusados por outras sub-tarefas dentro do fluxo, com a finalidade de concluir o experimento.
- Análise Pós-Execução: Etapa que consiste na análise dos resultados obtidos com a execução do experimento.

Dentre as principais vantagens em se utilizar *workflows* pode-se citar o aninhamento, uma vez que um *worflow* pode conter *subworkflows*, o que se torna interessante quando trata-se de experimentos ou tarefas complexas. No âmbito de *e-science*, pode-se citar a eficiência, confiabilidade e coordenação experimental como as principais benefícios do uso de *workflows* uma vez que o controle do experimento e a captura de metadados pode ser feito de maneira mais simplificada e transparente ao usuário.

2.2 Proveniência

A proveniência diz respeito ao histórico dos registros de determinados contextos. No que tange *workflows* científicos, está relacionada a metadados que justifiquem e expliquem resultados, e etapas de derivação [6, 21]. Proveniência impacta diretamente no processo de utilização de *workflows*, principalmente se tratando de bases de dado e experimentos complexos. Isso acontece porque em sua grande parte, *workflows* possuem diferentes tipos de dependências que podem interferir de forma direta ou indireta nos resultados experimentais [4].

Outra característica importante da proveniência é o seu nível de detalhamento, denominada granularidade, que possibilita melhor filtragem e manipulação dos metadados capturados. A proveniência pode ser caracterizada em dois tipos distintos, sendo eles o tipo prospectivo e o tipo retrospectivo [6], que são descritos e caracterizados nas subseções a seguir.

2.2.1 Proveniência Prospectiva. A proveniência prospectiva está diretamente relacionada aos metadados referentes a estruturação do *workflow*, fornecida em forma de grafo de atividades que compõem o fluxo e parâmetros necessários para sua execução. Este tipo de proveniência auxilia o usuário a entender a composição do *workflow*. Deste modo, a proveniência prospectiva é definida como a captura da especificação de tarefas computacionais correspondentes a etapas que devem ser seguidas para gerar determinado resultado ou classe de resultados [6].

Existem diversos Sistemas Gerenciadores de *Workflows* Científicos (SGWfCs) que proporcionam a captura deste tipo de proveniência como o YesWorkflow[2], que da suporte a scripts desenvolvidos nas linguagens R, Phyton e MATLAB, além de fazer uso de grafo dirigido acíclico (GDA) para representar a proveniência prospectiva capturada, que por sua vez é armazenada em um banco de dados relacional [14].

2.2.2 Proveniência Retrospectiva. Diferente da proveniência prospectiva, a proveniência retrospectiva diz respeito ao processo de captura de metadados relacionados ao ambiente, junto as etapas que compõe o arranjo do *workflow*, que por sua vez justifica determinado dado ou conjunto de dados [6]. Desta forma, a proveniência retrospectiva pode ser definida como a descrição de forma detalhada de determinado conjunto de dados, que explicita e explica processos e agentes envolvidos no processo de derivação. De forma análoga a proveniência prospectiva, existem diversos SGWfCs que possibilitam a captura da proveniência retrospectiva, como por exemplo o noWorkflow, que captura ambos os tipos de proveniência, além de ser voltado a manipulação de scripts construídos utilizando a linguagem de programação Python, coleta a proveniência e a armazena em um banco de dados relacional, além de possibilitar exportar a proveniência para a linguagem prolog [19].

2.3 Ontologia

Ontologia pode ser entendida como um conjunto de conceitos fundamentais e suas relações, que visam representar determinado domínio de maneira formal (eliminando inconsistências), possibilitando a compreensão do contexto por humanos e computadores [17]. Desta forma, tem-se que o uso de uma estrutura ontológica permite descrever determinada área do conhecimento ou atividade de forma apurada - sem ambiguidades - tornando possível a manipulação de dados e produção de resultados corretos por parte de agentes.

Uma estrutura ontológica é composta por conjuntos de classes, axiomas, instâncias e relações. Classes representam os conceitos dentro do domínio e são criadas a partir de axiomas, que por sua vez são definidos por fórmulas lógicas responsáveis por moldar e definir classes, relações conectam classes a outras classes bem como a instâncias, e instâncias são criadas a partir de classes [10]. De maneira mais simplificada, ontologias podem ser entendidas como um grafo, onde classes e instâncias são representados como vértices que são ligados por relações representados como arestas. Fazendo analogias a árvores, classes são nós mais superiores, e instâncias os nós-folha.

Levando em consideração os diversos domínios e contextos de aplicação, existem diferentes tipos de ontologia, que são mais adequadas para determinadas situações [9]:

- **Ontologias leves** (*lightweight ontologies*): indicadas quando não existe a preocupação de detalhamento de conceitos. São muito utilizadas para categorização de grandes quantidades de informações como por exemplo aquelas manipuladas por motores de busca.
- **Ontologias densas** (*heavyweight ontologies*): indicadas quando é necessário grande nível de detalhamento para conceitos, e sua aplicação é indicada para criação de bases que priorizam o reuso e o compartilhamento.
- **Ontologias de domínio** (*domain ontology*): definem determinado contexto bem como atividades e atores relacionados a ele, define um vocabulário voltado ao compartilhamento dentro de um domínio específico cujos principais objetivos são o compartilhamento e o reuso.

[2] https://github.com/yesworkflow-org/

- **Ontologias de tarefa** (*task ontology*): fornece um vocabulário aplicavel a questões relacionadas a uma tarefa específica, que pode ou não ser independente de domínio.

Dentre os vários benefícios da aplicação de ontologias, cita-se a formalização de domínios e tarefas, bem como a inferência de conhecimento. A partir da formalização de um domínio ou tarefa através de uma estrutura ontológica, é possível aplicar por exemplo um agente raciocinador, como por exemplo Pellet Reasoner[3], com a finalidade de gerar conhecimento novo a partir de informações relacionadas na estrutura.

3 TRABALHOS RELACIONADOS

O trabalho de Pérez et al. [18] visa realizar um levantamento dos principais sistemas de captura de proveniência com dois objetivos principais. O primeiro é identificar suas principais características, e o segundo determinar quais são os mais utilizados e suas características comuns. Como resultados os autores criam uma taxonomia e levantam os 25 sistemas mais utilizados. Os autores apontam que uma das principais características de sistemas baseados em ontologia é a vinculação dos dados, visto seu sistema relacional.

Lopez-Herrejon et al. [12] por sua vez, realizam um mapeamento sistemático com a finalidade de analisar práticas em *Software Product Line* (SPL) tendo em vista as vantagens que o uso dessa prática impacta no produto, como por exemplo o tempo para lançamento no mercado. A finalidade do trabalho é levantar questões sobre técnicas aplicadas, a natureza dos estudos de caso, e quais fases do Teste de Interacção Combinatorial (CIT) foram aplicadas, objetivando identificar tendências comuns, lacunas e oportunidades.

4 MÉTODO

Este estudo possui abordagem qualitativa uma vez que não serão utilizados testes estatísticos no processo de análise de dados. Também é caracterizado como mapeamento sistemático, uma vez que visa identificar determinadas características relacionadas a um tópico [11].

O método utilizado foi composto por duas etapas básicas, que constam em definir o objetivo da pesquisa bem como o protocolo a ser seguido. O objetivo geral "Identificar como pesquisadores utilizam ontologias no contexto da proveniência em experimentos científicos baseados em *workflows* tendo em vista o tipo de proveniência".

A segunda etapa destinada a definir um protocolo de pesquisa foi composto por quatro sub-atividades, apontadas e descritas a seguir:

- **Definir questões de pesquisa:** atividade destinada a definição de perguntas-chave a serem respondidas com a análise dos trabalhos. Deste modo, buscou-se responder quatro perguntas básicas:
  - Q1: Qual o tipo de proveniência abordada na pesquisa analisada?
  - Q2: Qual o tipo de ontologia aplicada na pesquisa analisada?
  - Q3: Quais as principais vantagens em se utilizar ontologias no contexto de proveniência em experimentos baseados em *workflows* na pesquisa analisada?
  - Q4: Qual o nível de granularidade de proveniência abordada na pesquisa analisada?
- **Definir fontes de pesquisa:** atividade destinada a definição da base de dados utilizada para coleta de dados. Nesta atividade ficou definida a base de dados Scopus[4];
- **Definir *string* de busca:** atividade destinada ao desenvolvimento da *string* de busca utilizada para triagem inicial dos dados. Os critérios para selecionar as pesquisas foram todos os trabalhos disponíveis no repositório publicados entre os anos de 2016 e 2017, que contivessem as palavras "*ontology*" e "*workflow*" e "*provenance*" em seu *abstract*. Tais critérios foram traduzidos na seguinte *string*: ABS (ontology AND workflow AND provenance) AND (LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016));
- **Definir critérios de seleção:** atividade voltada ao processo de triagem secundária. Foram definidos como critério de inclusão:
  - I1: A pesquisa deve ser escrita em inglês;
  - I2: A pesquisa deve ser fornecida de forma gratuita;
  - I3: Os termos da string de busca devem ser relacionados na problemática da pesquisa;

  Foi definido como critério de exclusão:
  - E1: A pesquisa não cumpre pelo menos um critério de inclusão.

5 RESULTADOS

5.1 Referente ao tipo de proveniência

Durante a análise deste primeiro item percebeu-se que a informação não é explícita nas pesquisas, sendo assim, o tipo de proveniência foi definido com base em informações implícitas. Após analisar as pesquisas com base em Q1, notou-se que 50% dos trabalhos analisados abordam a proveniência retrospectiva, ou seja preocupam-se com aspectos relacionados a captura de metadados que justifiquem e expliquem os objetos de dados resultantes durante a execução de um *workflow* ou módulo. Foi percebido também que 25% dos trabalhos abordam proveniência prospectiva, e 25% abordam ambos os tipos de proveniência.

Os trabalhos que abordam apenas a proveniência retrospectiva possuem quase em sua totalidade o ano de publicação como característica comum, 3/4 dos trabalhos que abordam este tipo de proveniência foram publicados em 2016. Frank and Zander [5], objetivam a captura de proveniência retrospectiva de experimentos que fazem uso de dados heterogêneos do domínio espaço-temporal buscando principalmente normalização dos dados. Wu et al. [24] por sua vez, propõem um sistema genérico acoplável a sistemas de análise de dados para captura de proveniência, e Stillerman et al. [22], propoem um sistema de recuperação de relações entre dados e metadados. Merkys et al. [15] diferentemente das pesquisas citadas anteriormente, publicaram sua pesquisa em 2017, e abordaram a captura de proveniência retrospectiva relacionada a ciência dos materiais. O motivo de grande parcela dos trabalhos publicados em

[3] https://github.com/stardog-union/pellet

[4] https://www.scopus.com

2016 focarem na proveniência retrospectiva pode ser justificado pelo recente interesse científico na captura de proveniência para a reprodutibilidade de experimentos científicos baseados em *workflow*, e também pelo perfil da proveniência retrospectiva (justificar dados resultantes e suas transmutações).

Diferentemente das pesquisas que abordam a proveniência retrospectiva, os estudos que abordam a proveniência prospectiva foram publicados em anos diferentes. Abla et al. [1], desenvolve um sistema para documentação automática de *workflows*. Valdez et al. [23], desenvolve a ferramenta ProvCaRe, um sistema de domínio específico voltado a fornecer informações referentes a método, ferramentas e dados de pesquisas.

A última parcela de pesquisas analisadas neste item, é referente aos trabalhos que abordam ambos os tipos de proveniência, sendo eles os trabalhos de Miksa and Rauber [16] e Zhang et al. [26]. Enquanto Zhang et al. [26] utiliza uma estrutura específica para coordenar *web services*, capturar e armazenar a proveniência do domínio geopespacial, Miksa and Rauber [16] propõem e integraram ontologias ja existentes para armazenar informaçõs sobre *software*, *hardware* e arquivos referentes a *workflows*.

A partir dos resultado adquiridos com a análise de Q1 no conjunto de dados em questão, nota-se interesse científico em caputar ambos os tipos de proveniência utilizando estruturas que manipulem *workflows* ou possam ser acopladas a ferramentas de análise de dados independente de domínio.

5.2 Referente ao tipo de ontologia

O segundo passo para análise das pesquisas constou em identificar o tipo de ontologia utilizados nos estudos, e assim como na primeira análise, tal informação não é disponibilizada de forma explícita. A etapa constou em classificar as estruturas ontológicas em dois tipos, de domínio ou de tarefa, onde ontologias de domínio são aquelas desenvolvidas a partir de vocabulário específico de determinada área do conhecimento, como a biomedicina por exemplo, e ontologias de tarefa são aquelas voltadas a armazenar informações para a execução de uma tarefa específica, como por exemplo a captura de proveniência de experimentos científicos baseados em *workflow*.

Após análise notou-se que ontologias de tarefa foram mais aplicadas que ontologias de domínio no ano de 2016. Foi notado também que em trabalhos mais recentes foram abordadas ontologias de domínio com maior frequência.

Tomando como base a ordem cronológica, foram publicadas em 2016 três pesquisas que fazem uso de ontologia de tarefa [1, 22, 24] contra uma pesquisa que faz uso de ontologia de domínio [5]. Wu et al. [24] aplicam um sistema de captura de metadados denominado, para análise de *workflows* heterogêneos distribuídos, onde o usuário possui total controle da instrumentação para capturar diferentes níveis de granularidade da proveniência. Stillerman et al. [22] por sua vez, não documentam o desenvolvimento de uma ferramenta mas sim apresentam uma proposta de desenvolvimento de um sistema que faz uso de ontologia e trata questões como armazenamento e recuperação das relações entre os dados e metadados. Por fim, Abla et al. [1] realizam a descrição do sistema de captura de proveniência baseado em ontologia MPO, explicitando questões como sua arquitetura, e base de dados. O uso da ontologia nestes casos é justificado pela problemática abordada pelos autores, que é centrada em uma atividade específica, que é a captura de proveniência de *workflows*.

Enquanto Frank and Zander [5] utilizam uma ontologia de domínio para normalização de dados heteronêneos do domínio geográfico. O uso desse tipo de ontologia pode ser justificado pelo reuso, visto que a existência de várias ontologias específicas de domínio que podem ser reutilizadas e reaplicadas em contextos diferentes de um mesmo domínio. O que possibilita por exemplo, identificar dependências necessárias a determinadas atividade.

Pesquisas mais recentes por sua vez utilizaram em sua maioria ontologias de domínio. Enquanto Merkys et al. [15], Valdez et al. [23], Zhang et al. [26] fizeram uso de ontologia de domínio para relacionar informações de proveniência advindos dos domínios da biomedicina, geoespacial e ciência dos materiais, respectivamente. Enquanto Miksa and Rauber [16] fizeram uso de uma ontologia de tarefa para descrever *workflows* e ambientes em que são executados. O uso deste tipo de ontologia neste caso pode ser justificado pelo objetivo de capturar ambos os tipos de proveniência. Ontologias de domínio mapeam e conectam atores de um mesmo domínio por meio de relações, desta forma, o processo de identificar itens faltantes em determinada relação, ou mesmo normalizar dados advindo de diferentes procedências, pode se tornar menos custoso computacionalmente, visto o perfil formal de ontologias. Uma justificativa para a utilização de ontologias de tarefa para proveniência retrospectiva, é que a mesma está diretamente ligada a execução do *workflow*, capturando informações que justifiquem os resultados obtidos em cada execução.

A partir disso, percebeu-se que o foco dos trabalhos publicados em 2016 foi generalização da tarefa de captura de proveniência, possibilitando que *workflows* de diferentes domínios fossem aplicados a soluções que capturam a proveniência do tipo prospectivo. Por outro lado, foi notado que pesquisas mais recentes fizeram uso de ontologia de domínio, e apesar de não permitirem aplicação de forma global como as pesquisas que manipulam metadados prospectivos, abordam ambos os tipos de proveniência.

5.3 Referente as vantagens da utilização de ontologias e granularidade

A vantagem da utilização de ontologias pode ser percebido principalmente pelo seu perfil de relacionamentos e formalidade. Sua estrutura hierárquica possibilita o comportamento de herança entre classes, subclasses e instâncias, fazendo com que componentes em níveis mais baixos da hierarquia, herdem características definidas formalmente de componentes em níveis mais altos.

Essa característica aplicada no contexto de proveniência, pode por exemplo, possibilitar identificar itens ausentes relacionados a aquisição de determinado objeto de dados. Vale ressaltar também a normalização de dados heterogêneos, e o reuso, visto a existência de diferentes tipos de ontologias em diferentes domínios. Outra vantagem da aplicação da ontologias nesta problemática é a possibilidade de adequar ontologias existentes para outro fim, e até mesmo mesclar ontologias de diferentes tipos para a captura de proveniência (por exemplo, uma ontologia de domínio agregada a uma ontologia de tarefa).

Após análise dos trabalhos, foi notado que a granularidade é pouco abordada na documentação científica, visto que apenas um trabalho abordou este tópico [24].

6 CONCLUSÃO E TRABALHOS FUTUROS

Enquanto estudos menos recentes são focados na manipulação de proveniência retrospectiva, estudos mais recentes focam na proveniência prospectiva e em ambas.

O tipo de ontologia aplicada está diretamente ligada ao tipo de proveniência abordada, entretanto não é regra. Enquanto estudos relacionados a captura de proveniência retrospectiva são aplicadas ontologias de tarefa, estudos relacionados a captura de proveniência prospectiva e ambas, utilizam geralmente ontologias de domínio.

Relacionado as vantagens de se utilizar ontologias para proveniência, além de formalizar o domínio minimizando inconsistência e auxiliar na gestão do conhecimento, ontologias podem trazer benefícios consideráveis quando se trata de normalização de dados heterogêneos. Através deste estudo verificou-se também que questões que abordam a granularidade da proveniência pouco abordadas na documentação científica, logo não se pode afirmar benefícios sobre a utilização de ontologias relacionando-as a granularidade da proveniência.

Como trabalho futuro espera-se adicionar mais informações no mapeamento, visando identificar quais ontologias de domínio são mais reutilizadas, e quais tipos de estudo de caso são mais realizados. Planeja-se também executar um experimento baseado em ontologias e raciocinadores[5], com o objetivo de verificar seu desempenho na recuperação dos componentes da relação entre dados heterogêneos e metadados, bem como suas características no melhor caso, caso médio, e pior caso.

Uma ontologia contendo os resultados deste mapeamento foi desenvolvida, e está disponível juntamente com a lista de trabalhos analisados, e anotações produzidas durante o processo de pesquisa em um repositório do Github, para ter acesso utilize o QR Code abaixo:

REFERÊNCIAS

[1] Gheni Abla, Elizabeth N Coviello, Sean M Flanagan, Martin Greenwald, Xia Lee, Alex Romosan, David P Schissel, Arie Shoshani, Joshua Stillerman, John Wright, et al. 2016. The MPO system for automatic workflow documentation. *Fusion Engineering and Design* 112 (2016), 1007–1013.

[2] Khalid Belhajjame, Jun Zhao, Daniel Garijo, Matthew Gamble, Kristina Hettne, Raul Palma, Eleni Mina, Oscar Corcho, José Manuel Gómez-Pérez, Sean Bechhofer, et al. 2015. Using a suite of ontologies for preserving workflow-centric research objects. *Web Semantics: Science, Services and Agents on the World Wide Web* 32 (2015), 16–42.

[3] Ryan R Brinkman, Mélanie Courtot, Dirk Derom, Jennifer M Fostel, Yongqun He, Phillip Lord, James Malone, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, et al. 2010. Modeling biomedical experimental processes with OBI. *Journal of biomedical semantics* 1, 1 (2010).

[4] Susan B Davidson and Juliana Freire. 2008. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 1345–1350.

[5] Matthias Frank and Stefan Zander. 2016. Smart Web Services for Big Spatio-Temporal Data in Geographical Information Systems.. In *SALAD@ ESWC*.

[6] Juliana Freire, David Koop, Emanuele Santos, and Cláudio T Silva. 2008. Provenance for computational tasks: A survey. *Computing in Science & Engineering* 10, 3 (2008).

[7] Kristina M Hettne, Harish Dharuri, Jun Zhao, Katherine Wolstencroft, Khalid Belhajjame, Stian Soiland-Reyes, Eleni Mina, Mark Thompson, Don Cruickshank, and Lourdes Verdes-Montenegro. 2014. Structuring research methods and data with the research object model: genomics workflows as a case study. *Journal of biomedical semantics* 5 (2014).

[8] David Hollingsworth. 1994. Workflow management coalition: The workflow reference model. *The Workflow Management Coalition* (Outubro 1994).

[9] Seiji Isotani and Ig Ibert Bittencourt. 2015. *Dados Abertos Conectados*. Novatec Editora, São Paulo.

[10] Atanas Kiryakov, Damyan Ognyanov, and Dimitar Manov. 2005. OWLIM–a pragmatic semantic repository for OWL. In *International Conference on Web Information Systems Engineering*. Springer, New York, 182–192.

[11] B. Kitchenham and S Charters. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering.

[12] Roberto E Lopez-Herrejon, Stefan Fischer, Rudolf Ramler, and Alexander Egyed. 2015. A first systematic mapping study on combinatorial interaction testing for software product lines. In *Software Testing, Verification and Validation Workshops (ICSTW), 2015 IEEE Eighth International Conference on*. IEEE, 1–10.

[13] Bertram Ludäscher, Mathias Weske, Timothy McPhillips, and Shawn Bowers. 2009. Scientific workflows: Business as usual?, In International Conference on Business Process Management. *International Journal of Digital Curation*, 31–47.

[14] Timothy McPhillips, Tianhong Song, Tyler Kolisnik, Steve Aulenbach, Khalid Belhajjame, Kyle Bocinsky, Yang Cao, Fernando Chirigati, Saumen Dey, Juliana Freire, et al. 2015. YesWorkflow: a user-oriented, language-independent tool for recovering workflow information from scripts. *arXiv preprint arXiv:1502.02403* (2015).

[15] Andrius Merkys, Nicolas Mounet, Andrea Cepellotti, Nicola Marzari, Saulius Gražulis, and Giovanni Pizzi. 2017. A posteriori metadata from automated provenance tracking: integration of AiiDA and TCOD. *Journal of cheminformatics* 9, 1 (2017), 56.

[16] Tomasz Miksa and Andreas Rauber. 2017. Using ontologies for verification and validation of workflow-based experiments. *Web Semantics: Science, Services and Agents on the World Wide Web* 43 (2017), 25–45.

[17] Riichiro Mizoguchi. 2004. Tutorial on ontological engineering Part 2: Ontology development, tools and languages. *New Generation Computing* 22, 1 (2004), 61–96.

[18] Beatriz Pérez, Julio Rubio, and Carlos Sáenz-Adán. 2018. A systematic review of provenance systems. *Knowledge and Information Systems* (2018), 1–49.

[19] Joao Felipe Pimentel, Leonardo Murta, Vanessa Braganholo, and Juliana Freire. 2017. noWorkflow: a tool for collecting, analyzing, and managing provenance from python scripts. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1841–1844.

[20] Queli Terezinha Schmitz and Marcelo Macedo. 2011. Ferramenta de Workflow para Apoio ao Processo de Gestão do Conhecimento. *XXXI Encontro Nacional de Engenharia de Produção* (Outubro 2011).

[21] Umber Sheikh, Abid Khan, Bilal Ahmed, Abdul Waheed, and Abdul Hameed. 2018. Provenance inference techniques: Taxonomy, comparative analysis and design challenges. *Journal of Network and Computer Applications* (2018).

[22] Joshua Stillerman, Thomas Fredian, Martin Greenwald, and John Wright. 2016. A general purpose tool-set for representing data relationships: Converting data into knowledge. In *Scientific Data Summit (NYSDS), 2016 New York*. IEEE, 1–6.

[23] Joshua Valdez, Michael Rueschman, Matthew Kim, Sara Arabyarmohammadi, Susan Redline, and Satya S Sahoo. 2017. An Extensible Ontology Modeling Approach Using Post Coordinated Expressions for Semantic Provenance in Biomedical Research. In *OTM Confederated International Conferences On the Move to Meaningful Internet Systems*. Springer, 337–352.

[24] Kesheng Wu, Elizabeth N. Coviello, S. M. Flanagan, Martin Greenwald, Xia Lee, Alex Romosan, David P. Schissel, Arie Shoshani, Josh Stillerman, and John Wright. 2016. MPO: A System to Document and Analyze Distributed Heterogeneous Workflows. In *Provenance and Annotation of Data and Processes*, Marta Mattoso and Boris Glavic (Eds.). Springer International Publishing, Cham, 166–170.

[25] Jia Yu and Rajkumar Buyya. 2005. A taxonomy of scientific workflow systems for grid computing. *ACM Sigmod Record* 34, 3 (2005), 44–49.

[26] Mingda Zhang, Joshua Lieberman, and Peng Yue. 2017. Ontology for processing service orchestration. In *Agro-Geoinformatics, 2017 6th International Conference on*. IEEE, 1–5.

[5] https://protegewiki.stanford.edu/wiki/Using_Reasoners