convibra
www.convibra.org

ARCABOUÇO TECNOLÓGICO PARA ABERTURA DE DADOS DA PESQUISA EM CIÊNCIA DO SOLO NO BRASIL

Marcos Alexandre dos Anjos[(1)], Alessandro Samuel-Rosa[(2)], Matheus Ferreira Ramos[(3)]

[(1)] Bolsista de extensão tecnológica, discente do curso de bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná (UTFPR), Santa Helena, Paraná, Brasil. E-mail: marcosanjos@alunos.utfpr.edu.br.
[(2)] Docente dos cursos de bacharelado em Agronomia e Ciência da Computação da Universidade Tecnológica Federal do Paraná (UTFPR), Santa Helena, Paraná, Brasil. E-mail: alessandrorosa@utfpr.edu.br.
[(3)] Discente do curso de bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná (UTFPR), Santa Helena, Paraná, Brasil. E-mail: matheusramos@alunos.utfpr.edu.br.

Introdução

O solo ocupa posição central na complexa rede de relações das esferas terrestres. Seu uso sustentável é fundamental para assegurar uma vida saudável e promover o bem-estar de todas as formas de vida na Terra [1]. Essa importância é reconhecida na maioria dos 17 Objetivos do Desenvolvimento Sustentável (ODS) [2]. Por isso, a Food and Agriculture Organization (FAO) e a International Union of Soil Sciences (IUSS) lançaram a Global Soil Partnership (GSP), com ações junto aos segmentos da sociedade para conscientizar sobre a necessidade de juntar esforços para preservar o solo.

As ações da GSP são organizadas em pilares. O foco do pilar IV é aumentar a quantidade e qualidade dos dados e informações do solo. Para isso, muitos países estão desenvolvendo sistemas de informação de solos que se comunicarão com um sistema global de informação (GLOSIS). O Brasil é referência em dados e informações do solo em regiões tropicais [3], ainda não possui um sistema de informação de solos. Grande parte dos dados produzidos no Brasil é difícil de encontrar e/ou acessar e, portanto, difícil de reutilizar. Quando acessíveis, costumam estar incompletos ou organizados de maneira inapropriada [4]. Além de impedir a replicação das pesquisas, isso resulta na subutilização de recursos públicos, atrasa o avanço do conhecimento sobre o solo e impede o alcance dos ODS.

Diversos esforços institucionais ou individuais tentaram resolver o cenário descrito acima. Contudo, uma solução duradoura para o problema de salvaguardar dados da pesquisa em ciência do solo e promover seu reúso não foi concretizada. Diante disso, em 2016, foi criado o Repositório Brasileiro Livre para Dados Abertos do Solo (FEBR). Seu diferencial foi a adoção de métodos baseados em experiências internacionais, uma política de dados abertos e a seleção de tecnologias de fácil acesso, manutenção e uso. Hoje o FEBR é o maior repositório de dados da pesquisa em ciência do solo do Brasil, com dados de mais de 20 mil locais em todo o país. Por isso, o FEBR precisa estar alinhado aos desenvolvimentos tecnológicos internacionais da área.

O objetivo deste trabalho é definir, a partir de uma profunda revisão bibliográfica e estudo de repositórios nacionais e internacionais de dados da pesquisa, o arcabouço tecnológico necessário para que o FEBR atinja o nível de confiabilidade requerido de repositórios de dados da pesquisa.

Repositório de Dados Abertos da Pesquisa: Princípios e Fundamentos

A chegada da *World Wide Web* alterou a forma de obter e compartilhar as informações, e permitiu a redução das barreiras do acesso à informação [5]. Atualmente, a facilidade em compartilhar informações através das redes de computadores trouxe um aumento significativo no volume de dados. Os dados podem ser categorizados, por exemplo, em dados abertos, que de acordo com definição da *World Wide Web Consortium* (W3C) " [...] pode ser acessado por qualquer pessoa que for livre para usá-lo, modificá-lo e distribuí-lo" [5].

Essa definição implica em alguns pontos essenciais, as quais foram estabelecidas através de três leis: (1) dados não disponíveis e não indexado na Web não existem, (2) dado aberto em formato não acessível por máquina não pode ser reaproveitado, e (3) se algum dispositivo legal não permitir sua replicação, o dado não é útil [6]. As três leis foram propostas para dados abertos governamentais, porém suas definições podem ser aplicadas aos dados abertos em forma geral.

Governos e comunidades nacionais passaram a estudar maneiras de visualizar o valor sobre os dados compartilhados. Dessa forma, o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) vem promovendo ações para o avanço da ciência por meio do Manifesto de Acesso Aberto a dados da pesquisa brasileira.

Em meados de 2014 surgiram as primeiras manifestações aos princípios FAIR que significa *Findable, Accessible, Interoperable e Reusable.* FAIR aderiu uma solução para construção de *backbone*, ou seja uma rede de comunicação para uso e distribuição dos dados pela *World Wide Web* através da rede de computadores [6]. Assim, relacionando estes princípios do FAIR é possível garantir a interoperabilidade entre repositórios unido a um conjunto de metadados definidos para uso tanto por buscadores de metadados quanto por pessoas. O FAIR está presente na ciência aberta e vai ao encontro com o movimento de abertura dos dados. Isso quer dizer que os dados estão disponíveis para uso e reuso onde que está atrelado a uma licença de uso, créditos e citações relacionadas.

Para garantir o acesso e confiabilidade dos conjuntos de dados abertos, é necessário que esses dados estejam em um lugar seguro, acessível e disponível. Repositórios de dados são considerados uma boa prática para a ciência, compartilhando métodos de acesso às informações [7]. Estes repositórios utilizam softwares para gerenciamento de metadados seguindo as normas e padrões internacionais de interoperabilidade. Para avaliação de confiabilidade de um repositório existe o modelo de auditoria *DSA-WDS Core Trustworthy Data Repositories Requirements (TRUST),* constituído por 16 requisitos, podendo o repositório obter um certificado nível básico caso consiga cumprir todos os requisitos [10]. De maneira geral existem processos de auditoria que consiste em avaliar as características do repositório entre elas verificando a confiabilidade do repositório a partir de uma certificação dando credibilidade ao repositório.

Esquema de Metadados

Diante desta realidade, para disponibilização da informação, torna-se necessário a utilização de padrões melhor descritos naquela informação. Os esquemas de metadados para domínio na Web representam as características e atributos de um objeto real com intuito de identificar um modelo para posteriormente realizar a recuperação [9]. Os sistemas de gerenciamento de metadados evoluem a partir desses esquemas de metadados. A representação esquema metadados é codificada com linguagem de marcação *eXtensible Markup Language* (XML). Uso de esquemas padronizados de metadados ajuda a definir que categorias de informações registrar no banco de dados e como estruturar essas informações[10]. Dessa forma, conhecer os esquemas de metadados e a forma como são a forma como são manipulados torna-se essencial para saber quais modelos utilizar.

A interoperabilidade está relacionada diretamente com a escolha dos metadados que irão identificar a descrição dos recursos em um sistema. Modelos de esquemas para metadados que melhor representam o modelo do conjunto de dados são DataCite e OpenAIRE. O modelo de representação dos metadados com DataCite contém 18 atributos que variam em obrigatórias, recomendadas e opcionais, e oferece suporte aos pesquisadores para citar dados de outras pesquisas [11]. O esquema OpenAIRE apresenta os mesmos objetivos que o DataCite, fornecendo um esquema de metadados e garantindo a interoperabilidade [12]. O principal destaque do OpenAIRE é a aceitação de outros esquemas de identificador único, e não apenas o DOI. Neste contexto, o IBICT em meados de 2020 adotou o esquema de metadados OpenAIRE na implementação do repositório para dados da pesquisa e publicação.

Identificador Regular Persistente

O modelo de padrão dos metadados necessita apresentar um identificador único para suprir as necessidades dos protocolos de comunicação [13]. Dessa forma, cada conjunto de dados no repositório deve conter um identificador único usado para requisição dos métodos de uma informação. A padronização dos metadados devem ser persistentes e identificáveis, sendo assim, um item importante dos princípios do FAIR.

Para transmissão de dados devemos seguir alguns protocolos como *Hypertext Transfer Protocol* (HTTP), *Uniform Resource Locator* (URL), *Uniform Resource Identifier* (URI). O desenvolvimento de aplicações na plataforma Web implica na adoção de tecnologias padronizadas. HTTP é o protocolo de comunicação que implica em obter a semântica e infraestrutura de desenvolvimento [14]. URL se refere ao local, *host* (ponto de acesso) quer acessar determinado recurso [15]. A URL *https://www.pedometria.org,* possibilita a execução de duas operações HTTP sobre o recurso GET e POST, ou seja, pode receber e enviar requisições. URI como o próprio nome sugere, se refere ao identificador do recurso [16]. De maneira geral, todos os recursos como imagens e páginas estão atrelados ao identificador único. A URL *https://www.pedometria.org/dadosX/{ID}* é um exemplo de URI template, que possibilita a construção de URLs com identificador único (ID) variáveis.

Protocolos e Interoperabilidade

Os repositórios vêm ganhando destaque em relação ao armazenamento e preservação dos dados. Dessa forma, os protocolos de comunicação têm como objetivo facilitar o acesso e reuso de dados [17]. Existem diversas ferramentas e protocolos de comunicação de interoperabilidade de registros de metadados, divididos em provedores de dados e provedores de serviços. Os protocolos para provedores de dados apresentam o mesmo comportamento dos buscadores de metadados, os quais realizam a coleta para adicionar na sua base de dados. Para softwares que usam conceito de provedores de serviços, faz-se uso de protocolos para entregar os metadados aos buscadores de metadados. Para garantir a interoperabilidade entre estes softwares utilizam-se protocolos de comunicação.

Podemos entender que interoperabilidade é a capacidade de sistemas da tecnologia da informação (TIC) e todos seus processos se comunicarem, auxiliando a troca de informações [18]. *Open Archives Initiative* (OAI) um protocolo de comunicação que surgiu a partir de uma necessidade da comunicação entre bases de repositórios estabelece um modelo padrão de comunicação via redes de computadores [19]. Em meados de 2001 foi publicada a primeira versão (1.0) do *Open Archives Initiative - Protocol for Metadata Harvesting* (OAI - PMH). No ano seguinte, o protocolo foi revisado e lançado na versão 2.0 com melhorias na interoperabilidade [20]. Atualmente o protocolo está em estudos mas sem grandes mudanças, permanecendo na versão 2.0.

Conclusões

Este estudo demonstrou que existem soluções tecnológicas que atendem às demandas do FEBR como, por exemplo, o DataVerse. Contudo, elas possuem demandas de instalação ou manutenção elevadas, exigindo uma equipe dedicada a essa finalidade. Assim, se conclui que a estratégia mais viável para o FEBR seja o desenvolvimento de uma solução própria selecionando e instalando os componentes estritamente necessários. Isso inclui (1) o esquema de metadados DataCite ou OpenAIRE, (2) o protocolo OAI-PMH para expor os metadados e (3) um identificador uniforme de objetos digitais baseado em URL/URI.

Bibliografia

[1] L. Montanarella *et al.*, “World’s soils are under threat”, *SOIL*, vol. 2, nº 1, p. 79–82, fev. 2016, doi: 10.5194/soil-2-79-2016.

[2] S. D. Keesstra *et al.*, “The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals”, *SOIL*, vol. 2, nº 2, p. 111–128, abr. 2016, doi: 10.5194/soil-2-111-2016. [Online]. Disponível em: https://soil.copernicus.org/articles/2/111/2016/. [Acessado: 20-nov-2020]

[3] F. A. O. Camargo, V. H. Alvarez, e P. C. Baveye, “Brazilian soil science: from its inception to the future, and beyond”, *Rev. Bras. Ciênc. Solo*, vol. 34, p. 589–599, 2010, doi: 10.1590/S0100-06832010000300001.

[4] A. Samuel-Rosa, R. S. D. Dalmolin, J. M. Moura-Bueno, W. G. Teixeira, e J. M. F. Alba, “Open legacy soil survey data in Brazil: geospatial data quality and how to improve it”, *Sci. Agric.*, vol. 77, nº 1, 2020, doi: 10.1590/1678-992x-2017-0430.

[5] C. BIZER, T. HEATH, e T. BERNERS-LEE, “Linked data – the story so far. International journal on semantic web and information systems”, vol. 5, p. 3, jan. 2009, doi:

10.4018/jswis.2009081901.
[6] DATA FAIRPORT, “Jointly designing a data FAIRPORT”, 2014 [Online]. Disponível em: https://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602
[7] S. M. de S. Costa e F. C. L. Leite, *Insumos conceituais e práticos para iniciativas de repositórios institucionais de acesso aberto à informação científica em bibliotecas de pesquisa*. EDUFBA, 2010 [Online]. Disponível em: https://repositorio.unb.br/handle/10482/5470. [Acessado: 25-mar-2021]
[8] “CoreTrustSeal - E-LIS repository”. [Online]. Disponível em: http://eprints.rclis.org/34431/. [Acessado: 25-mar-2021]
[9] R. C. V. Alves, “Metadados como elementos do processo de catalogação”, p. 134.
[10] E. MÉNDEZ RODRÍGUEZ, “Metadados y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales”, p. 429, 2002.
[11] “DataCite”. [Online]. Disponível em: https://datacite.org/. [Acessado: 12-abr-2021]
[12] “Diretrizes OpenAIRE para repositórios de dados.pdf”. [Online]. Disponível em: https://livroaberto.ibict.br/bitstream/123456789/1086/2/Diretrizes%20OpenAIRE%20para%20reposit%C3%B3rios%20de%20dados.pdf. [Acessado: 12-abr-2021]
[13] “Open Archives Forum. Implementing oai-pmh”. [Online]. Disponível em: http://www.oaforum.org/tutorial/english/page4.htm. [Acessado: 20-dez-2020]
[14] J. WEBBER, S. PARASTATIDIS, e I. ROBINSON, *REST in Practice: Hypermedia and Systems Architecture*. .
[15] I. L. Salvadori, “UNIVERSIDADE FEDERAL DE SANTA CATARINA DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA”, p. 158, 2015.
[16] A. F. D. Santos, “CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO”, p. 61, 2013.
[17] C. H. Marcondes e L. F. Sayão, “Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira”, *Ciênc. Informação*, vol. 30, nº 3, 2001 [Online]. Disponível em: http://revista.ibict.br/ciinf/article/view/909. [Acessado: 29-mar-2021]
[18] C. J. A. de Araújo, “Um modelo para interoperabilidade entre instituições heterogêneas”, text, Universidade de São Paulo, 2012 [Online]. Disponível em: http://www.teses.usp.br/teses/disponiveis/45/45134/tde-08022013-111002/. [Acessado: 29-mar-2021]
[19] p GARCIA e M. SUNYE, “O protocolo OAI-PMH para interoperabilidade em bibliotecas digitais. In: CONGRESSO DE TECNOLOGIA PARA GESTÃO DE DADOS E METADADOS DO CONE SUL”, vol. 1, 2003.
[20] “Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): descripción, funciones y aplicación de un protocolo - E-LIS repository”. [Online]. Disponível em: http://eprints.rclis.org/4093/. [Acessado: 29-mar-2021]