# Dados Conectados na Educação

**Crystiam K. Pereira[1], Sean W. M. Siqueira[1], Bernardo Pereira Nunes[1,2]**

[1]Programa de Pós-Graduação em Informática, Universidade Federal do Estado do Rio de Janeiro (UNIRIO) – Rio de Janeiro– RJ – Brasil

[2]Departamento de Informática, PUC-Rio – Rio de Janeiro – RJ – Brasil

{crystiam.kelle, sean}@uniriotec.br, bnunes@inf.puc-rio.br

***Abstract.*** *The adoption of Linked Data in Education can bring advances to Technology Enhanced Learning, even though some obstacles still need to be addressed. In this paper, we discuss these obstacles from different perspectives, such as privacy and data ownership, the need for public policies to encourage the openness and reuse of data, the difficulties in data publishing and consumption as well as the construction of a scenario of technological innovation in Education through the use of Linked Data. Finally, we also present relevant actions and evaluation indicators to further advance in this field.*

***Resumo.*** *O uso de Dados Conectados na Educação pode trazer avanços para a área de Informática na Educação. Entretanto, alguns obstáculos ainda precisam ser superados. Neste artigo nós os discutidos sob diferentes perspectivas, tais como a propriedade e a privacidade dos dados, a necessidade de políticas públicas para incentivar sua abertura e reutilização, as dificuldades na publicação e consumo de dados, bem como na construção de um cenário de inovação tecnológica na Educação através do uso de Dados Conectados. Finalmente, apresentamos também ações e indicadores de avaliação para avançar nesse cenário.*

## 1. Contextualização

Dados Conectados podem ser resumidos como o uso da Web para criar conexões entre dados que podem estar armazenados originalmente em diversos bancos de dados, mantidos por diferentes organizações e distribuídos em diferentes localizações geográficas [Bizer et al. 2009].

D'Aquin [2012] explica que há potencial para o uso de Dados Abertos Conectados na Educação em virtude da natureza aberta e acessível dos dados e dos recursos educacionais produzidos por muitas universidades. Dietze et al. [2013] destacam que a publicação e o consumo desses dados podem promover a integração de dados e serviços educacionais, buscando resolver a heterogeneidade entre padrões de APIs (*Application Programming Interface)* e repositórios e, ainda, enriquecer dados através da adoção de vocabulários e da interligação entre dados de fontes heterogêneas.

Em [Alcantara et al. 2015] são apresentados alguns desafios no uso de Dados (Abertos) Conectados na realidade da Educação Brasileira. Os autores mostram que ainda existem poucas iniciativas de abertura de dados educacionais no Brasil, concluindo que a cultura de publicá-los praticamente não existe. No presente trabalho, ampliamos a visão

apresentada em [Alcantara et al. 2015]: (*i*) os desafios aqui apresentados foram compilados a partir de um mapeamento sistemática de propostas que relataram a publicação e/ou o uso de Dados Conectados na educação em diferentes iniciativas ao redor do mundo, tendo assim desafios específicos da educação, mas também, desafios pertinentes à área de Dados Conectados que se refletem na educação; (*ii*) damos ênfase a questões técnicas do uso de Dados Conectados na Educação; (*iii*) focamos em desafios para a publicação dos dados nos padrões de Dados Conectados, já que a publicação de dados em outros formatos (CSV, PDF, DOC, etc) vem evoluindo, mesmo que a passos lentos; (*iv*) detalhamos desafios existentes em etapas específicas do processo de publicação como a etapa de interligação dos dados; (*v*) discutimos aspectos não técnicos como a privacidade pessoal; e (*vi*) apresentamos ações, soluções e tecnologias adotadas para evoluir tais desafios. Além disso, foram considerados os desafios relacionados à inovação tecnológica na Educação, através do uso de dados conectados, com base no ecossistema sugerido em [Kapoor et al. 2015][Siqueira et al. 2016].

As ações sugeridas e associadas a cada um dos desafios são propostas como questões a serem discutidas para implantação a curto, médio e longo prazo, sendo que algumas caminham em paralelo aos desafios gerais da área de Dados Conectados. A visão dos desafios enfrentados por iniciativas ao redor do mundo, bem como possíveis soluções e encaminhamentos podem auxiliar em futuras discussões e direcionar pesquisas e práticas direcionadas para a realidade brasileira.

## 2. Políticas Públicas de Abertura de Dados Conectados

O incentivo à publicação de dados passa pelo âmbito do desenvolvimento de políticas públicas, de diretrizes de segurança e de definições em relação à propriedade intelectual dos dados que serão publicados e reutilizados. Desafios associados ao incentivo e à cultura de abertura dos dados podem ser sintetizados em: quais são as políticas públicas (e como implantá-las) para incentivar a publicação de dados e desenvolver a cultura de abertura e reutilização de dados?

A realidade brasileira aponta a necessidade de ampliar as discussões e colocar em prática ações imediatas para incentivar a adoção de Dados Abertos Conectados na Educação. Uma auditoria do Tribunal de Contas da União (TCU)[1] (AC-3022-48/15-P), realizada em 2015, avaliou a efetividade das iniciativas de algumas instituições (Ministério da Educação, Fundo Nacional de Desenvolvimento da Educação, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) quanto à abertura e transparência dos seus dados governamentais. Os resultados apontaram, entre outros problemas, baixa atuação para promover a abertura de dados públicos, falta de ações estratégicas para a abertura dos dados, disponibilização de dados em formatos que não permitem a leitura por aplicações automáticas e ausência de dados públicos importantes, que ainda não estavam sendo disponibilizados pelas instituições.

A LAI – Lei de Acesso à Informação, Lei nº 12.527/2011, regulamenta o direito constitucional de acesso às informações públicas. Entretanto, dada à importância do amplo acesso às informações relativas à Educação, observa-se que políticas públicas mais específicas precisam ser desenvolvidas para que dados educacionais em seus mais diversos níveis sejam disponibilizados de modo aberto e conectado. Isso permitiria não apenas entender as

[1]https://contratospublicos.com.br/tcu-auditoria-operacional-efetividade-de-iniciativas-de-abertura-de-dados-governamentais-na-area-de-educacao-recomendacoes-determinacoes-arquivamento/jurisprudencia

estratégias bem e mal sucedidas, mas, principalmente, promover um ambiente de inovação nessa área, resultando em Educação de melhor qualidade com apoio de tecnologia.

O número de instituições educacionais que disponibilizam seus dados de forma aberta e conectada é um indicador importante da efetividade das políticas públicas. A quantidade de *datasets* governamentais e a existência de auditorias e acórdãos no sentido de fiscalizar as iniciativas também são indicadores importantes, tanto do comprometimento governamental com a disponibilização dos dados, quanto da cobrança para que as políticas existentes sejam seguidas.

## 3. Privacidade pessoal em ambientes de dados educacionais conectados

Antes mesmo de discutir questões operacionais e tecnológicas, há um grande desafio em relação à privacidade pessoal e à propriedade dos dados educacionais. Uma vez que os dados educacionais passem a ser publicados de forma aberta, algumas informações sensíveis dos usuários podem ser expostas, ademais, a conexão dos dados pode revelar informações pessoais, ferindo assim direitos de privacidade pessoal. Esse é um debate necessário, que requer engajamento por parte de diferentes atores e setores. Algumas questões importantes envolvidas são: (*i*) a quem pertencem os dados educacionais das instituições?; (*ii*) como e quais dados podem ser disponibilizados?; e (*iii*) como os dados disponíveis serão utilizados?

Muitos dados educacionais produzidos pelas instituições pertencem aos indivíduos (alunos, professores, tutores) ligados a ela, por isso, podem ser sensíveis à divulgação. No entanto, se por um lado há dados que podem ser particulares a seus usuários, por outro lado os dados educacionais são importantes para promover o entendimento do cenário atual e o desenvolvimento de novas políticas educacionais. Desse modo, deve haver uma política nacional de abertura e conexão dos dados educacionais, mas também uma atenção à privacidade dos dados.

Jeremić et al. [2013] destacam a importância de um envolvimento da comunidade global da Web, não apenas dos educadores, no debate sobre a percepção da privacidade e da propriedade dos dados neste novo contexto de aprendizagem em rede. Corsar et al. [2014] reforçam que o assunto deve ser discutido por diferentes grupos da nossa sociedade e cita pelo menos quatro deles: (1) a comunidade de Dados Conectados (incluindo pesquisadores, desenvolvedores e profissionais); (2) os desenvolvedores de software que obtêm e usam informações pessoais de usuários; (3) os indivíduos que compartilham suas informações pessoais; e (4) os comitês de ética.

Algumas ações que podem ser encaminhadas para avançar nesse debate são: (a) pensar, discutir e implantar políticas de privacidade, diretrizes sobre o que e como algumas informações podem ser publicadas e reutilizadas; (b) criar ferramentas que auxiliem na verificação automática de possíveis problemas de privacidade dos dados (através da análise de algumas propriedades de vocabulários e ontologias, por exemplo); (c) criar alternativas para publicar dados educacionais, mantendo o anonimato dos indivíduos; e (d) buscar alternativas para publicar dados de múltiplos usuários, tornando mais difícil inferir características individuais.

São indicadores importantes para avaliar os avanços a respeito da privacidade e propriedade dos dados: existência de políticas de privacidade dos dados; a existência de licenças indicando como os dados podem ser publicados, modificados e reutilizados;

cumprimento de restrições impostas pelo usuário (se houver); existência e efetividade de mecanismos que garantam o sigilo dos dados pessoais; número de ferramentas que ajudem a detectar possíveis indícios de problemas de privacidade de dados; número de diretrizes com foco na proteção à privacidade dos dados.

## 4. Publicação de Dados

Grande parte dos dados produzidos por sistemas educacionais estão armazenados em diferentes formatos e fontes de dados. Nesse sentido, alguns desafios se mostram relevantes: (*i*) como permitir que os dados sejam transformados e estruturados como Dados Abertos Conectados?; (*ii*) como considerar as especificidades educacionais / pedagógicas desses dados?; (*iii*) como encontrar vocabulários/ontologias para serem reutilizados na publicação?; (i*v*) como integrar/interoperar diferentes vocabulários de forma que seja possível atender as especificidades de diferentes domínios-alvo?; e (*v*) sabendo-se que a educação é um processo contínuo e passível de constantes mudanças, como tratar a atualização dos dados disponibilizados e as mudanças estruturais?

Apesar dos esforços no desenvolvimento e na adoção de várias ferramentas (D2RQ[2], DB2RDF[3], SIMILE RDFIzer[4] e Triplify[5]) para apoiar o processo de publicação de dados, esse ainda requer conhecimentos técnicos muito específicos, por exemplo, o conhecimento de tecnologias semânticas (formatos, vocabulários, ferramentas de conversão e manipulação de dados). Alcantara et al. [2015] propõem como solução a utilização de ferramentas que flexibilizem a interface entre o provimento e a disponibilização dos dados, podendo disponibilizar os dados em diversos formatos que atendam aos diferentes públicos. Outro direcionamento importante é a integração de ferramentas aos principais sistemas educacionais, tendo como objetivo tornar o processo de publicação mais transparente ao usuário, não exigindo conhecimento ou esforços específicos para adaptação aos padrões de publicação de Dados Abertos Conectados.

Os vocabulários/ontologias são especialmente importantes para o uso de Dados Conectados na Educação. Embora existam esforços consideráveis na construção de vocabulários específicos para o domínio educacional, muitos trabalhos acabam criando novos vocabulários (ou estendendo algum já existente) para atender a aspectos específicos de um *dataset* ou por dificuldade em localizá-los.

Jeremić et al. [2013] discutem que o mapeamento entre vocabulários mais específicos e alguns mais amplos e mais utilizados é um direcionamento importante para integrar/interoperar diferentes vocabulários. Essa questão já foi objeto de pesquisa em [Zablith et al. 2011][Dietze et al. 2013], mas diversos problemas persistem como o uso de vocabulários heterogêneos, interpretações e usos distintos para o mesmo padrão de metadados, bem como a falta de consenso nos descritores e descrições.

Jeremić et al. [2013] complementam destacando a necessidade de construir ontologias de domínio para diferentes domínios-alvo, buscando assim ampliar a diversidade de domínios. Disponibilizar os vocabulários/ontologias em bases de conhecimento de anotações de serviços é essencial para melhorar a localização e reutilização desses vocabulários.

[2] http://d2rq.org/
[3] https://sourceforge.net/projects/db2rdf/
[4] http://simile.mit.edu/RDFizers/
[5] http://semanticweb.org/wiki/Triplify.html

As mudanças contínuas existentes nos repositórios de recursos educacionais são discutidas em [Dietze et al. 2013]. O desafio se estende às mudanças constantes da estrutura de dados ao longo das atividades de aprendizagem, que exigem constantes adaptações para representar corretamente o estado de organização de pessoas, grupos e atividades. É importante também ressaltar que as ontologias usadas precisam continuar representando esse domínio em constante mudança. Em [Vasconcellos et al. 2014], o tema do refinamento automático de ontologias é apresentado como uma possível forma de permitir atualizações em ontologias de forma que as mesmas retratem o domínio atual. Ainda assim, diversas questões continuam em aberto, particularmente: como manter a consistência e manutenção das versões de ontologias e as aplicações em uso e como garantir a confiabilidade de tais refinamentos automáticos.

Faz parte da publicação de dados a etapa de interligação de dados, intra e inter *datasets*. Os desafios dessa etapa na área de Dados Conectados são discutido por [Nunes 2014][Nunes et al. 2013][Nguyen et al. 2013][Wölger et al. 2011]. Os desafios dessa etapa estão presentes na área de Dados Conectados em diferentes domínios e se estendem ao contexto educacional [Dietze et al. 2013]. O uso de vocabulários comuns e o mapeamento entre diferentes vocabulários são ações que ajudam a estabelecer conexões entre os dados abertos educacionais.

O número de *datasets* educacionais disponibilizados, o grau de interligações entre dados ou recursos educacionais em um mesmo *dataset* e entre *datasets* diferentes, o número de vocabulários educacionais e para domínios específicos (diferentes áreas, por exemplo), quantitativo de repositórios de *dataset* e de vocabulários, a completude dos dados disponibilizados (medida pela capacidade do *dataset* em responder consultas, por exemplo) são indicadores que podem ajudar a avaliar as iniciativas de publicação de Dados Abertos Conectados na Educação.

## 5. Consumo de Dados Educacionais Abertos Conectados

Uma vez que os dados sejam disponibilizados de forma aberta e conectada, há grande potencial na sua utilização. No entanto, para que seja possível usufruir dos benefícios provenientes do consumo de Dados (Abertos) Conectados, alguns desafios também são encontrados: (i) como localizar Dados Abertos Conectados de forma efetiva?; (ii) como encontrar e conectar fontes de dados com assuntos específicos dentro da prática educacional; (iii) como identificar *datasets* para domínios específicos; e (iv) como lidar com a indisponibilidade de *datasets*.

Recuperar Dados Abertos Conectados ainda é um processo que exige conhecimento técnico dos padrões, linguagens e *interfaces*. Ações no sentido de criar interfaces com maior usabilidade e integradas aos ambientes educacionais são necessárias para a construção de sistemas inteligentes que possam impactar de forma eficaz na prática educacional.

A divulgação de novos *datasets* com uma notação semântica apropriada e a criação de repositórios/catálogos específicos para esse fim são ações que podem contribuir para a localização de dados e de recursos educacionais. Disponibilizar dados em formatos que possam ser explorados por outras aplicações também é essencial para a criação de aplicações inteligentes que possam, entre outras coisas, associar recursos a temas específicos, sem a exigência de exaustivas buscas manuais.

A indisponibilidade de *datasets* torna-se um problema de forte impacto quando aplicações passam a explorar conjuntos de dados. Esse ainda é um desafio na área. Hogan et al. [2016] mostram que muitos *datasets* descritos em artigos se tornam indisponíveis pouco tempo após a divulgação/aceitação. Ações focadas na manutenção e atualização constantes de *datasets* são importantes para incentivar o desenvolvimento de aplicações externas e garantir que elas tenham acesso estável e permanente aos dados.

Além dos desafios já citados, há uma grande barreira a ser transposta no sentido de gerar aplicações inovadoras a partir do consumo de Dados Abertos Conectados na Educação. Uma vez que os dados produzidos sejam disponibilizados, despontam-se grandes oportunidades para a elaboração de soluções inovadoras, seja através de análise do contexto educacional atual e tomada de decisão mais consciente e direcionada aos problemas revelados, seja pelo desenvolvimento de aplicações que explorem os dados e recursos disponíveis para melhorar a qualidade da educação brasileira.

Avaliar o consumo de Dados Abertos Conectados na Educação seria possível através de indicadores como: quantidade de aplicações que se beneficiam de dados abertos conectados para trazer soluções e modelos inovadores; quantidade de relatos (publicações científicas, relatórios, experiências etc.) de soluções educacionais desenvolvidas usando dados abertos conectados. A taxa de disponibilidade dos *datasets,* tempo de resposta às consultas, a disponibilidade dos *endpoints* para consultas SPARQL, a capacidade de consumo dos dados por aplicações automáticas e a escalabilidade dos *datasets* também podem ajudar a avaliar a situação e as melhorias no consumo de Dados Abertos Conectados Educacionais.

## 6. Considerações Finais

Ainda que haja inúmeros desafios, em diferentes perspectivas, o uso de dados conectados na educação está em constante evolução e existem muitas iniciativas buscando enfrentar os desafios listados neste trabalho.

Alguns projetos importantes na publicação e utilização de dados conectados na educação são: o dataset TED Talks[6] que possui diversas conferências sobre diferentes temas; os datasets do British Museum[7] e da Bibliothèque Nationale De France[8], que disponibilizam dados dos seus acervos; a *Open University* [Qing et al. 2012], que publica informações sobre seus cursos; as iniciativas da University of Southampton[9], Greek University Open Data[10] e Linking Italian University Statistics Project[11] que publicam seus dados, incluindo informações administrativas, acadêmicas, produção científica, indicadores educacionais; o mEducator - Linked Educational Resources e o PubMed, que fornecem recursos educacionais num formato de Dados Conectados. Vale ressaltar ainda o projeto LinkedUp[12], que mantém um catálogo e um repositório de dados conectados, ambos relevantes e úteis no cenário educacional [D'Aquin 2016].

Diferentes abordagens/tecnologias estão sendo exploradas e evoluídas para enfrentar

6 http://data.linkededucation.org/request/ted/sparql

7 http://collection.britishmuseum.org/

8 http://data.bnf.fr/

9 http://sparql.data.southampton.ac.uk/

10 http://www.auth.gr/sparql

11 http://sw.unime.it:8890/sparql

12 http://data.linkededucation.org/linkedup/catalog/

os desafios na área de Dados Conectados. Para a etapa de publicação de dados, por exemplo, alguns catálogos podem auxiliar o usuário na busca de vocabulários em diferentes domínios, entre eles estão o LOV[13], o BioPortal[14], o JoinUp[15] e o SCHEMA.ORG[16]. Outros vocabulários úteis no contexto educacional podem ser encontrados em https://linkededucation.wordpress.com/data-models/schemas/. Na etapa de interligação de dados, existem propostas explorando abordagens/tecnologias distintas para melhorar os resultados de conexão entre dados inter e intra *datasets*, são algumas delas: técnicas de similaridade, busca em grafos, clusterização, inferência (*reasoner*) baseada em lógica e técnicas de aprendizado de máquina [Nikolov et al. 2012][Soru and Ngonga Ngomo 2014].

Os desafios no uso de Dados Abertos Conectados no contexto educacional estão muito relacionados com aqueles percebidos na grande área de pesquisa de Dados Abertos Conectados. Porém, alguns aspectos educacionais requerem ações e direcionamentos de pesquisa específicos e precisam ser trabalhados pela comunidade de Informática na Educação (IE). Esta proposta visa listar estes grandes desafios para a comunidade de IE no Brasil, indicando também possíveis projetos e indicadores de avaliação.

A visão dos desafios enfrentados por iniciativas ao redor do mundo, bem como possíveis soluções e encaminhamentos podem auxiliar em discussões futuras e direcionar pesquisas e práticas direcionadas para mudar a realidade brasileira atual.

## Agradecimentos

Este trabalho foi parcialmente financiado pela FAPERJ (projeto E-26-102.256/2013).

## Referências

Alcantara, W., Bandeira, J., Barbosa, A. and Lima, A. (2015). Desafios no uso de Dados Abertos Conectados na Educação Brasileira. *Anais do 4º DesafIE - Workshop de Desafios da Computação Aplicada à Educação*,

Bizer, C., Heath, T. and Berners-lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, v. 5, n. 3, p. 1–22.

Corsar, D., Edwards, P. and Nelson, J. (2014). Personal privacy and the web of linked data. *CEUR Workshop Proceedings*, v. 1121, p. 1–11.

D'Aquin, M. (2016). On the Use of Linked Open Data in Education: Current and Future Practices. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 9500, p. 153–165.

D'Aquin, M. (2012). Linked Data for Open and Distance Learning.

Dietze, S., Kaldoudi, E., Dovrolis, N., et al. (2013). Socio-semantic integration of educational resources - The case of the mEducator project. *Journal of Universal Computer Science*, v. 19, n. 11, p. 1543–1569.

Dietze, S., Sanchez-Alonso, S., Ebner, H., et al. (2013). Interlinking educational resources and the web of data: A survey of challenges and approaches. *Program*, v. 47, n. 1, p. 60–91.

[13] http://lov.okfn.org/dataset/lov/

[14] http://bioportal.bioontology.org/

[15] https://joinup.ec.europa.eu/catalogue/repository

[16] http://schema.org/docs/schemas.html

Hogan, A., Hitzler, P. and Janowicz, K. (2016). Linked Dataset Description Papers at the Semantic Web Journal : A Critical Assessment. v. 7, p. 1–3.

Jeremić, Z., Jovanović, J. and Gašević, D. (2013). Personal Learning Environments on the Social Semantic Web. *Semant. web*, v. 4, n. 1, p. 23–51.

Kapoor, S., Mojsilovi, A., Strattner, J. N. and Varshney, K. R. (2015). From Open Data Ecosystems to Systems of Innovation : A Journey to Realize the Promise of Open Data. *Bloomberg Data for Good Exchange Conference*,

Nguyen, K., Ichise, R. and Le, B. (2013). Interlinking Linked Data Sources Using a Domain-Independent System. In *Joint International Semantic Technology Conference*. Springer Berlin Heidelberg. http://ri-www.nii.ac.jp/SLINT/JIST2012.pdf.

Nikolov, A., D'Aquin, M. and Motta, E. (2012). Unsupervised learning of link discovery configuration. *The Semantic Web: Research and Applications*, v. 7295 LNCS, p. 119–133.

Nunes, B. P. (2014). Towards a well-interlinked Web through matching and interlinking approaches. Pontificia Universidade Católica do Rio de Janeiro.

Nunes, B. P., Dietze, S., Casanova, M. A., et al. (2013). Combining a co-occurrence-based and a semantic measure for entity linking. In *Extended Semantic Web Conference*. Springer Berlin Heidelberg.

Qing, H., Dietze, S., Giordano, D., et al. (2012). The Open University's repository of research publications Linked education : interlinking educational resources and the web of data. In *The 27th ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications*.

Siqueira, S., Bittencourt, I. I., Isotani, S. and Nunes, B. P. (2016). Sistemas de Informação baseados em Dados Abertos (Conectados). In *I GranDSI-BR – Grandes Desafios de Pesquisa em Sistemas de Informação no Brasil 2016 a 2026*.

Soru, T. and Ngonga Ngomo, A.C. (2014). A Comparison of Supervised Learning Classifiers for Link Discovery. In *Proceedings of the 10th international conference on semantic systems*. . Association for Computing Machinery.

Vasconcellos, S., Revoredo, K. and Bai, F. (2014). How Can Ontology Design Patterns Help Ontology Refinement ? v. 12, p. 4–16.

Wölger, S., Hofer, C., Siorpaes, K., et al. (2011). Interlinking data-approaches and tools. Technical report, STI Innsbruck, University of Innsbruck.

Zablith, F., D'Aquin, M., Brown, S. and Green-Hughes, L. (2011). Consuming Linked Data within a large educational organization. *CEUR Workshop Proceedings*, v. 782.