SISTEMA DE MÉTRICAS DE PUBLICAÇÕES ACADÊMICAS BASEADO NO JOURNAL ARTICLE TAG SUITE: O CASO DO PORTAL EDUC@ NA ÁREA DA EDUCAÇÃO

Metrics System for Academic Publications based on the Journal Article Tag Suite: the case of the Educ@ Portal in the education area
Sistema de Métricas para Publicaciones Académicas basado en Journal Article Tag Suite: el caso del Portal Educ@ en el área de educación

Ronnie Fagundes de Brito
Doutor em Eng. e Gestão do Conhecimento, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil.
Tecnologista do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil.
Lattes: http://lattes.cnpq.br/9015008667871372

Higor Alexandre Duarte Mascarenhas
Mestre em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil.
Professor do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Divinópolis, MG, Brasil.
Lattes: http://lattes.cnpq.br/8723936697065308

Bernardo Panerai Velloso
Bel. em Ciência da Computação, Universidade do Vale do Itajaí (UNIVALI), Itajaí, SC, Brasil.
Lattes: http://lattes.cnpq.br/6952994098627651

Gabriella Fernandes Rampinelli
Graduada em Letras pela Universidade de São Paulo (USP), São Paulo (SP), Brasil. Exerce atividades relacionadas a editoração e divulgação científica na Fundação Carlos Chagas (FCC), São Paulo (SP), Brasil.
Lattes: http://lattes.cnpq.br/4233547700367506

Nelson Gimenes
Mestre e Doutor em Psicologia da Educação pela Pontifícia Universidade Católica de São Paulo (PUC-SP), São Paulo, SP, Brasil. Pesquisador da Fundação Carlos Chagas e professor do mestrado profissional em Educação: Formação de Formadores da PUC-SP.
Lattes: http://lattes.cnpq.br/0869573191791125

Ronaldo Ferreira de Araujo
Doutor em Ciência da Informação, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil. Visiting Researcher at Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, Netherlands.
Lattes: http://lattes.cnpq.br/3328212638040851

Resumo

Introdução: Métricas sobre a produção acadêmica na forma de artigos permitem caracterizar e acompanhar a produção de autores e instituições, tendo sua importância destacada no acompanhamento de políticas editoriais. **Objetivos**: Com o objetivo de disponibilizar um portal de métricas na área de educação, foi desenvolvido um sistema de extração de dados. **Metodologia**: Esse sistema extrai dados a partir da representação XML dos artigos publicados no Portal Educ@, os quais complementa com dados de fontes externas em um modelo relacional, que em seguida são exportados para um índice e exibidos em uma ferramenta de visualização de informação. **Resultados**: Como resultado, foi disponibilizado um sistema de métricas aos editores das revistas do Portal, o qual possibilita o acompanhamento do desempenho de suas revistas. **Conclusão**: Verificou-se que os dados consolidados atendem as demandas dos editores, contudo se faz necessário um sistema de curadoria para normalização e padronização de registros visando métricas mais consistentes.

Palavras-chave: Métricas; Bibliometria; Educação; Portal.

Abstract

***Introduction**: Metrics on academic production in the form of articles allow characterizing and monitoring the production of authors and institutions, with their importance highlighted in monitoring editorial policies. **Objectives**: With the aim of making a metrics portal available in education, a data extraction system was developed from the XML representation of the articles published on the Educ@ Portal, which are complemented by data from external sources in a relational model and exported to an index and presented in a graphics presentation tool. **Results**: As a result, a metrics system was made available to the editors of the journals on the Portal, which permitted the monitoring of the performance of their journals. **Conclusion**: It was found that the consolidated data meet the demands of publishers, but a curation system is needed to normalize and standardize records for more consistent metrics.*

Keywords: *Metrics; Bibliometrics; Education; Portal.*

Resumen

Introducción: *Las métricas sobre producción académica en forma de artículos permiten caracterizar y monitorear la producción de autores e instituciones, destacando su importancia en el seguimiento de las políticas editoriales.* ***Objetivos***: *Con el objetivo de disponer de un portal de métricas en el área de educación, se desarrolló un sistema de extracción de datos a partir de la representación XML de los artículos publicados en el Portal Educ@, los cuales se complementan con datos de fuentes externas en un modelo relacional y exportado a un índice y desplegado en una herramienta de presentación de gráficos.* ***Resultados***: *Como resultado, se puso a disposición de los editores de las revistas en el Portal un sistema de métricas que les favorece monitorear el desempeño de sus revistas.* ***Conclusión***: *Se constató que los datos consolidados satisfacen las demandas de los editores, pero se necesita un sistema de curación que normalice y estandarice los registros para obtener métricas más coherentes.*

Palabras clave: *Métrica; Bibliometría; Educación; Portal.*

DOI: 10.22477/vi.widat.09

1. Introdução

A tomada de decisões e a definição de políticas no âmbito da editoração de revistas científicas demandam dados e informações que fundamentem direções a serem seguidas e ações a serem realizadas. Métricas e indicadores sobre produção acadêmica servem de base para o acompanhamento de políticas editoriais de periódicos, permitindo acompanhar características intrínsecas à publicação, como a afiliação institucional de autores, colaboração e internacionalização, assim como seu impacto em bases específicas (CUSCHIERI, 2022). As soluções que buscam sistematizar a produção e o impacto de determinada literatura científica costumam se utilizar de dois caminhos complementares na aferição de métricas: o desenvolvimento de ferramentas próprias de monitoramento e a agregação de serviços existentes (DESANTO; NICHOLS, 2017).

O presente trabalho descreve as atividades desenvolvidas no âmbito do projeto de construção de indicadores para o portal de periódicos Educ@, da Fundação Carlos Chagas,1 voltado a indexar revistas das áreas de educação e ensino. Apresentam-se a definição geral das métricas e indicadores potenciais, as tecnologias e linguagens de marcação de suporte aos registros, o modelo de arquitetura para o sistema de extração e apresentação dos dados de produção do portal, assim como reflexões sobre o processo de desenvolvimento e seus resultados.

1.1 Métricas e indicadores

Métricas e indicadores são essenciais à gestão editorial e profissionalização da editoração científica, sendo amplamente utilizados por bases indexadoras, gestores de portais de periódicos, editores, equipe editorial e pesquisadores do campo dos estudos métricos da comunicação científica. Os indicadores de produção, ligação e impacto são os mais utilizados em contextos de análise de produtividade e avaliação de desempenho de revistas e áreas de conhecimento (COSTA *et al.*, 2012).

Indicadores de produção científica são construídos pela contagem do número de publicações por tipo de documento (livros, artigos, publicações científicas, relatórios, etc.), por instituição, área de conhecimento, país, entre outros parâmetros.

Por sua vez, indicadores de ligação são construídos pelas coocorrências de autoria, citações e palavras, sendo aplicados na elaboração de mapas de estruturas de conhecimento e de redes de relacionamento entre pesquisadores, instituições e países. Empregam técnicas de análise estatística de agrupamentos.

1 http://educa.fcc.org.br/

Já os indicadores de impacto são construídos pela contagem dos registros de atividades que o item recebe. Essas atividades podem variar desde o acesso até o uso, ou seja, das visualizações (HTML, PDF) ao número de citações recebidas por uma publicação de artigo de periódico, por exemplo.

Tais indicadores devem ser combinados e utilizados de acordo com a demanda de avaliação e baseados em métricas que se pretende medir. O Quadro 1 apresenta exemplos das métricas modeladas para o Portal Educ@, os quais serviram para priorizar sua implementação, começando pelos indicadores de menor complexidade e maior prioridade.

Quadro 1 – Exemplos de métricas para o Portal Educ@

MÉTRICA	OBJETIVO	TIPO
Cobertura temática	Verificar os temas abordados na produção do portal ou revista	Produção
Quantidade de revistas	Acompanhar a evolução do Portal Educ@	Produção
Quantidade de participações em artigos por instituição	Identificar o volume de autores por instituição	Ligação
Índice de autocitação	Monitorar a endogenia da revista	Ligação
Menções em redes sociais	Monitorar o efeito da produção em redes sociais	Impacto
Citações em bases bibliográficas	Mensurar o impacto na produção bibliográfica/acadêmica em determinado conjunto de documentos	Impacto

Fonte: Elaborado pelos autores (2023).

O agrupamento das métricas em classes de produção, ligação e impacto orientou a estruturação do painel (*dashboard*) com essas informações.

Um recurso essencial para a elaboração das métricas do portal foram os arquivos dos artigos codificados em formato que facilitasse seu processamento computacional, o qual é descrito a seguir.

1.2 JATS (Journal Article Tag Suite)

O Journal Article Tag Suite (JATS) é um esquema de marcação em formato XML que permite estruturar a informação referente a um artigo científico de acordo com uma configuração predefinida e padronizada. No Brasil, o JATS foi adotado pela SciELO, que o estendeu e adaptou suas *tags* de acordo com demandas específicas, a partir de *tags* proprietárias (KIMURA; MACHUCA, 2014).

As revistas que fazem parte do Portal Educ@ vêm adotando em seus artigos esse conjunto de *tags*. Tal forma de representação torna possível o tratamento computacional dos textos dos artigos, pois, ao armazenar cada

elemento de informação como dado explicitado dentro de uma *tag*, viabiliza a extração e consolidação de dados específicos dos documentos publicados.

Como exemplos, podem-se citar as *tags* "*title-group*", que armazenam os títulos originais e traduzidos do artigo, e "*contrib-group*", que apresentam dados sobre os autores e suas respectivas afiliações institucionais. O conjunto de *tags* do JATS ofereceu a base para a estruturação dos dados referentes às métricas de produção do Educ@.

1.3 A produção de conhecimento em educação no Brasil e a Coleção Educ@

Em todas as áreas do conhecimento, os periódicos constituem um dos principais meios para a disseminação da produção científica. Na área da educação, embora caiba destacar que no Brasil ainda se verifica uma quantidade expressiva da produção intelectual publicada em livros e capítulos de livros (SOUSA; WERLE, 2014), nos últimos dez anos, segundo Souza *et al*. (2018), o número de periódicos avaliados pela Capes passou de 1.100 para 2.900, com aumento de mais de 160%, sendo a maioria deles nacionais. Além desse aumento, também devem ser destacadas as diversas mudanças na atividade editorial, sobretudo devido aos avanços tecnológicos, entre eles o surgimento de plataformas *on-line* de gestão editorial de manuscritos, a possibilidade de indexação em diferentes bases e coleções, as discussões sobre ciência aberta, a marcação XML dos artigos e o uso das mídias sociais para a divulgação científica.

A Coleção Educ@, implementada em 2010 por iniciativa da Fundação Carlos Chagas, é uma plataforma de acesso aberto que agrega 61 periódicos das áreas de educação e ensino, totalizando mais de 36 mil artigos, sendo quase ¾ deles pertencentes aos extratos A1 ou A2 do Qualis Periódicos do quadriênio 2017-2020. Ela tem como finalidade: 1) ampliar a divulgação da produção acadêmica por meio da admissão e permanência de periódicos científicos; 2) disponibilizar e disseminar *on-line*, em acesso aberto, artigos científicos que utilizam o procedimento de avaliação por pares; e 3) contribuir para o desenvolvimento de estudos por meio do aperfeiçoamento e da ampliação da capacidade de comunicação dos resultados veiculados por periódicos de qualidade, assim como subsidiar práticas investigativas em sistemas e redes de ensino.

2. Procedimentos Metodológicos

O sistema de métricas foi desenvolvido sob uma perspectiva de Datawarehouse (YULIANTO, 2019; AZIZ *et al*., 2012), segundo a qual os dados são extraídos dos arquivos JATS por meio do processamento do XML, transformados e complementados, e posteriormente carregados em um índice, de modo a facilitar seu acesso. O sistema para geração de dados-base para as métricas e indicadores do Educ@ foi implementado seguindo os passos de: extração de dados, transformação, inserção em modelo relacional, carga de dados em modelo dimensional, geração de índice e apresentação em *dashboard*.

De forma geral, os dados são extraídos por meio de *scripts* a partir dos XML referentes aos artigos e do próprio Portal Educ@. Esses *scripts* efetuam uma leitura de todas as *tags* dos XML, fazendo com que cada atributo seja alocado na sua devida tabela em um modelo relacional. Os *scripts* são executados em duas etapas: na primeira, são extraídos os dados relativos ao artigo em si, como, por exemplo, autores, instituição, ano de publicação, data de aceite, palavras-chave, número, volume, revista, entre outros, ignorando as referências. Posteriormente, o *script* analisa as referências, efetuando a extração de todas as suas *tags*, com dados como instituição de origem de uma referência, natureza (ex.: periódicos, anais de eventos, jornais, teses, entre outros), autores e ano. Logo essa referência também é inserida como um documento no banco de dados, porém com um atributo explicitando que corresponde a uma referência.

Depois da extração inicial os dados são transformados, principalmente por meio da normalização, para que estejam concisos e com mais alta confiabilidade. Para a normalização dos dados foram adotados a técnica de Distância *Levenshtein* e o cálculo da razão (YUJIAN; BO, 2007), o que permitiu padronizar as informações constantes nos arquivos XML. Por exemplo, em determinadas situações, o periódico *Eccos Revista Científica* era mencionado nas referências de diferentes formas, tais como "Eccos Rev. Cient"; "Eccos Revista Científica. São Paulo" e "EccoS Revista científica. Uninove". De forma análoga, as instituições mencionadas na afiliação dos autores também precisaram ser normalizadas. Outros *scripts* mais específicos de correção e normalização dos dados também são executados durante o processamento dos XML, a exemplo da remoção de caracteres especiais em alguns campos de texto ou sua lematização.

Após a normalização das informações, estas são complementadas no processo de transformação de dados por meio de consulta a bases externas, tais como Doaj, Altmetrics, Lattes, WikiData e CrossRef.

Seguinte à transformação/complementação, os dados são carregados em um modelo dimensional. Esse modelo permite criar consultas específicas na forma de visões no banco de dados. Contudo, o tempo de processamento demonstra-se relativamente elevado, prejudicando a interação do usuário com os gráficos e filtros do *dashboard*. Assim, a partir das visões, decidiu-se por realizar a carga em índices específicos, de modo a otimizar o tempo necessário para as consultas.

Como etapa final no fluxo de tratamento dos dados, ocorre a depuração, na qual se verifica a consistência entre a informação apresentada nos gráficos e tabelas e os dados existentes nas edições publicadas no Portal Educ@. Para a correção de dados inconsistentes foi desenvolvido um sistema de auditoria e curadoria dos dados, que permite rastrear a origem de determinada informação e eventualmente retificá-la.

3. Resultados

Com a aplicação dos métodos e técnicas apresentadas foi possível desenvolver um sistema de métricas para o Portal Educ@2.

Para apresentação das métricas e indicadores, foi adotada uma arquitetura de informação que permitiu a navegação por diferentes dimensões.

As informações disponibilizadas nesse sistema de métricas estão organizadas a partir de um conjunto de indicadores e métricas bibliográficas de modo a atender demandas da comunidade acadêmica no que se refere à avaliação e ao monitoramento da produção científica e à sua divulgação e disseminação nos diferentes âmbitos da sociedade, sobretudo entre pesquisadores, pós-graduandos, graduandos e demais profissionais da educação.

No sistema, o usuário pode acessar informações relativas ao conjunto total de periódicos, a uma parte dele ou a apenas um periódico, considerando os anos de publicação dos volumes de interesse. São apresentados indicadores que denominamos de produção científica, ligação, impacto e análise das referências.

No conjunto de indicadores de produção científica estão as informações relativas à contagem de artigos publicados, principalmente as que dizem respeito ao tempo médio de avaliação, distribuição das publicações por revista e por ano de publicação, idiomas publicados, países das instituições de vínculo dos autores e quantidade de participação de autores em artigos publicados na Coleção Educ@ segundo instituições brasileiras e estrangeiras.

O segundo conjunto de indicadores, denominado de ligação, corresponde àqueles que procuram apresentar, entre outras informações, coocorrências de autoria, citações e palavras, podendo ser utilizado para a elaboração de mapas de estruturas de conhecimento e de redes de relacionamento entre pesquisadores, instituições e países. Emprega técnicas de análise estatística de agrupamentos. No sistema de métricas da Coleção Edc@ estão presentes indicadores que retratam a colaboração científica entre países, instituições e autores de artigos publicados e indexados na Coleção.

Também são disponibilizados os indicadores de impacto, que têm como objetivo registrar a quantidade de atividades que o artigo recebe. Essas atividades compreendem desde o acesso até o uso dos documentos, incluindo *downloads*, citações e menções recebidas por uma publicação. Trata-se de uma ampla gama de métricas, das tradicionais às alternativas. Nesse sentido, o sistema dispõe de subagrupamentos, que agregam informações relativas: 1) às citações de artigos pertencentes à base em outros periódicos que também fazem parte da Coleção Educ@; 2) às citações de artigos da Coleção em outros artigos que possuam DOI e que estejam na base

2 Disponível em: http://educa-ibict.fcc.org.br:8088/superset/dashboard/12/.

da CrossRef; 3) às menções dos artigos da Coleção em fontes de métricas alternativas; 4) ao impacto em bases de políticas públicas; e 5) à quantidade de acessos aos artigos na Coleção Educ@ – considerando o número de visualizações das revistas, dos resumos, dos textos –, bem como ao número de *downloads*.

Além dessas, também foi possível propor uma nova categoria de indicadores, denominada análise das referências. Tem como base analisar referências bibliográficas citadas pelos artigos da Coleção Educ@. Assim, estão disponíveis informações sobre o tipo de documento referenciado (livros, capítulos, artigos, teses, legislação, relatórios, etc.), a média de idade dessas publicações e o volume de autocitação dos periódicos com artigos publicados na Coleção. Tais informações são de suma importância para o debate na área, uma vez que se verifica uma cultura instituída de uso predominante de referências oriundas de livros e capítulos de livros e com elevado tempo médio de publicação.

4. Considerações Finais

Uma questão que se apresentou relevante na elaboração das métricas foi, por exemplo, a definição de "artigo". Tal problemática surgiu em função da contagem de traduções de determinado documento para diferentes idiomas: discutiu-se se cada versão deveria ser contada uma ou mais vezes nos indicadores de produção. Ficou definido, para os fins deste trabalho, que o artigo consiste no documento original, elaborado pelo autor, sendo desconsideradas as versões traduzidas, exceto nas métricas voltadas à análise de internacionalização do conteúdo e de traduções.

Cabe notar que, apesar de garantir a estrutura do documento, a marcação XML não garante que a semântica esteja correta, ou seja, é possível que ocorram dados inconsistentes em seu conteúdo. Como exemplo, averiguou-se uma referência bibliográfica que corresponde a um artigo marcada como sendo parte de um livro, além de arquivos XML que mencionam data de publicação inválida (29 de fevereiro em um ano não bissexto, por exemplo).

A correção dos dados demonstrou-se mais complexa do que inicialmente estimado, inviabilizando a automatização dessa tarefa. Sendo assim, optou-se pelo desenvolvimento de um sistema de curadoria, que permite ao usuário gerenciar os dados extraídos pelo sistema, aplicar correções e resolver inconsistências.

O XML JATS se mostrou adequado para a extração de dados iniciais, contudo se fez necessária a complementação dessas informações por meios alternativos, como *web scraping*, normalização de dados e consultas a APIs externas.

Os métodos e técnicas desenvolvidos e adotados no projeto podem ser aplicados em outros portais ou bases de artigos, desde que adotem a codificação de artigos em JATS.

Referências

AZIZ, Azwa A.; WAHID, Abdul H. A.; HAMID, Nazirah A.; ROZAIMEE, A. Integration of heterogeneous databases in academic environment using open source ETL tools. **The International Conference on Informatics and Applications**, p. 433-439, 2012. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b61c5e30f128afb51d037d726b2d057c3f22ed75. Acesso em: 25 abr. 2023.

CUSCHIERI, Sarah. What are research metrics and why should I care? *In*: CUSCHIERI, Sarah (org.). **A roadmap to successful scientific publishing**: the dos, the don'ts and the must-knows. Cham: Springer International Publishing, 2022. p. 97-106.

COSTA, Teresa; LOPES, Sílvia; FERNÁNDEZ-LLIMÓS, Fernando; AMANTE, Maria João; LOPES, Pedro Faria. A bibliometria e a avaliação da produção científica: indicadores e ferramentas. **Atas Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas**, Lisboa, n. 11, 2012. Disponível em: https://comum.rcaap.pt/handle/10400.26/4620. Acesso em: 25 abr. 2023.

DESANTO, Dan; NICHOLS, Aaron. Scholarly metrics baseline: a survey of faculty knowledge, use, and opinion about scholarly metrics. **College & Research Libraries**, Chicago, v. 78, n. 2, p. 150-170, 2017.

KIMURA, Herbert; MACHUCA, Nadia Cristina de Araujo. O formato XML SciELO na RAC. **Revista de Administração Contemporânea**, Rio de Janeiro, v. 18, n. 5, set./out. 2014. Disponível em: https://arquivo.anpad.org.br/periodicos/arq_pdf/1_18_05_CartaLeitor.pdf. Acesso em: 13 jun. 2023.

SOUSA, Clarilza Prado de; WERLE, Flávia Obino Corrêa. Avaliação de livros na área de educação. **Revista Diálogo Educacional**, Curitiba, v. 14, n. 41, p. 289-308, jan./abr., 2014. https://doi.org/10.7213/dialogo.educ.14.041.dc01. Acesso em: 13 jun. 2023.

SOUZA, Ângelo Ricardo de *et al*. Qualis: a construção de um indicador para os periódicos na área da Educação. **Práxis Educativa**, Ponta Grossa, v. 13, n. 1, p. 219-231, jan./abr., 2018. https://doi.org/10.5212/PraxEduc.v.13i1.0013. Acesso em: 13 jun. 2023.

YUJIAN, Li; BO, Liu. A normalized Levenshtein distance metric. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 29, n. 6, p. 1091-1095, 2007.

YULIANTO, Ardhian Agung. Extract transform load (ETL) process in distributed database academic data warehouse. **APTIKOM Journal on Computer Science and Information Technologies**, v. 4, n. 2, p. 61-68, 2019.