XXI ENANCIB
Encontro Nacional de Pesquisa em Ciência da Informação
50 anos de Ciência da Informação no Brasil: diversidade, saberes e transformação social
Rio de Janeiro • 25 a 29 de outubro de 2021

XXI Encontro Nacional de Pesquisa em Ciência da Informação – XXI ENANCIB

GT-8 – Informação e Tecnologia

ANÁLISE AUTOMATIZADA DO AUTOARQUIVAMENTO NA CIÊNCIA DA INFORMAÇÃO: UM MÉTODO REPRODUTÍVEL

AUTOMATED ANALYSIS OF SELF-ARCHIVING IN INFORMATION SCIENCE: A REPRODUCIBLE METHOD

Sarah Rúbia de Oliveira Santos - Universidade Federal de Minas Gerais (UFMG)
Dalgiza Andrade Oliveira - Universidade Federal de Minas Gerais (UFMG)

Modalidade: Trabalho Completo

Resumo: A abertura da ciência tem se tornado cada vez mais desejável. Desde o acesso aberto às publicações científicas até a abertura dos dados e do fluxo de trabalho dos pesquisadores, busca-se a transparência para que as pesquisas se tornem acessíveis e replicáveis. A presente pesquisa apresenta uma proposta de método reprodutível para análise do autoarquivamento na Ciência da Informação, a partir da produção científica disponível no repositório *Eprints in Library and Information Science*. O estudo é aplicado, descritivo e de caráter exploratório. Aplica a linguagem de programação Python, o ambiente computacional Jupyter Notebook e a plataforma de Ciência de Dados Anaconda para a criação dos instrumentos de coleta e análise de dados. Tem como resultado a criação de um *script* para coleta e análise de dados do repositório. Com ele foi possível extrair dados de páginas *web*, organizar, analisar e visualizar os dados obtidos. A partir dos 125 conjuntos de dados coletados com informações de 23.245 publicações, foram gerados 57 gráficos e 40 tabelas sobre as categorias de análise: *types, subjects, date, datestamp, linguabib, publication, conference* e *keywords*. O autoarquivamento foi realizado entre 2002 e 2020, com publicações feitas entre 1965 e 2020. Os artigos de periódicos, artigos de conferências e apresentações são as tipologias mais depositadas e suas temáticas refletem as tendências e perspectivas da área. A partir desse estudo, observa-se que é possível pensar maneiras de tornar o percurso metodológico de pesquisas nas ciências sociais mais aberto e transparente.

Palavras-Chave: Ciência Aberta; Pesquisa reprodutível; Autoarquivamento; Bibliometria.

Abstract: *The opening of science has become increasingly desirable. From open access to scientific publications to the opening of data and researchers' workflow, transparency is sought so that research becomes accessible and replicable. This research presents a proposal for a reproducible method for the analysis of self-archiving in Information Science, based on the scientific production available in the Eprints in Library and Information Science repository. The study is applied, descriptive and exploratory. It uses the Python programming language, the Jupyter Notebook computational environment and the Anaconda Data Science platform for the creation of data collection and analysis instruments. The result is the creation of an script for collecting and analyzing data from the repository. With that it was possible to extract data from web pages, organize, analyze and visualize the data obtained. From the 125 datasets collected with information from 23,245 publications, 57 graphs and 40 tables were generated on the categories of analysis: types, subjects, date, datestamp, linguabib, publication, conference and keywords. The self-archiving was carried out between 2002 and 2020, with publications made between 1965 and 2020. Journal articles, conference articles and presentations are the most common types of documents and their themes reflect the trends and perspectives of the area. From this study, it was observed that it is possible to think of ways to make the methodological path of research in the social sciences more open and transparent.*

Keywords: *Open Science; Reproducible research; Self-archiving; Bibliometrics.*

1 INTRODUÇÃO

Em meio às iniciativas abertas que surgiram no início dos anos 2000, seguindo os postulados do Movimento de Acesso Aberto e da Iniciativa de Arquivos Abertos, o repositório *EPrints in Library and Information Science* (E-LIS)[1] foi criado. O E-LIS é um repositório digital para documentos relacionados à Biblioteconomia, Ciência da Informação e áreas correlatas, que está hospedado pelo Sistema de Biblioteca (CAB) da Universidade de Nápoles Federico II, na Itália. Esse repositório foi estabelecido, é gerenciado e mantido por uma equipe voluntária de bibliotecários e cientistas da informação pertencentes a cerca de 45 países (E-LIS, 2020).

Desde a sua criação e com o crescimento do acervo, alguns trabalhos analisando a produção científica disponível no repositório foram desenvolvidos, no entanto, os percursos metodológicos adotados, em especial, as formas de coleta e análise de dados utilizadas pelos pesquisadores, não foram abertos ou transparentes o suficiente para que esses resultados fossem reproduzidos. O que constitui uma preocupação que perpassa movimentos dentro da Ciência Aberta.

Essa abertura tem relação, entre outros aspectos, com as publicações científicas, os dados que possibilitaram chegar aos resultados das pesquisas e a transparência quanto ao fluxo de trabalho, metodologias e ferramentas utilizadas. A transparência no fazer científico permite que as pesquisas sejam reprodutíveis. Dá-se atenção, então, a todas as atividades que compõem o fazer científico e observa-se as formas de tornar esse processo replicável. Além de uma responsabilidade moral com respeito ao campo científico, a reprodutibilidade também pode mitigar o fardo do próprio pesquisador. Está relacionada não só com as tecnologias utilizadas, mas também com os hábitos adotados pelos pesquisadores para tornar esse processo mais eficiente (SANDVE *et al*, 2013).

Diante do exposto, o presente estudo busca desenvolver um método automatizado para análise bibliométrica da produção científica disponível no E-LIS, a partir dos documentos depositados por pesquisadores da área de Ciência da Informação no repositório E-LIS, de forma que esse processo seja transparente e reprodutível.

2 REVISÃO DE LITERATURA

[1] http://eprints.rclis.org/

O E-LIS é o primeiro servidor eletrônico aberto especializado em campos relacionados à Biblioteconomia e a Ciência da Informação (SANTILLÁN-ALDANA, 2009) e pode ser entendido como um dos principais recursos abertos dessa área na Internet (LE COADIC, 2004). Seu desenvolvimento foi estimulado pelo conceito de Acesso Aberto e seu acervo é composto de documentos técnico-científicos depositados, geralmente, pelos próprios autores. Esse processo em que os próprios autores depositam uma cópia de suas publicações em plataformas de acesso aberto, como os repositórios digitais temáticos ou institucionais, é denominado autoarquivamento (BOAI, 2002).

Com base na literatura, observou-se que muitos trabalhos sobre o E-LIS foram desenvolvidos desde a sua criação em 2003. O primeiro trabalho a mencioná-lo foi publicado ainda nesse ano. Nele, Barrueco-Cruz e Subirats-Coll (2003) abordam a criação do repositório como parte do projeto *Research in Computing, Library and Information Science* (RCLIS) e apresentam duas outras iniciativas de arquivos abertos para a Biblioteconomia e Ciência da Informação: o @rchiveSIC, projeto francês; e o *Digital Library of Information Science and Technology* (DLIST), com foco em competência informacional e métodos informétricos. Apesar da existência desses repositórios, os autores evidenciam a necessidade de se criar um verdadeiro esforço internacional sem barreiras geográficas ou de idiomas, que seria a proposta do E-LIS. Os artigos seguintes, seguiram uma linha de divulgação para que pesquisadores entendessem essa proposta e trabalhos da área fossem depositados no repositório (DE ROBBIO, 2003; MEDEIROS, 2004; SUBIRATS-COLL; BARRUECO, 2004; ARENCIBIA-JORGE; SANTILLÁN-ALDANA; SUBIRATS-COLL, 2005; MORRISON *et al*, 2007).

No Brasil, Moreno, Leite e Arellano (2006) e Weitzel (2006), ao abordarem o acesso livre a publicações e o papel dos repositórios digitais na comunicação científica, citam o E-LIS como um dos repositórios da área que, desde então, vinha ganhando atenção internacionalmente. Weitzel, Leite e Arellano (2008, p. 14) afirmaram que a "disseminação do conteúdo dos periódicos brasileiros da área no E-LIS é fundamental para a visibilidade da produção científica nacional", reconhecendo a comunidade desse repositório e o impacto que a disponibilização da produção científica nele poderia acarretar.

Alguns trabalhos utilizam o E-LIS como uma de suas fontes de informação. Pujol e Vivó (2010), por exemplo, analisaram o estado da produção e disseminação de teses de doutorado das universidades espanholas em repositórios nacionais e internacionais, entre os anos de 1997 e 2008. Miller (2017), no que lhe concerne, buscou fornecer uma revisão

sistemática dos papéis emergentes ou recentemente adotados pelos profissionais da informação a partir da literatura profissional de Biblioteconomia e Ciência da Informação.

Outros, recorrem ao E-LIS para estudos bibliométricos. Alguns autores buscaram caracterizar a produção científica depositada por pesquisadores de diferentes países, como México (VILLEGAS, 2006) e Portugal (NEVES; FERREIRA, 2014). Osório (2014), por sua vez, pesquisou sobre a cobertura de assunto dos documentos disponíveis no repositório. E, de forma a fazer uma análise mais abrangente, outros autores realizaram uma análise do acervo completo em dado momento. Santillán-Aldana (2009), por exemplo, investigou características do acervo depositado até 2007, enquanto Santos e Oliveira (2019) realizaram uma análise preliminar em 2019 sobre a quantidade de depósitos realizados, a tipologia dos documentos e os assuntos mais abordados.

A partir dessa literatura, principalmente dos trabalhos que analisaram a produção científica depositada no E-LIS, identificou-se que os métodos utilizados para coleta e análise de dados não eram transparentes o suficiente. Isto quer dizer que, a partir do documentado pelos autores, a reprodução ou replicação das pesquisas, desde a coleta até a análise de dados, se torna inviável já que o fluxo de trabalho não é explícito e os dados não estão disponíveis para verificação. É nesse sentido que se entende como necessário pensar e repensar como a ciência é feita e como tornar esse processo mais aberto.

Ao abordarem a transparência e reprodutibilidade das pesquisas nas Ciências Sociais, Miguel *et al* (2014) afirmam que essas práticas ainda são incipientes e devem ser incentivadas, de forma que os pesquisadores se beneficiem das mudanças que podem resultar para o progresso científico geral. Entende-se que cada área do conhecimento terá suas particularidades para tornar sua pesquisa reprodutível. Se a pesquisa considerar aspectos computacionais, algumas regras básicas listadas por Sandve *et al.* (2013) podem ser aplicadas a qualquer campo.

Entre elas, incentiva-se que os pesquisadores: a) registrem cada etapa da pesquisa, observando como cada resultado foi produzido; b) evitem a manipulação manual de dados para que não haja adulteração de resultados; c) arquivem as versões exatas de todos os programas externos e *scripts* utilizados na pesquisa; d) registrem todos os resultados intermediários, quando possível, em formatos padronizados; e) sempre armazenem dados brutos antes de criar representações gráficas, para que outras representações possam ser feitas; f) conectem declarações textuais aos seus resultados de tal forma que outros

pesquisadores possam chegar às mesmas conclusões; g) forneçam acesso público a *scripts*, execuções e resultados da sua investigação (SANDVE *et al*, 2013).

Conforme apresentado e considerando a importância da análise periódica da produção científica para verificação de indicadores bibliométricos, busca-se criar um instrumento para caracterizar a produção científica e auxiliar no acompanhamento das práticas de autoarquivamento na Biblioteconomia e Ciência da Informação.

3 PERCURSO METODOLÓGICO

O estudo é aplicado, descritivo e de caráter exploratório. Ao emprestar filosofias, ferramentas e fluxos de trabalho criados principalmente para o desenvolvimento de *softwares*, busca-se desenvolver um instrumento capaz de coletar e analisar automaticamente os dados da produção científica disponível no E-LIS.

Para a coleta e análise dos dados, utilizou-se a linguagem de programação Python 3 em conjunto com o ambiente computacional Jupyter Notebook, aplicação baseada na *web* e contida na plataforma Anaconda. Optou-se pela utilização dessa ferramenta porque ela permite interação entre texto e código de programação, transformando o trabalho legível também para humanos. Se bem documentado, o trabalho no Jupyter Notebook pode descrever o fluxo de trabalho do pesquisador, com os passos e os códigos, de forma que a análise possa ser reproduzida por outros pesquisadores, ou até mesmo pelo próprio pesquisador futuramente. Nele, é fácil modificar uma análise, sendo possível estendê-la ou refiná-la, mudando os códigos existentes ou adicionando novos blocos de código. Pode-se, também, compartilhar as análises realizadas, pois ele permite que o documento seja exportado em diferentes formatos, como o PDF, HTML e apresentações em *slide* (WASSER, 2018).

O trabalho foi desenvolvido considerando as seguintes especificações técnicas: Sistema Operacional Ubuntu 20.04 LTS (64-bits); linguagem de programação Python 3.7.3 no ambiente computacional Jupyter Notebook 6.0.0; plataforma de Ciência de Dados Anaconda; e navegador *web* Mozilla Firefox. As ferramentas empregadas aqui são gratuitas e podem ser baixadas em seus respectivos *sites*, estando disponíveis para Linux, Windows e MacOS.

Considerando-se o grau de incursão dispensado à criação dos instrumentos de coleta e análise de dados, eles serão apresentados como resultados no tópico a seguir.

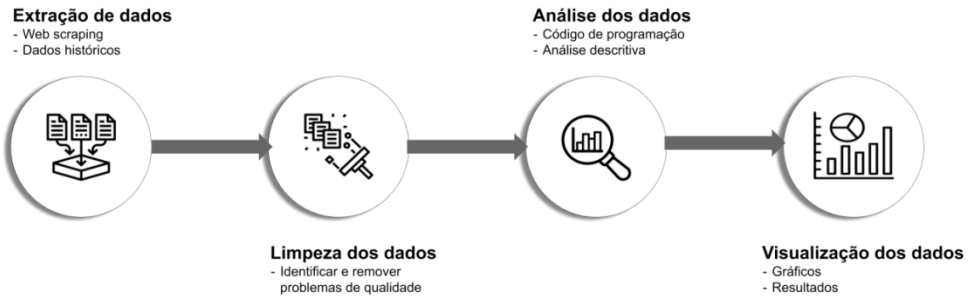
4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Criou-se, então, um *script* para coleta e análise de dados da produção científica disponível no repositório E-LIS, chamado **ELIScript**. Com ele, é possível extrair, tratar e visualizar os dados, atividades complementares que possibilitaram chegar aos resultados deste trabalho. O ELIScript[2] está organizado em quatro pastas e dois arquivos. O esquema das pastas e arquivos é apresentado abaixo:

- **db:** db, de *database,* é uma pasta com todos os arquivos coletados pelo ELIScript em CSV, isto é, todas as planilhas com os metadados das 23.245 publicações, separadas por continentes;
- **db_generated:** são as planilhas e tabelas geradas pelo ELIScript, que permitirão gerar os gráficos para análise;
- **plots**: nessa pasta estão disponíveis todos os gráficos gerados pelo ELIScript;
- **source**: os códigos-fonte do ELIScript divididos em três partes, o *ELIScript_collection* para a coleta de dados; e o *ELIScript_organization* para organização dos dados para posterior análise; e o *ELIScript_analysis* para análise de dados propriamente dita;
- **LICENSE**: nesse arquivo consta o texto da licença GPL v3.0, sob a qual o E-LIScript foi distribuído;
- **README.md**: é um arquivo de texto no formato *markdown*, com as informações sobre o ELIScript; introdução sobre o projeto, responsáveis pela criação do *script*, dependências a serem instaladas e agradecimentos.

Para explicação das funções do ELIScript, ele foi dividido em três partes. A primeira é utilizada para extração de dados. A segunda, para organização dos dados por meio de agrupamento de planilhas para facilitar as análises por continente. E a terceira para análise dos dados, na qual é feita a limpeza e, a partir da análise, a visualização dos dados (Figura 1).

Figura 1: *Framework* do *script*



Fonte: Santos (2021).

A seguir esses elementos são descritos.

[2] O repositório do ELIScript está disponível publicamente no endereço: https://github.com/sarahrubia/eliscript. Nele, é possível ter acesso a todos os dados coletados para essa pesquisa, gráficos e tabelas gerados, bem como os códigos para coleta e análise dos dados.

4.1 Extração de Dados em Páginas da Web

O E-LIS possui uma página[3] com uma lista dividida por continente e países, no qual se encontram as planilhas com os metadados dos documentos do repositório. O *script* acessa diretamente essa página e recupera o *Uniform Resource Locator* (URL) referente a cada país. Após recuperar os URL, que ficarão salvos em uma lista, o navegador Mozilla Firefox acessa cada um deles e executa comandos de exportação das planilhas. Sequencialmente, o *script*: a) abre o navegador *web*; acessa um URL na lista de URL recuperados; b) localiza o menu suspenso com os formatos de arquivos para exportação; c) seleciona o formato *multiline CSV*, que corresponde às planilhas CSV; d) clica no botão *Export*; e e) faz o *download* da planilha no diretório definido nas preferências do navegador.

O navegador será aberto apenas uma vez e será atualizado cada vez que abrir uma página. Ao final da execução, o navegador será fechado. O processo descrito foi realizado para cada URL na lista de URL salvos até que todos tenham sido acessados. A coleta de dados foi realizada no dia 9 de setembro de 2020. Verificou-se a existência de 125 conjuntos de dados, cada um correspondente a um país.

4.2 Organização e Limpeza dos Dados

A organização e a limpeza dos dados podem ser entendidas como uma etapa de pré-processamento para identificação de problemas com os dados recuperados. Os problemas com os dados sempre existem quando se lida com dados da realidade. A natureza e severidade desses problemas, no entanto, dependem de diversos aspectos: podem ser erros na entrada ou na saída dos dados; uma quantidade demasiada ou insuficiente de dados, dados corrompidos ou incompatíveis, entre outros. A não realização dessa etapa pode afetar a análise dos dados e, consequentemente, os resultados de um trabalho (FAMILI *et al*., 1997).

A preparação para a análise de dados, então, consiste do entendimento da natureza dos dados e o uso de técnicas para pré-processamento destes. As técnicas de pré-processamento utilizadas nos dados do E-LIS são divididas em dois grupos: 1) transformação de dados e 2) compilação de informações. A primeira corresponde a técnicas de filtragem, ordenação e edição dos dados. A segunda, por sua vez, abrange técnicas de visualização, eliminação, seleção e amostragem de dados (FAMILI *et al.*, 1997).

[3] http://eprints.rclis.org/view/countries/

As planilhas com os metadados de cada publicação que foram baixadas estão divididas por país. Para possibilitar uma análise comparativa entre continentes, as planilhas dos países que pertencem ao mesmo continente foram concatenadas, gerando uma planilha para cada continente, a saber: África (187 documentos); América, dividida em: a) América do Norte e Central (3.345 documentos); e b) América do Sul (4.125 documentos), divisão política feita pelo próprio repositório; Antártica (2 documentos); Ásia (2.169 documentos); Europa (13.253 documentos); e Oceania (164 documentos). Totalizando metadados de 23.245 documentos para análise.

Precedendo cada análise, os processos de transformação dos dados foram empregados. Por meio da filtragem e ordenamento foi possível extrair e compilar as informações necessárias para gerar a visualização dos dados com gráficos e tabelas. Para facilitar o acompanhamento do que foi desenvolvido, as técnicas de pré-processamento dos dados serão descritas em conjunto com os processos de análise, dado que é difícil determinar onde termina um e onde começa o outro.

4.3 Análise e Visualização dos Dados

O *script* é formado por várias funções que executam um conjunto de instruções, de forma que os dados do repositório E-LIS sejam analisados. Algumas categorias de análise foram consideradas baseando-se nos metadados dispostos como colunas nas planilhas CSV. Essas categorias consistem na tipologia do documento (*types*), assunto do documento (*subjects*), data de publicação do item (*date*), data de depósito do item (*datestamp*), idioma da publicação (*linguabib*), periódico em que o artigo foi publicado (*publication*), evento em que artigo, apresentação ou pôster foram apresentados (*event_title; conference*) e palavras-chave mais utilizadas (*keywords*).

De forma geral, o *script* acessa a planilha de cada continente, seleciona a coluna correspondente a cada categoria de análise, extrai seus dados, gera tabelas com os dados divididos por continentes. Essas tabelas são usadas como entrada para gerar gráficos comparativos. No caso do assunto e do idioma das publicações, houve ainda a conversão dos códigos que, na primeira categoria, são dados pela notação do esquema de classificação JITA[4]

[4] O nome da classificação JITA é um acrônimo derivado da letra inicial do nome de cada autor do esquema: José Manuel Barrueco-Cruz, Imma Subirats-Coll, Thomas Krichel e Antonella De Robbio (E-LIS, 2020). Esse esquema é utilizado especificamente para classificar documentos da Biblioteconomia e Ciência da Informação.

; e na segunda, são representados pelos códigos da norma técnica International Organization for Standardization (ISO) 639-1.

4.4 Alguns Resultados

A partir dos 125 conjuntos de dados, concatenados em sete planilhas CSV, uma para cada continente, foram gerados 57 gráficos e 40 tabelas. Por meio dos dados levantados de 23.245 publicações, foi possível identificar quais tipologias foram mais depositadas no E-LIS (Quadro 1).

Quadro 1: Tipologias de documento no E-LIS

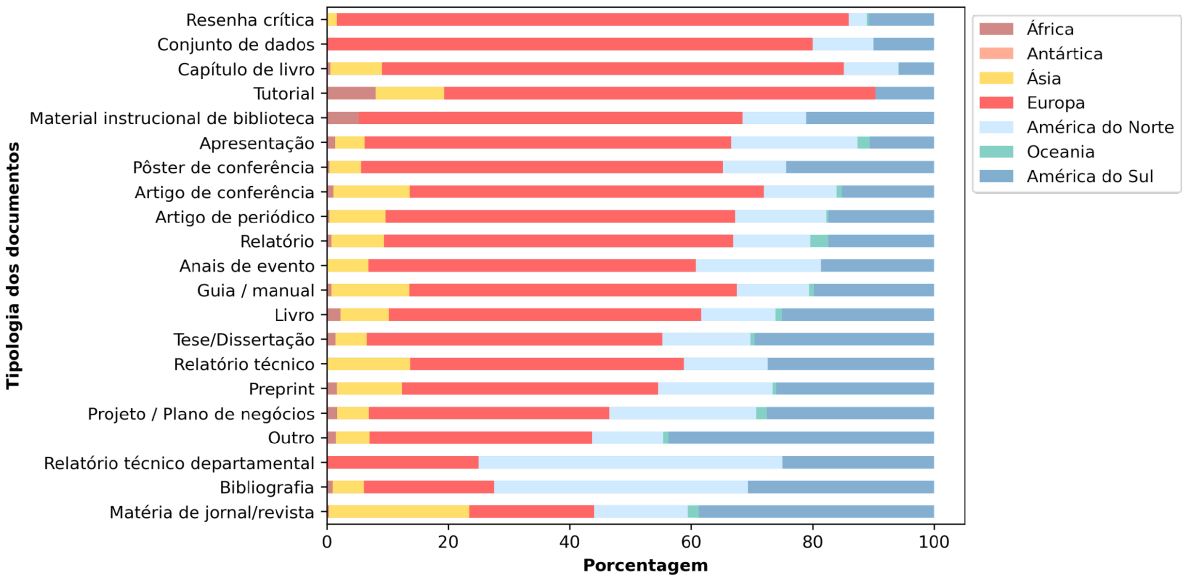
Tipo de documento	Ano (2007)*	% (2007)	Ano (2020)**	% (2020)
Artigo de periódico	3323	48	10861	47
Artigo de conferência	1662	24	4431	19
Apresentação	699	10	1826	8
Preprint	233	3	1072	5
Tese/Dissertação	192	3	731	3
Capítulo de livro	161	2	1171	5
Relatório técnico departamental	111	2	4	0
Matéria de jornal/revista	89	1	320	1
Pôster de conferência	84	1	530	2
Revisão de Literatura	80	1	299	1
Livro	56	1	391	2
Guia/Manual	50	1	243	1
Relatório técnico	42	1	51	0
Bibliografia	28	0	98	0
Anais de evento	25	0	102	0
Projeto / Plano de negócios	21	0,4	58	0
Material instrucional de biblioteca	20	0,4	19	0
Tutoriais	17	0,4	62	0
Ementa	1	0	0	0
Relatório	0	0,4	511	2
Conjunto de dados (d*ataset*)	0	0,4	10	0
Outro	0	0,4	455	2
TOTAL	6894	100	23245	100

*Valores do artigo de Santillán-Aldana (2009);
** Valores obtidos com o ELIScript.
Fonte: dados da pesquisa (2020).

Identificou-se que os artigos de periódico (47%), artigos submetidos para eventos científicos (19%) e apresentações (8%) são as categorias de documento mais frequentemente depositadas, tipologias que também aparecem como as mais compartilhadas em 2007 (SANTILLAN-ALDANA, 2009), evidenciando os produtos científicos mais comumente produzidos.

No Gráfico 1, faz-se uma comparação entre os tipos de documentos depositados por cada continente. Considera-se que cada barra corresponde a 100% de uma tipologia de documento e cada cor corresponde a um continente. nota-se a predominância da Europa no depósito de praticamente todos os tipos de documento, com exceção dos Relatórios Técnico Departamentais e Bibliografias, sendo ultrapassado pela América do Norte e Central; e Matéria de Jornal/Revista e Outro, que foi mais depositado por pesquisadores da América do Sul. Apenas a Europa (vermelho), a América do Norte e Central (azul-claro), e a América do Sul (azul-escuro) realizaram depósito de todas as tipologias de documentos. A Ásia (amarelo) também aparece em destaque, embora não tenha sido localizado nenhum depósito de Conjunto de Dados, nem de Relatório Técnico Departamental. A Antártica só possui dois documentos depositados, um Capítulo de Livro e um *Preprint*, tornando-se imperceptível no gráfico.

Gráfico 1: Distribuição de tipologias por continente



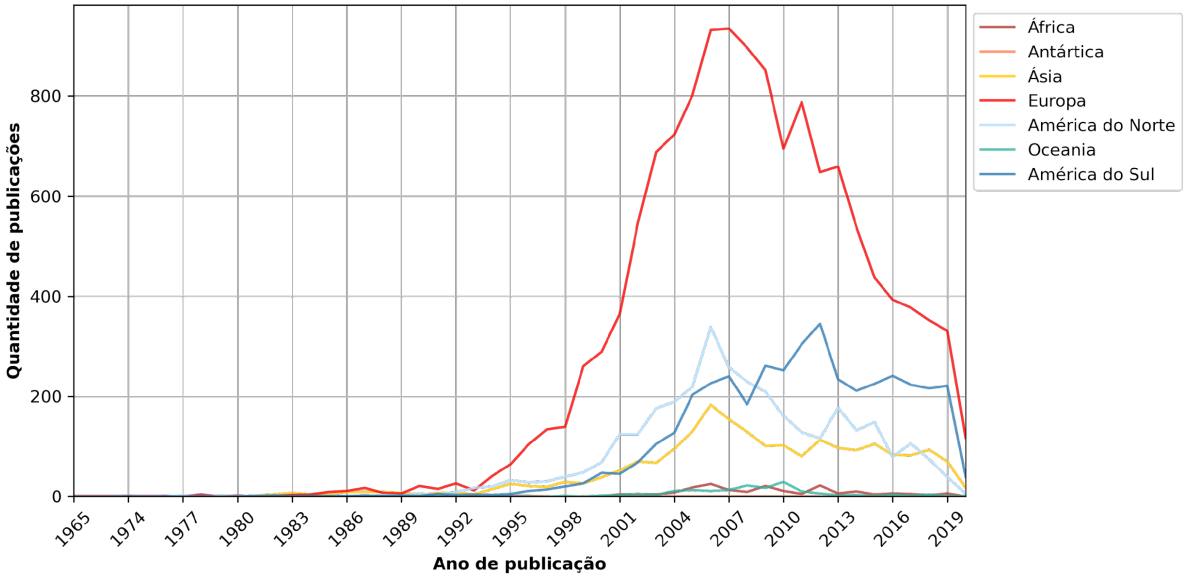
Fonte: dados da pesquisa (2020).

As práticas de autoarquivamento por pesquisadores da Biblioteconomia e Ciência da Informação compreendem o período entre 2002 e 2020, isto é, desde a criação do

repositório até o momento da coleta de dados. Em todos os anos houve autoarquivamento, no entanto, em 2007 mais de 3,5 mil documentos foram depositados, possivelmente por influência das atividades de divulgação do repositório (MORRISON *et al.*, 2007).

Nenhum dos trabalhos que analisaram a produção científica disponível no E-LIS, seja o acervo inteiro ou a produção específica de um país, abordou a atualidade dos documentos (VILLEGAS, 2006; SANTILLAN-ALDANA, 2009; NEVES; FERREIRA, 2014). Os documentos depositados no repositório foram publicados entre 1965 e 2020 (Gráfico 2). Observou-se que mais trabalhos podem ser encontrados entre 1983 e 2020. No período entre 1965 e 1983, há um número pouco expressivo de trabalhos publicados. A partir de 1983, no entanto, percebe-se um crescimento. A maior parte dos trabalhos foi publicada em 2007, ano em que mais depósitos foram realizados, evidenciando que os trabalhos costumam ser depositados no mesmo ano em que são publicados. Laakso (2014) aponta que o momento em que o autoarquivamento pode ser realizado depende das políticas editoriais do periódico, no caso de artigos ou dados publicados em revistas científicas.

Gráfico 2: Ano de publicação dos documentos depositados no E-LIS



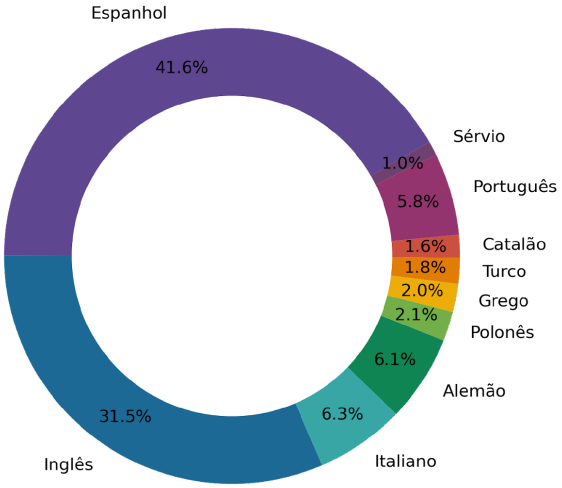
Fonte: dados da pesquisa (2020).

Entre os assuntos mais abordados nos documentos, identificou-se, de forma geral, o "Uso da Informação e Sociologia da Informação" (1.764 documentos), a "Tecnologia da Informação e Tecnologia de Biblioteca" (1.594 documentos), e "Tratamento da Informação para Serviços de Informação" (1.164 documentos). Em relação às classes mais específicas de

assunto na classificação JITA, tem-se as temáticas "Disseminação e difusão da informação" (1.343 documentos) em primeiro lugar, seguida de “Repositórios (baseado ou não em OAI)” (1.128 documentos) e, por fim, "Métodos bibliométricos" (967 documentos).

Os autores podem caracterizar sua produção escolhendo entre 38 idiomas ao submeter um documento ao E-LIS. Na pesquisa foram identificados documentos em 37 idiomas. Os 10 idiomas principais foram elencados no Gráfico 3.

Gráfico 3: Idioma das publicações



Fonte: dados da pesquisa (2020).

Observou-se a prevalência do espanhol (41,6%) como o idioma com maior representação nas publicações. Em seguida, estão o inglês (31,5%), o italiano (6,3%), o alemão (6,1%) e o português (5,8%). Morrison *et al.* (2007) caracterizam o inglês como o idioma principal do repositório. A interface do E-LIS está em inglês, além disso, cada documento submetido em outro idioma que não seja o inglês deve ser acompanhado de um resumo em inglês. Todavia, observou-se que a Espanha e os países falantes da língua espanhola da América do Sul e da América do Norte e Central são os responsáveis pela maior quantidade de documentos do repositório.

A pesquisa possibilitou identificar, ainda, os periódicos com mais publicações depositadas no E-LIS. Os cinco periódicos com mais publicações são o El Profesional de La Información (610 artigos), o ACIMED (472 artigos), o VÖB-Mitteilungen (463 artigos), o Comunicar (300 artigos) e o Anales de Documentación (291). Desses, apenas o ACIMED,

agora denominado Revista Cubana de Información en Ciencias de la Salud, não é editado em um país europeu. Uma das características observada é que grande parte dos periódicos com publicações depositadas no E-LIS estão indexados em bases de dados importantes da Biblioteconomia e Ciência da Informação, como a LISTA, a LISA e a ISTA. Além delas, os periódicos estão indexados por bases generalistas como a Web of Science e a Scopus, indicando a relevância dos documentos depositados.

Com relação aos eventos com mais publicações depositadas, foram identificados problemas na análise devido, principalmente, à falta de padronização na inserção dos títulos e o número da edição dos eventos. No momento da submissão de um documento apresentado ou publicado em evento, não existe um campo específico para adicionar a edição do evento, levando os autores a inserirem a edição junto ao título desse. Apesar disso, foi possível verificar a existência de eventos regionais, nacionais e internacionais. O evento com mais publicações depositadas foi o 69th Annual Meeting of the American Society for Information Science and Technology (ASIST), na América do Norte e Central.

Apesar de existirem mais trabalhos em língua espanhola no repositório, ao analisar as palavras-chave, observou-se a incidência de mais termos em inglês. Os termos com maior representação foram “biblioteca”, “informação”, “acesso aberto”, “usuário”, “digital”, “serviço” e “dados” (em inglês). Por mais gerais que sejam esses termos, eles representam o propósito do repositório de reunir documentos pertinentes para as áreas de Biblioteconomia e Ciência da Informação. Eles evidenciam, assim, o cerne da área, que é a preocupação com as formas de trabalhar com os fluxos da informação, independente do meio.

5 CONSIDERAÇÕES FINAIS

A partir desse estudo, observou-se que é possível pensar maneiras de tornar o percurso metodológico de pesquisas nas ciências sociais mais aberto e transparente. Não é necessário pensar em programação ou em recursos computacionais para isso, mas documentar minuciosamente o processo, quais ferramentas foram utilizadas, o que foi descoberto ao longo da pesquisa, as falhas e o acertos, de forma a gerar fluxos de trabalho mais claros e fáceis de serem seguidos, seja pelos avaliadores dos artigos, por pesquisadores que querem utilizar esses dados para pesquisa ou para o próprio autor do trabalho caso precise revisar resultados.

O ELIScript foi criado para possibilitar análises reprodutíveis e servir como uma

ferramenta para análise periódica do acervo do E-LIS. Com ele, pôde-se coletar os dados pertinentes a todo o acervo do repositório, fazer a limpeza e organização desses dados, e analisá-los conforme as categorias de análise propostas. Embora o instrumento seja capaz de realizar a coleta de dados do acervo inteiro em poucas horas e as análises em alguns segundos, gerando cerca de 50 gráficos e 40 tabelas, o seu desenvolvimento envolveu um longo período de testes e de falhas até chegar ao produto final. Para além de sua criação, o *script* foi aplicado e sua viabilidade confirmada, o que permitiu levantar dados significativos sobre o acervo, bem como acerca do cenário das práticas de autoarquivamento na área de Biblioteconomia e Ciência da Informação.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

REFERÊNCIAS

ARENCIBIA-JORGE, R. SANTILLÁN-ALDANA, J.; SUBIRATS-COLL, I. Iniciativas de acceso abierto en Ciencias de la Información y Documentación: evolución y perspectivas de E-LIS. **Revista Española de Documentación Científica**, v. 28, n. 2, p. 221-232, 2005. Disponível em: http://hdl.handle.net/10760/6633. Acesso em: 24 maio 2019.

BARRUECO-CRUZ, J. M.; SUBIRATS-COLL, I. RCLIS: towards a digital library for Information Science. **Biblioteconomia i Documentació**, p. 1-11, 2003. Disponível em: http://bid.ub.edu/11barru2.htm. Acesso em: 24 maio 2019.

BOAI. **Budapest Open Access Initiative**. Budapeste, 2002. Disponível em: https://www.budapestopenaccessinitiative.org/read. Acesso em: 03 maio. 2021.

DE ROBBIO, A. E-LIS: un open archive per library and information science. **AIB Notizie**, v. 15, n. 2, p. 1-2, 2003. Disponível em: http://hdl.handle.net/10760/5366. Acesso em: 24 maio 2019.

E-LIS. **E-Prints in Library & Information Science.** 2020. Disponível em: http://eprints.rclis.org/. Acesso em: 3 maio 2021.

FAMILI, A. *et al*. Data preprocessing and intelligent data analysis. **Intelligent Data Analysis**, v. 1, n. 1, jan. 1997, p. 3-23. Disponível em: https://content.iospress.com/articles/intelligent-data-analysis/ida1-1-02. Acesso em: 03 jul. 2020.

LAAKSO, M. Green open access policies of scholarly journal publishers: a study of what, when, and where self-archiving is allowed. **Scientometrics**, v. 99, n. 2, p. 475-494, 2014.

Disponível em: https://link.springer.com/article/10.1007/s11192-013-1205-3. Acesso em: 27 maio 2019.

LE COADIC, Y. F. **A ciência da informação.** 2. ed. rev. e atual. Brasília: Briquet de Lemos, 2004. 124p.

MEDEIROS, N. A repository of our own: the E-LIS e-prints archive. **OCLC Systems & Services: International digital library perspectives**, v. 20, n. 2, p. 58-60, jun. 2004. Disponível em: http://www.emeraldinsight.com/doi/10.1108/10650750410539040. Acesso em: 21 abr. 2019.

MIGUEL, E. *et al*. Promoting transparency in social science research. **Science**, v. 343, n. 6166, p. 30-31, 2014. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103621/. Acesso em: 20 mar. 2020.

MILLER, R. E. Literature Suggests Information Professionals Have Adopted New Roles. **Evidence Based Library and Information Practice**, v. 12, n. 1, p. 137-139, 2017. Disponível em: https://ejournals.library.ualberta.ca/index.php/EBLIP/article/view/28739. Acesso em: 24 maio 2019.

MORENO, F. P.; LEITE, F. C. L.; ARELLANO, M. Á. M. Acesso livre a publicações e repositórios digitais em ciência da informação no Brasil. **Perspectivas em Ciência da informação**, Belo Horizonte, v. 11, n. 1, p. 82–94, jan/abr. 2006. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/447/258. Acesso em: 27 maio 2019.

MORRISON, B. H. et al. E-LIS: The Open Archive for Library and information Science. **The Charleston Advisor**, v. 9, n. 1, p. 56-61, 2007. Disponível em: http://eprints.rclis.org/10158/. Acesso em: 27 maio 2019.

NEVES, B.; FERREIRA, C. Caracterização da produção científica portuguesa em Ciência da Informação disponibilizada em acesso aberto no E-LIS. **Cadernos BAD**, v. 1, n. 2, p. 95-98, 2014. Disponível em: https://www.bad.pt/publicacoes/index.php/cadernos/article/download/1184/1191. Acesso em: 27 maio 2019.

OSORIO, N. L. An Analysis of Subject Coverage and Worldwide Involvement of E-LIS: the International Repository for Library and Information Science. **Library Philosophy and Practice**, n. 1067, 2014. Disponível em: https://digitalcommons.unl.edu/libphilprac/1067/. Acesso em: 13 jan. 2021.

PONTIKA *et al*. Fostering open science to research using a taxonomy and an eLearning portal. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE TECHNOLOGIES AND DATA-DRIVEN BUSINESS, 15. Graz, Austria, 2015. **Proceedings** [...]. Disponível em: https://dl.acm.org/doi/pdf/10.1145/2809563.2809571. Acesso em: 26 abr. 2021.

PUJOL, E. F.; VIVÓ, L. A. Las tesis doctorales en españa (1997-2008): Análisis, estadísticas y repositorios cooperativos. **Revista Española De Documentacion Cientifica**, v.33, n. 1, p. 63-89, 2010. Disponível em:

https://search-proquest.ez27.periodicos.capes.gov.br/docview/212061030?accountid=134127. Acesso em: 24 maio 2019.

ROGEL-SALAZAR, J. **Data Science and Analytics with Python**. Boca Raton: CRC Press, 2017. 376 p.

SANDVE, G. K. *et al.* Ten simple rules for reproducible computational research. **PLoS computational biology**, v. 9, n. 10, 2013. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3812051/. Acesso em: 20 mar. 2020.

SANTILLÁN-ALDANA, J. The open access movement and the library world seen from the experience of the E-LIS project. **OCLC Systems & Services**: **International digital library perspectives**, v. 25, n. 2, p. 135-147, maio 2009. Disponível em: https://www.emeraldinsight.com/doi/full/10.1108/10650750910961938. Acesso em: 24 maio 2019.

SANTOS, S. R. O. **Método automatizado para análise do autoarquivamento na Ciência da Informação**: ELIScript. 2021. 187 f. Dissertação (Mestrado em Gestão e Organização do Conhecimento) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte. 2021. Disponível em: https://repositorio.ufmg.br/handle/1843/36664. Acesso em:

SANTOS, S. R. O.; OLIVEIRA, D. A. Autoarquivamento na Ciência da Informação: Uma análise dos documentos depositados no repositório digital e-LIS. **Múltiplos Olhares em Ciência da Informação**, v. 9, n. 2, 2019. Disponível em: https://periodicos.ufmg.br/index.php/moci/article/view/19126. Acesso em: 03 maio. 2021.

SUBIRATS-COLL, I.; BARRUECO, J. M. Un archivo abierto en ciencias de la documentación e información. **El profesional de la información**, v. 13, n. 5, p. 346-352, 2004. Disponível em: http://eprints.rclis.org/5578/. Acesso em: 03 maio. 2021.

VILLEGAS, M. A. S. Iniciativas de acceso abierto y perspectivas de E-LIS en México. *In*: CONGRESO INTERNACIONAL DE INFORMACIÓN, 17, 2006, La Habana, Cuba. **Anais** [...]. La Habana, 2006. Disponível em: http://eprints.rclis.org/7528/. Acesso em: 03 maio. 2021.

WASSER, L. **Earth Analytics Python Course**: earthlab/earth-analytics-python-course: Version 1.0. Zenodo, 2018. DOI: 10.5281/zenodo.2209415. Disponível em: https://www.earthdatascience.org/courses/earth-analytics-python/. Acesso em: 17 out. 2019.

WEITZEL, S. R. O papel dos repositórios institucionais e temáticos na estrutura da produção científica. **Em Questão**, Porto Alegre, v. 12, n. 1, p. 51-71, 2006. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/19. Acesso em: 24 maio 2019.

WEITZEL, S. R.; LEITE, F. C. L..; ARELLANO, M. Á. M. E-LIS: um repositório digital para a biblioteconomia e ciência da informação no Brasil. *In:* SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 15., 2008, São Paulo. **Anais** [...]. São Paulo: CRUESP, 2008. p. 1-16. Disponível em: http://eprints.rclis.org/12537/. Acesso em: 21 abr. 2019.