Uma Plataforma para o Tratamento e Integração de Dados Científicos

Washington Segundo
IBICT
Brasília, Brasil
washingtonsegundo@ibict.br

Thiago Dias
CEFET-MG
Divinópolis, Brasil
thiagomagela@cefetmg.br

Tales Moreira
CEFET-MG
Divinópolis, Brasil
tales.info@gmail.com

Adilson Pinto
UFSC
Florianópolis, Brasil
adilson.pinto@ufsc.br

Vivian Silva
IBICT
Brasília, Brasil
vivian.ss@gmail.com

Josir Gomes
IBICT
Rio de Janeiro, Brasil
josir@irdx.com.br

Luc Quoniam
UFMS
São Paulo, Brasil
quoniam.luc@gmail.com

Lautaro Matas
La Referencia
Espanha
lmatas@gmail.com

Ary Dias
IBICT
Brasília, Brasil
arygabrieldias@gmail.com

Juliana Schneider
IBICT
Brasília, Brasil
jschineider.js@gmail.com

***Abstract* — In the last years few initiatives that aimed at creating systems that manage the academic production of an institution, country or area of knowledge have received attention from several areas. Such systems are known by the acronym CRIS (Current Research Information Systems) and aim to aggregate information from diverse databases in order to provide reports and consolidated data for researchers to analyze. Therefore, this work presents the development process of the BrCris platform with the objective of providing technological tools in order to provide the Brazilian academic community with consolidated data from the national scientific production.**

Keywords – Scientific Data, Big Data, Open Science.

***Resumo* — Nos últimos anos várias iniciativas que visavam a criação de sistemas que gerenciam a produção acadêmica de uma instituição, país ou área de conhecimento tem recebido atenção de diversas áreas. Tais sistemas são conhecidos pela sigla CRIS (*Current Research Information Systems*) e têm como objetivo agregar informações de bases de dados diversas com intuito de fornecer relatórios e dados consolidados para que pesquisadores possam analisar. Logo, este trabalho apresenta o processo de desenvolvimento da plataforma BrCris com o objetivo de fornecer ferramentas tecnológicas com o intuito de munir a comunidade acadêmica brasileira com dados consolidados da produção científica nacional.**

***Palavras-chave* — Dados Científicos, Big Data, Ciência Aberta.**

I. INTRODUÇÃO

A produção científica brasileira tem crescido expressivamente e, em perspectiva às especificidades de campos disciplinares distintos, heterogênea quanto à tipificação de sua produção tanto em termos quantitativos como qualitativos. E o resultado desta produção se materializa em forma de artigos em periódicos, teses e dissertações, além de produtos diversos como: *softwares*, patentes, obras e instalações artísticas, entrevistas e projetos cinematográficos.

Para o campo da Ciência da Informação, e em especial da Cientometria, quantificar essa produção é uma tarefa árdua pois a disponibilização de bases de dados abertas muitas vezes é restrita ou simplesmente inexistente em determinados contextos. Bases de dados proprietárias como a Scopus, Web of Science, Google Acadêmico e Microsoft Research Data permitem o acesso mas este é sempre limitado ao número de registros que podem ser obtidos, contemplam poucos repositórios e periódicos nacionais e ainda existe o grave problema da opacidade dos algoritmos utilizados por estas plataformas que determinam o que é ou não relevante.

A partir desse cenário, começaram a surgir iniciativas que visavam a criação de sistemas que gerenciam a produção acadêmica de uma instituição, país ou área de

conhecimento. Tais sistemas são conhecidos pela sigla CRIS (*Current Research Information Systems*) e têm como objetivo agregar informações de bases de dados diversas com intuito de fornecer relatórios e dados consolidados para que pesquisadores da área possam analisar como se dá a produção em seus países ou áreas de conhecimento.

CRIS define um sistema de informação sobre todo o ecossistema do processo científico. São organizadas em um só lugar todas informações do ciclo da pesquisa Científica, desde o Fomento, passando pelos Projetos, Pesquisadores, Instituições de Pesquisa e Laboratórios, até os outputs de uma pesquisa científica, tais como artigos científicos, teses, dissertações, livros, capítulos de livro, patentes e conjuntos de dados científicos [1].

Neste contexto, a idealização do Projeto do Sistema BrCris, que é o CRIS no contexto da Ciência Brasileira, data de 2014, quando inspirado no modelo proposto por Portugal de um CRIS nacional (o PTCRIS - https://ptcris.pt), o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) iniciou uma sequência de estudos e parcerias interinstitucionais para a execução do Projeto. Em 2020, houve a implementação formal de um Projeto de Pesquisa para a construção do BrCris. O intuito é fornecer ferramentas tecnológicas visando munir a comunidade acadêmica brasileira com dados consolidados da produção científica nacional. Tomando como base outros projetos CRIS e padrões internacionais disponibilizados pelo OpenAire e COAR.

Logo, o BrCris tem por objetivo estabelecer um modelo único de organização da informação científica de todo o ecossistema da pesquisa brasileira. Entre os agentes deste ecossistema estão os pesquisadores, os projetos, infraestruturas, laboratórios e instituições de pesquisa, os financiadores, além dos resultados da pesquisa expressos principalmente por publicações científicas, teses, dissertações, conjuntos de dados científicos, *software* e patentes (ver Figura 1).

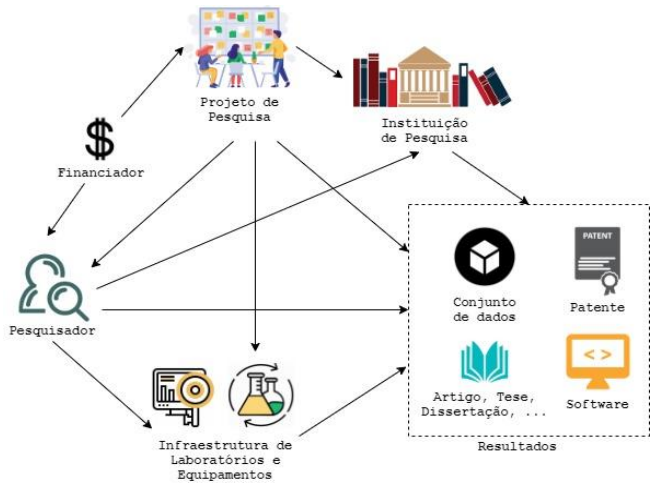

Fig. 1. Ecossistema da pesquisa científica.

Diante disso, com a integração dos dados em um repositório de dados padronizado, o uso do dashboard em forma de visualização elucida alguns benefícios, como a redução de complexidade de dados, auxilia na percepção das propriedades existentes, ajuda na detecção de erros aparentes, consegue englobar a representação em pouco conteúdo, amplia a percepção cognitiva, dentre outros.

Outro aspecto que vale ser salientado na função do dashboard é a geração de índice e indicadores visando atribuir a mensuração de fenômenos, sejam de natureza social, econômica ou científico/tecnológica. No contexto deste trabalho, valoriza-se o foco em sua representação e facilidade de identificação dos cenários.

Tendo em vista o que foi exposto, este trabalho tem como objetivo apresentar o processo de desenvolvimento do projeto BrCris, que visa coletar, integrar e disponibilizar informações relativas ao universo de pesquisa científica no Brasil, catalogando e traçando relacionamentos entre pesquisadores, as organizações às quais pertencem, os projetos dos quais participam e como são financiados, e todos os produtos por ele gerados, como publicações, patentes e *software*.

II. DESENVOLVIMENTO

O BrCris concentra um amplo ecossistema de dados, de diversas fontes, como por exemplo, dados curriculares de indivíduos, sobre organizações, programas de pós-graduação, publicações, orientações, revistas científicas, dentre outros, sendo necessário todo um esforço para tratamento dos dados de interesse. Neste contexto, tendo em vista as diversas fontes de dados que irão compor o BrCris se faz necessário a transformação dos dados em formato padronizado, sendo necessário a transformação baseada em um modelo que será importado para a plataforma LA Referencia.

Em se tratando do modelo de dados do BrCris, iniciou-se pela adoção de nove entidades de dados, seguindo padrões amplamente utilizados na comunidade científica internacional. São elas:

- Project: projetos de pesquisa executados, ou em execução;
- Service: revistas científicas, repositórios digitais, bibliotecas digitais e outras fontes de informação científica;
- Program: programas de pós-graduação brasileiros;
- Course: cursos nacionais, ou internacionais de pós-graduação stricto ou lato sensu;
- OrgUnit: instituições, faculdades, departamentos de pesquisa;

• Person: pesquisadores, assistentes de pesquisa e pessoas de apoio técnico à pesquisa;

• Patent: patentes como resultado da pesquisa;

• Dataset: conjuntos de dados de pesquisa coletados por pesquisadores e demais agentes no âmbito de um projeto ou pesquisa científica;

• Publication: artigos científicos, teses, dissertações, livros, capítulos de livro e relatórios científicos.

O modelo de dados é definido por um conjunto de entidades e relações, que por sua vez possuem identificadores e atributos pré-definidos. A utilização de um descritivo visa facilitar a identificação de atributos de cada entidade (Figura 2) e suas relações (Figura 3), possibilitando que com o auxílio de uma rotina desenvolvida especificamente para esta funcionalidade, o modelo possa incorporar todas as mudanças realizadas diretamente no modelo. Esta estratégia visa facilitar de forma significativa a incorporação de novos atributos e relações, sem a necessidade de alterações diretamente no modelo de dados.

Field	Description	Diadorim	DOAJ
identifier.diadorim	id diadorim	id	
identifier.latindex	id LatinDex		
identifier.ulrichsweb	id ULrichsweb		
identifier.issn	id issn	dc.identifier.issn[pt_BR]	Journal EISSN (online version)
identifier.issnl	id issnl	dc.identifier.issnl[pt_BR]	Journal ISSN (print version)
identifier.oai	id oai	dc.identifier.oai	
identifier.uri	id uri	dc.identifier.uri	Journal URL
identifier.brcris	hash gerado com título + publisher	dc.title[pt_BR] + dc. publisher.name[pt_BR]	Journal title + Publisher
identifier.eissn	(Será considerado como um issn)		Journal EISSN (online version)
identifier.opendoarId			
identifier.other			
compatibility			
acronym			
status	status da revista (active / inactive)	dc.relation situation[pt_BR]	
accessType	tipo (direito) de acesso	dc.rights.access[pt_BR]	Permite acesso ao texto completo
ccLicence	Permissões	dc.rights.cc[pt_BR]	Journal license
rightsType		dc.rights.type[pt_BR]	
researchArea	área de pesquisa		

Fig. 2. Entidades do Descritivo.

Name	From Entity	From Label	To Entity	To Label	Description
Affiliation	OrgUnit	hasMember	Person	isMemberOf	Is a relation beteween Person and OrgUnit
IsUnitOf	OrgUnit	isUnitOf	OrgUnit	hasUnit	The unit related to an organization.
OrgUnitProgram	OrgUnit	hasProgram	Program	isProgramOf	The program related to an organization.
OrgUnitProject	OrgUnit	hasProject	Project	isProjectOf	Is a relation beteween Project and OrgUnit.
ThesisSponsorship	OrgUnit	sponsors	Publication	isSponsoredBy	Is a relation between a publication of type thesis and a sponsor OrgUnit.
CourseOrgUnit	OrgUnit	hasCourse	Course	isProvidedBy	The OrgUnit(s) that provided the Course.
Authorship	Publication	hasCreator	Person	isCreatorOf	The author of this content or rating.
Advisoring	Publication	hasAdvisor	Person	isAdvisorOf	The advisor of this content or rating.
CoAdvisoring	Publication	hasCoAdvisor	Person	isCoAdvisorOf	The coadvisor of this content or rating.
Referee	Publication	hasRefeee	Person	isRefereeOf	The advisor of this content or rating.
Publisher	Publication	hasPublisher	Service	isPublishedIn	The editor of this content or rating.
PartOf	Publication	hasPartOf	Publication	isPartOf	The publisher of this content or rating.
ProgramThesis	Publication	hasPublication	Program	isProgramOf	Is a relation beteween two publications.
SeniceOrgUnit	Service	hasOrgUnit	OrgUnit	isOrgUnitOf	The publication (thesis) related to an program.
Editoring	Service	hasEditor	Person	isEditorOf	The Course that is associated with a thesis (publication).
CourseThesis	Course	isAssociatedTo	Publication	hasAssociationWith	Is a relation between Patent and Person.
Inventor	Patent	hasInventor	Person	isInventorOf	Is a relation related to an OrgUnit of Patent.
PatentOrgUnit	Patent	isPatent	OrgUnit	hasPatent	Is a relation between Project consortion and member.

Fig. 3. Relações do Descritivo.

Como pode ser observado, o descritivo de uma Entidade apresenta inicialmente o atributo definido no Modelo de Dados bem como sua respectiva descrição. Logo, para cada conjunto de dados que é fonte de informações para a Entidade, são descritos os atributos dos conjuntos de dados relacionados aos do modelo.

Para o tratamento dos dados foi desenvolvida uma biblioteca na linguagem de programação Python, contendo uma estrutura de dados preparada para facilitar o processamento de dados originários de todas as fontes para o formato exigido pela plataforma LA Referencia. Logo, a biblioteca desenvolvida é responsável por toda a transformação e exportação dos dados, utilizando como base o "Modelo de Dados" da plataforma LA Referencia, validando as entidades, campos e relacionamentos aceitos pelo modelo.

Além disso, a biblioteca desenvolvida também é responsável por gerar Identificadores BrCris, criados com o intuito de realizar uma pré-desambiguação dos dados, evitando entidades duplicadas na plataforma LA Referencia. A geração do Identificador BrCris é realizada de forma distinta para cada entidade, em geral realizando um hash de seus próprios campos de dados, como por exemplo nas entidades:

• OrgUnit: hash a partir da concatenação do nome da organização e nome da cidade de localização;

• Publication: hash a partir da concatenação do título, tipo e ano da publicação;

• Program: hash a partir da concatenação do nome do programa de pós-graduação, instituição e cidade de localização;

• Service: hash a partir da concatenação do país e do nome da editora de um periódico científico.

Por fim, o ferramental proposto possibilita realizar a exportação dos dados originais no formato XML (*Extensible Markup Language*), no padrão que será importado pela plataforma LA Referencia.

III. COLETA DOS DADOS

Inicialmente, foram elencados os principais repositórios de dados que seriam fonte de informações para o BrCris. Diversos critérios foram adotados para a seleção dos repositórios a serem utilizados, dentre eles: a consistência e atualização dos dados, o acesso aberto aos conjuntos de interesse, a amplitude dos repositórios e o reconhecimento dos dados pela comunidade científica brasileira. Como resultado, são agregados diversos repositórios, em diferentes formatos e com características distintas.

O principal repositório de dados para o BrCris são os currículos cadastrados na Plataforma Lattes do CNPq. Acessível em < http://lattes.cnpq.br/ >, a Plataforma Lattes foi criada e é mantida pelo CNPq, contando atualmente com mais de 7 milhões de currículos cadastrados (em 15/04/2021), além de grupos de pesquisa e diretórios de instituições. Os currículos de interesse para o BrCris são aqueles dos pesquisadores, assistentes, técnicos que têm grau de mestre ou doutor, ou que possuem algum tipo de relação com a pesquisa científica (ou tecnológica), dado que possuem a publicação de artigo, compartilhamento de conjunto de dados, ou depósito de patente, ou têm alguma relação com a pós-graduação brasileira, seja como discente ou docente. São estimados, aproximadamente, 2,5 milhões de currículos de interesse.

De acordo com Lane [2], em artigo publicado na revista Nature, a Plataforma Lattes é um poderoso exemplo de boas práticas para fornecimento de dados de alta qualidade. A autora relata também que órgãos federais, instituições e órgãos financiadores são incentivadores assíduos desta plataforma e que ela é uma das fontes de dados de pesquisadores mais confiáveis existente.

Além dos dados curriculares da Plataforma Lattes que subsidiaram informações para as entidades Person, OrgUnit, Publication, Patent, Event, Program, Course e Service, também são integrados dados dos seguintes repositórios:

• OasisBR: mantido pelo IBICT, fornece dados confiáveis sobre publicações científicas em acesso aberto. Os dados foram mapeados para as entidades Publication, Service, Person.

• BDTD: a exemplo do OasisBR, a BDTD também é mantida pelo IBICT. Fornece dados confiáveis sobre teses e dissertações brasileiras. Os dados foram mapeados para as entidades Publication, Course e Person.

• Plataforma Sucupira: concentra dados dos Programas de Pós-graduação do Brasil, fornecendo um conjunto de informações sobre os programas e cursos de pós-graduação. Todos os dados dos programas foram mapeados para as entidades Program , Course e OrgUnit;

• Instituições do INEP: assim como os programas de pós-graduação da Plataforma Sucupira, o INEP fornece uma base confiável, sobre as instituições de ensino do país em outros níveis de capacitação, sendo mapeados para a entidade OrgUnit.

• Dados Abertos da CAPES: fornece dados como publicações, orientações, entre outros que são mapeados para diversas entidades, como Person, OrgUnit, Publication, Program e Course.

• Revistas Científicas: o processamento de dados do conjunto de revistas científicas fornece informações diversas sobre elas, sendo mapeadas para a entidade Service. Exemplo de fonte de dados das revistas científicas são:

- Diadorim
- Latindex
- DOAJ
- UlrichsWeb

Como pode ser observado, as diversas fontes de dados mapeadas, se completam, possibilitando a criação de um conjunto padronizado e consistente, validado através de dados provenientes de diversas entidades brasileiras amplamente consolidadas e utilizadas. Ao se agregar todos os repositórios apresentados, é possível a adoção de técnicas que visam permitir a vinculação de conjuntos que inicialmente não era possíveis de se comunicarem, possibilitando dessa forma, a construção de um grande conjunto de dados, interligados, que facilitam a aplicação de consultas que inicialmente não seriam possíveis.

Diante do exposto, é possível verificar todo o conjunto de dados agregados no BrCris e também como estes foram selecionados e processados para uma integração, que engloba alguns elementos para possíveis desambiguações, utilizando para isso, identificadores implícitos ou gerados no processo de tratamento dos dados. Logo, com os dados agregados, diversas análises são viabilizadas, proporcionando um maior conjunto de análises bibliométricas.

IV. RESULTADOS E DISCUSSÃO

Os resultados da execução do Projeto já incluem o desenvolvimento da arquitetura do BrCris (ver Figura 4), o mapeamento das fontes de dados a serem agregadas pelo Sistema, a implementação de provas de agregação dos recursos mapeados, a definição e realização de testes de serviços a serem disponibilizados. Entre as fontes agregadas, destacam-se em âmbito nacional, o OasisBr, a BDTD, a Plataforma Lattes, a Plataforma Sucupira e o Portal de Dados Abertos da CAPES. Já entre as fontes internacionais, destaque é dado ao OpenAIRE Research Graph [3], DOIBoost [4], Portal Wikidata e ao DOAJ.

Fig. 4. Arquitetura do BrCris

Como esquematiza a Figura 4, a plataforma LA Referencia, um software desenvolvido de forma colaborativa, no âmbito da Rede LA Referencia (https://www.lareferencia.info), surge como fonte agregadora e organizadora da informação coletada das diversas fontes mapeadas em âmbito nacional e internacional. A plataforma LA Referencia é flexível quanto ao modelo de dados adotado para armazenamento das informações. Para consulta externa, os dados coletados no âmbito do projeto BrCris serão também representados em um modelo semântico baseado na Ontologia VIVO, um modelo bastante utilizado internacionalmente por diversos CRISs. A representação em RDF com base em uma ontologia permite que o BrCris disponibilize Linked Open Data (LOD), tornando os dados não só abertos, mas também acessíveis e interoperáveis, seguindo, desta forma, as boas práticas disciplinadas pelos princípios FAIR [5].

Todas as ações de mapeio, coleta, transformação entre formatos e carga na plataforma LA Referencia são executadas por módulos desenvolvidos em Python, o qual foi denominado, Módulo BrCris, que se caracteriza como o orquestrador do Sistema. Dos dados organizados na plataforma LA Referencia, é possível indexar as informações agregadas no motor de busca Elasticsearch . Esta ferramenta é um mecanismo de busca de texto completo de código aberto escrito em Java que foi projetado para receber grandes volumes de dados além de ser distributivo e escalável.

Para efeito de testes, já foram dadas cargas de dados e a partir destas cargas, foram disponibilizados os primeiros dashboards na ferramenta Kibana, plugada automaticamente aos índices do Elasticsearch, para que outros pesquisadores possam realizar análises dos dados importados sem a necessidade de softwares fechados ou que dependam de licenças de uso proprietárias ou pagas (ver exemplo nas Figuras 5 e 6).

Fig. 5. Dashboard com as Teses Defendidas entre 1987 a 2018

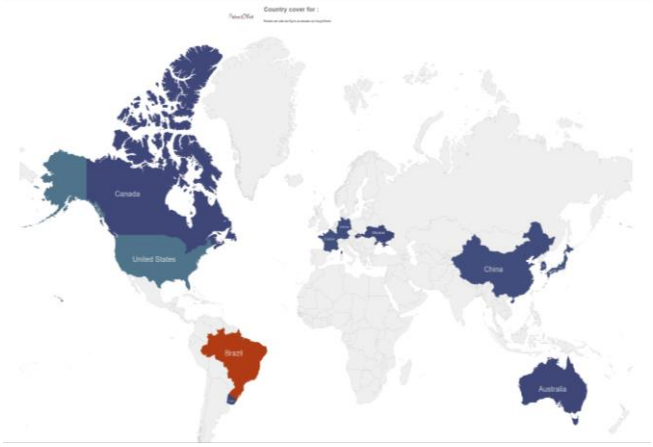

Fig. 6. Dashboard com distribuições geográficas

V. CONSIDERAÇÕES

O BrCris se configura como um importante espaço de pesquisa e análise de dados. As informações agregadas e organizadas segundo um modelo de dados semântico, permitem a geração de serviços para diversos atores, nos contextos de gestão e pesquisa acadêmica, assim como na área de informação para a inovação, que pretende ser o alvo da proposta apresentada. O BrCris é uma iniciativa que coleta e enriquece dados de repositórios e bases de dados abertas pela LA Referencia, utilizando protocolos OAI-PMH e múltiplos formatos de dados em XML e JSON. A próxima etapa do projeto é a aplicação dos sistemas de recomendações pelas métricas que podem ser explanadas em cada conjunto de dados.

Referências

1] Sivertsen, G. (2019) Developing Current Research Information Systems (CRIS) as data sources for studies of research. Springer handbook of science and technology indicators, p. 667-683.

[2] Lane, J. (2010), Let's make science metrics more scientific. Nature, v. 464, n. 7288, p. 488-489.

[3] Openaire. (2021), OpenAIRE Research Graph. Disponível em: https://graph.openaire.eu. Acesso em: 10 abril. 2021.

[4] La Bruzzo, S., Manghi, P., Mannocci, A. (2019), OpenAIRE's DOIBoost - Boosting CrossRef for Research. In: Manghi P., Candela L., Silvello G. (eds) Digital Libraries: Supporting Open Science. Communications in Computer and Information Science, vol 988.

[5] Wilkinson, M. D. et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, v. 3, n. 1, p. 1-9.