ORGANIZAÇÃO DA INFORMAÇÃO NOS REPOSITÓRIOS DE DADOS DE PESQUISA EM PORTUGAL

ORGANIZATION OF INFORMATION IN RESEARCH DATA REPOSITORIES IN PORTUGAL

Elis Gabriela Copa dos Santos

https://doi.org/10.21747/21836671/pagnesppk5

Resumo: Num momento em que se discute tão cotidianamente as questões relacionados ao acesso aberto à informação científica, entende-se válido compreender melhor os repositórios de dados científicos e como estes estão organizados. Dessa maneira, este estudo realizou um mapeamento dos repositórios de dados de pesquisa atualmente ativos em Portugal, com a finalidade de analisar sua organização. Foram consideradas questões sobre a temática do repositório, o tipo de conteúdo armazenado, os metadados utilizados na descrição dos documentos e sob quais coleções/categorias o conteúdo está agregado. Por meio da consulta aos portais re3Data e OpenDoar foram identificados 67 repositórios, dos quais sete foram selecionados para o estudo: apenas repositórios de acesso aberto que abrigassem dados provenientes de pesquisa científica, mantidos por instituições portuguesas. A pouca quantidade de repositórios, bem como de itens armazenados nos repositórios analisados, aponta para o fato de que o estado da arte dos repositórios de dados de pesquisa em Portugal ainda se encontra em estágio inicial.

Palavras-chave: Acesso aberto; Dados de pesquisa; Organização da informação; Repositórios de dados de pesquisa.

Abstract: At a time when issues related to Open Access to scientific information are daily discussed, it is considered valid to better understand research data repositories and how they are organized. Thus, this study carried out a mapping of the research data repositories currently active in Portugal, in order to analyze its organization. Were considered issues about the theme of the repository, the type of content stored, the metadata used in documents' description and under which collections/categories the content is aggregated. Through the consultation of the re3Data and OpenDoar portals, 67 repositories were identified, of which seven were selected for the study: only open access repositories that contained data from scientific research, maintained by Portuguese institutions. The small number of repositories, as well as items stored in the analyzed repositories, points to the fact that the state of the art of research data repositories in Portugal is still at an early stage.

Keywords: Open access; Research data; Information organization; Research data repositories.

Introdução

Uma das questões que tem se potencializado graças às possibilidades de compartilhamento e colaboração permitidas pela introdução de tecnologias digitais nos procedimentos científicos é a do Acesso Aberto (AA) à produção científica. São inúmeras as discussões e os pontos de vista sobre a questão. Fecher e Friesike (2014) sugerem que a abertura da ciência está apoiada em cinco escolas de pensamento complementar. Entre elas, a Escola Democrática, aponta a forma desigual como o acesso ao conhecimento está distribuído e propõe que esta desigualdade seja resolvida pelo acesso livre às publicações científicas e aos dados de pesquisa.

Os dados de pesquisa, ou dados de investigação, podem ser entendidos como registros factuais que foram coletados e/ou gerados no processo de pesquisa e usados como fontes primárias para a produção científica com a finalidade de validar os resultados da pesquisa (OECD, 2007). Estes dados podem ser textos, pontuações e escalas numéricas, coordenadas geográficas, imagens, áudios e sons, códigos-fonte...

> O registo das observações, ensaios e experiências, ou seja, a produção de dados, é já há vários séculos uma das características essenciais da ciência moderna. A forma e o volume desses registos ou dados científicos foram naturalmente evoluindo, crescendo em dimensão e complexidade, de acordo com a própria evolução da investigação científica, dos seus objectos, metodologias e instrumentos. De igual modo, foram-se registando alterações nas formas de armazenar, preservar, aceder e partilhar os dados produzidos no âmbito da actividade científica. (RODRIGUES *et al.*,2010).

A solução mais discutida atualmente no que diz respeito ao compartilhamento destes dados é a criação dos Repositórios de Dados de Pesquisa (RDP). Enquanto as bibliotecas preservam e disseminam um conhecimento já reconhecido e publicado, cabe aos repositórios organizar a nova produção intelectual da instituição ou da área ao qual está ligado. Assim, os RDP armazenam e preservam os dados científicos que compõem a base do conhecimento primário gerado pelas investigações. Nem sempre existem como repositórios independentes, em alguns casos estão integrados ou embutidos em outros repositórios. Conforme apontam Rousidis *et al.* (2014), “dado ao grande volume de e a diversidade dos dados científicos, repositórios de pesquisa estão se tornando uma parte integral do processo de comunicação e de colaboração entre pesquisadores e grupos de pesquisa”.

Este trabalho insere-se no contexto da Organização da Informação, que, conforme aponta Simões (2018) está contida no núcleo da Ciência da Informação e é fundamentada por processos como análise e representação, estabelecendo a mediação entre a informação produzida e o seu utilizador, com vista à recuperação de objetos informacionais. Deve permitir identificar a existência de todos os tipos de recursos informacionais relacionados e suscetíveis de preencher a necessidade de informação de quem a busca. Neste contexto, este estudo realizou um mapeamento e posterior análise do conteúdo, estrutura e organização dos RDP em Portugal.

Procedimentos metodológicos

Foram identificados os RDP mantidos por universidades e/ou instituições de pesquisa sediadas em Portugal por meio da pesquisa no modo de “Busca por país” nos portais agregadores: *re3data* e *OpenDOAR*. Para cada repositório identificado, foi criada uma ficha de caracterização com os seguintes campos: nome, instituição responsável, tema ou área temática e tipo de conteúdo. Após a identificação, procedeu-se a seleção dos repositórios que abrigam dados resultantes de pesquisa científica. Para os repositórios selecionados a ficha de caracterização foi expandida com os campos: “Atribui metadados aos arquivos?”, “O conteúdo está agregado sob coleções/categorias?” e “Tipos de dados armazenados”. Os resultados da análise dos repositórios selecionados estão apresentados a seguir.

Resultados

Em consonância com a comunidade acadêmica mundial, desde o início dos anos 2000 as universidades e instituições de pesquisa em Portugal têm se preocupado com as questões do acesso aberto à produção científica. Em 2008 foi criada a iniciativa nacional de acesso aberto – os Repositórios Científicos de Acesso Aberto de Portugal (RCAAP). Inicialmente o RCAAP focou-se na implantação de Repositórios Institucionais destinados ao armazenamento da produção intelectual das universidades e instituições de pesquisa de Portugal e partir de 2010 passou haver maior preocupação no domínio do acesso e curadoria dos dados resultantes de investigação e dos RDP (RODRIGUES *et al.*, 2010).

Os dados coletados por esta investigação confirmam tal situação, uma vez que foram identificados 67 repositórios mantidos por universidades e/ou instituições de pesquisa portuguesas, dos quais 50 de caráter institucional. Conforme os critérios adotados, foram selecionados para análise sete RDP, apresentados no Quadro 1.

Quadro 1 – Ficha de caracterização dos repositórios selecionados para a pesquisa

Nome do repositório	Instituição responsável	Tema/área
Antimicrobial Combination Networks (ACN)	Universidade do Minho	Biologia
DataRepositoriUM (DRUM)	Universidade do Minho	Geral
The Integron Database (INTEGRALL)	Universidade de Aveiro	Biologia
Kinetic Models of biological Systems (KiMoSys)	Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa	Biologia
Portulan Clarin Repository (PC)	Universidade de Lisboa/ Universidade de Évora	Ciências e tecnologias da linguagem
Repositório Dados Científicos (RDC)	RCAAP/FCT	Geral
Repositório de Dados Científicos do Instituto Politécnico de Castelo Branco[1](RDC-IPCB)	Instituto Politécnico de Castelo Branco	Geral

Fonte: Autoria própria, dados da pesquisa.

O primeiro fator que se entende necessário apontar é que embora sejam mantidos por instituições portuguesas, dos sete repositórios selecionados, quatro têm seus títulos, interface e grande parte do seu conteúdo em língua inglesa. São eles: ACN, INTEGRALL, KiMoSys e PC. São estes os mesmos quatro os repositórios cujo conteúdo é temático, ou seja, ligado a uma área do conhecimento. Os repositórios em língua portuguesa são os

[1] O RDC-IPCB não é um repositório independente, mas sim uma coleção dentro do repositório institucional do Instituto Politécnico de Castelo Branco. Está sendo considerado nesta pesquisa devido à quantidade de materiais depositados (1.972 registros na data de coleta de dados da pesquisa).

repositórios que abrigam os dados de pesquisa no âmbito de uma instituição específica, daí perceber-se uma grande similaridade com os repositórios institucionais.

Pressupunha-se que todos os repositórios tivessem uma certa conformidade em sua organização, condizente com a organização observada em outros tipos de repositórios digitais. Contudo, diferentemente de modelos mais tradicionais, onde estão armazenados documentos estáticos com arquivos para download ou visualização *online*, o ACN, o INTEGRALL e o KiMoSys são interfaces abastecidas por dados que mostram seus resultados a partir da interação do usuário com o sistema, gerando visualizações ou fórmulas. Já os demais (DRUM, RDC, RDC-IPCB e PC) seguem um padrão mais tradicional no qual estão abrigados arquivos (textos, imagens, áudios, códigos-fonte, arquivos tabulares (tabelas, quadros e planilhas), bases de dados...) que necessitam ser descarregados para ser utilizados.

Nos repositórios institucionais, costuma-se encontrar o conteúdo organizado sob coleções, geralmente seguindo a mesma organização da instituição, de acordo com os departamentos, faculdades e/institutos que as compõem. Esse mesmo tipo de organização foi observada no DRUM, no RDC-IPCB e no RDC, que são justamente os únicos repositórios que abrigam dados provenientes de pesquisas realizadas na instituição como um todo. Os demais repositórios, considerados temáticos, ou seja, cujo conteúdo é relativo a um tema ou área do conhecimento específica, não se organizam em coleções, mas sim em função do conteúdo. O ACN organiza seu conteúdo em função dos organismos sobre os quais armazena dados, o KiMoSys apresenta seu conteúdo de acordo com os organismos biológicos registrados e o INTEGRALL de acordo com sequências e arranjos genéticos.

Pode-se apontar que nos repositórios que seguem o padrão mais tradicional de organização, percebe-se também maior similaridade no modo de busca do conteúdo. No DRUM, no RDC e no RDC-IPCB é possível pesquisar por meio da caixa de busca presente na página inicial ou percorrendo o repositório por Comunidades e Coleções, Data de publicação, Autor, Título, Assunto, Tipo de Documento e Tipo de Acesso. No PC, a busca pode ser feita pela barra de busca textual ou por meio de uma lista com os materiais disponíveis. É necessário salientar que nem todos os registros possuem um item para *downloads*. Há registros que contém apenas a descrição do conteúdo e no lugar do botão de *download* está um atalho para contato com o detentor do conteúdo. No ACN a pesquisa, realizada em caixas de seleção, pode ser feita por organismo, agente microbiano, combinação microbiana, interação bem como pela combinação de filtros. O INTEGRALL permite a busca por meio de uma lista com os códigos identificadores das cadeias genéticas. No KiMoSys a busca pode ser feita por meio de consulta a uma lista apresentada com os organismos descritos ou por meio de uma caixa de pesquisa na qual o usuário pode inserir um termo ou um conjunto de termos. Ao encontrar o registro buscado, o usuário pode realizar o download dos artigos, arquivos e dados e arquivos de modelo para cada resultado.

Do ponto de vista da descrição dos conteúdos, há em todos os repositórios metadados que identificam o conteúdo. Nos repositórios que seguem o padrão mais tradicional, apresenta-se um registro e um arquivo para *download*. Nos registros observam-se informações como título, autoria ou responsabilidade, data e local de criação ou de coleta, formato, tamanho, permissões e condições de uso e requisitos de sistema necessários para o uso. No ACN, no INTEGRALL e no KiMoSys, embora a visualização do conteúdo seja diferente, para todos os organismos, agentes ou cadeias genéticas estão indicadas informações que os identificam e individualizam. Em todos os repositórios está indicada a importância da necessidade de citação do conteúdo ali disponibilizado.

Considerações finais

Um dos maiores obstáculos que se pode indicar com relação à implantação dos RDP é dificuldade ainda bastante persistente de identificar o que são dados de pesquisa. Os conceitos variam significativamente entre as áreas do conhecimento e, em muitos casos, até mesmo dentro de uma área específica há falta de consenso. Essa dinamicidade se reflete também nos RDP, que podem possuir diferentes formas de organização, estrutura e visualização, de acordo com as necessidades dos dados abrigados. Isto faz com que o cenário dos RDP seja muito diversificado, podendo imputar ao pesquisador dificuldades no momento de selecionar onde disponibilizar seus dados.

Os resultados dessa investigação permitem afirmar que o estado da arte dos RDP em Portugal ainda se encontra em estágio inicial, pois a baixa quantidade de repositórios e de itens armazenados nos repositórios analisados dá indícios de que os mesmos ainda tenham pouca visibilidade dentro de suas instituições. Contudo, a presença de conjuntos de dados nas coleções de alguns repositórios institucionais indica que há preocupação por parte de pesquisadores em disponibilizar seus dados. Portanto, com esforços de sensibilização para a importância da disponibilização de dados de pesquisa, bem como a implantação de mais RDP pode-se alcançar um ambiente científico mais aberto e transparente.

Referências bibliográficas

FECHER B.; FRIESIKE S.
2014 Open Science: one term, five schools of thought. In BARTLING, S.; FRIESIKE, S., ed. *Opening Science : the evolving guide on how the Internet is changing research, collaboration and scholarly publishing*. [Em linha]. Cham [etc.] : Springer Open, 2014. Disponível em: https://doi.org/10.1007/978-3-319-00026-8_2.

OECD
2004 *Declaration on access to research data from public funding*. [Em linha]. 2004. Disponível em: https://legalinstruments.oecd.org/en/instruments/157.

RODRIGUES, E. [et al.]
2010 *Os Repositórios de dados científicos: estado da arte*. [Em linha]. 2010. Disponível em: https://repositorio-aberto.up.pt/bitstream/10216/23806/2/44632.pdf.

ROUSIDIS, D. [et al.]
2014 Metadata for big data: a preliminary investigation of metadata quality issues in research data repositories. *Information Services & Use*. [Em linha]. 34:3/4 (2014) 279-286. Disponível em: https://doi.org/10.3233/ISU-140746.

SIMÕES, M. G. M.
2018 *Organização e gestão do conhecimento: [aula do Doutoramento em Ciência da Informação]*. Coimbra, 2018.

Elis Gabriela Copa dos Santos | elis.santos@student.fl.uc.pt
Universidade de Coimbra - Faculdade de Letras, Portugal