A informação patentária no contexto do BrCris

Thiago Magela Dias[1], Adilson Luiz Pinto[2], Jesús Pascual Mena-Chalco[3], Washington Luís Ribeiro de Carvalho Segundo[4], Josir Cardoso Gomes[5], Raulivan Rodrigo da Silva[6] e Luc Quoniam[7]

1 Introdução

CRIS é um acrônimo para *Current Research Information System*. Desde 2014 o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) vem desenvolvendo ações no sentido da construção de um CRIS nacional, o qual foi denominado BrCris.

Esse sistema 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).

1 Docente do Centro Federal de Educação Tecnológica de Minas Gerais, Doutor em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais.

2 Docente da Universidade Federal de Santa Catarina, Doutor em Documentación pela Universidad Carlos III de Madrid (Espanha).

3 Docente da Universidade Federal do ABC, Doutor em Ciência da Computação pela Universidade de São Paulo e pelo Instituto Nacional de Matemática Pura e Aplicada, em regime de cotutela.

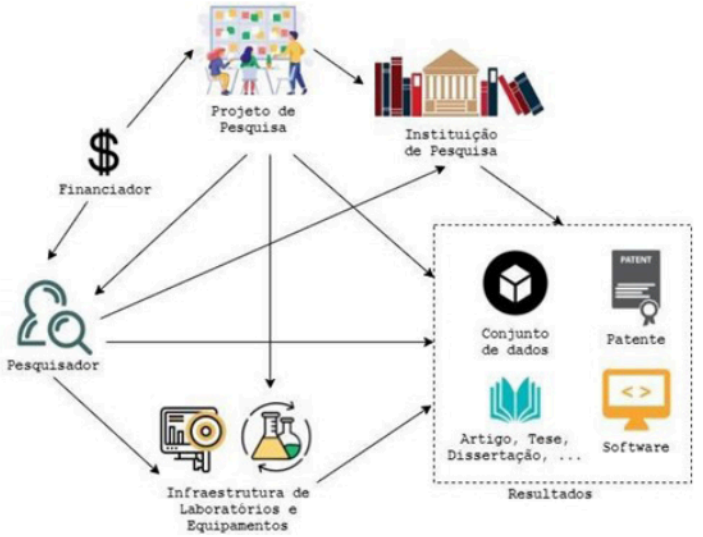
4 Coordenador do Laboratório de Metodologias de Tratamento e Disseminação da Informação do Instituto Brasileiro de Informação em Ciência e Tecnologia, Doutor em Informática pela Universidade de Brasília.

5 Diretor do Instituto RDX de Ensino, Doutorando em Ciência da Informação pelo Instituto Brasileiro de Informação em Ciência e Tecnologia e Universidade Federal do Rio de Janeiro.

6 Docente do Centro Federal de Educação Tecnológica de Minas Gerais, Mestrando em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais.

7 Professor Visitante da Universidade Federal de Mato Grosso do Sul, Doutor em Sciences de l´Information et de la Communication pela Université Aix Marseille III.

Figura 1 – Ecossistema da pesquisa científica



Fonte: Elaborado pelos autores.

A informação disponível para a construção do sistema BrCris está fornecida principalmente pelos repositórios institucionais, bibliotecas digitais de teses e dissertações, revistas eletrônicas de acesso aberto e repositórios de dados de pesquisa brasileiros, reunidos nos portais da BDTD (IBICT, 2021a) e do OasisBr (IBICT, 2021b). Além destas fontes, tem-se como tributária nacional de relevo a Plataforma Lattes (CNPq, 2021) assim como a Plataforma Sucupira (CAPES, 2021a) e o Portal de Dados Abertos da Capes (CAPES, 2021b) que reúnem informação sobre a avaliação dos programas de pós-graduação nacionais.

Está também em desenvolvimento no âmbito do projeto a Plataforma de Instituições de Ciência, Tecnologia e Inovação (PICTI), que se configura também como importante fonte de informação organizada das instituições de pesquisa que devem constar do BrCris. Também são agregadas informações de fontes internacionais abertas, tais como: OpenAIRE (OPENAIRE, 2021a), Wikidata (WIKIDATA, 2021) e Espacenet (EPO, 2021).

Com respeito ao modelo de dados do BrCris, iniciou-se pela adoção de dez entidades ao modelo de dados estabelecido. São elas: (i) *Project*: projetos de pesquisa executados, ou em execução; (ii) *Service*: repositórios digitais, bibliotecas digitais e outras fontes de informação científica; (iii) *Journal*: revistas científicas; (iv) *Graduate Program*: programas de pós-graduação brasileiros; (v) *Course*: cursos nacionais, ou internacionais de pós-graduação *stricto* ou *lato sensu*; (vi) *OrgUnit*:

instituições, faculdades, departamentos de pesquisa; (vii) *Person*: pesquisadores, assistentes de pesquisa e pessoas de apoio técnico à pesquisa; (viii) *Patent*: patentes como resultado da pesquisa; (ix) *Dataset*: conjuntos de dados de pesquisa coletados por pesquisadores e demais agentes no âmbito de um projeto ou pesquisa científica; (x) *Publication*: artigos científicos, teses, dissertações, livros, capítulos de livro e relatórios científicos. É importante ressaltar que a escolha das entidades a serem desenvolvidas foi motivada pelas *OpenAIRE Guidelines for CRIS Managers*, versão 1.1.1 (OPENAIRE, 2021b). Estas por sua vez tomam como base o esquema CERIF de descrição (EUROCRIS, 2021). No entanto, o modelo semântico em desenvolvimento do BrCris toma como base, principalmente, a *VIVO Ontology*, versão 1.11 (LIRASYS, 2021).

Observa-se neste contexto, que os dados sobre patentes depositadas por pesquisadores brasileiros são de difícil recuperação. Em contrapartida, as declarações realizadas por pesquisadores associadas ao produto patente, nos currículos da Plataforma Lattes, mostram-se fonte rica de análise, ainda que os dados pertencentes a esta plataforma sejam autodeclaratórios. Propõe-se a certificação dos dados de patentes do Currículo Lattes, por meio do acesso livre à *Application Programming Interface* (API) do Portal Espacenet, do Ofício Europeu de Patentes, o qual possui dados de mais de 120 milhões de patentes, de 105 países diferentes, incluindo os registros do Brasil. A vinculação e certificação entre Plataforma Lattes e Espacenet permitem a validação das informações declaradas no primeiro e geração de análises sobre Índice de Valor de Patentes tal como realizado especificamente para Patentes Verdes Brasileiras (CATIVELLI *et al.*, 2021).

Essa pesquisa, portanto, tem como objetivo a organização dos registros de patentes do Currículo Lattes, validados contra o Portal Espacenet, no modelo de dados do BrCris. Uma meta conseguinte é o desenvolvimento de métricas e indicadores sobre a massa de dados estabelecida.

2 Metodologia

Desenvolveu-se um modelo inicial de descrição de patentes, que permitirá, além de ter a descrição desta patente, tentar aproximar uma avaliação do potencial inovador desta. A lista a seguir apresenta os atributos selecionados:

a) Identificador interno ao BrCris, identificador Espacenet.
b) *Kind code* (tipo de registro: modelo de utilidade / invenção).
c) Título na Espacenet, título no Currículo Lattes.
d) Data de depósito e data de publicação.
e) Código do país designado pela patente.

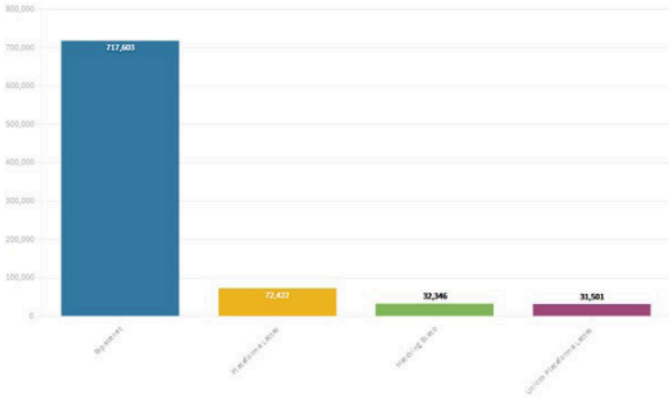
f) Inventor (relação entre as entidades *Patent* e *Person,* pessoa física).
g) *Applicant* (relação entre as entidades *Patent* e *OrgUnit,* pessoa jurídica).
h) Classificação segundo código IPC.
i) Classificação segundo código CPC.
j) *Abstrac*t obtido do Currículo Lattes e *Abstract* obtido na Espacenet.
k) Família: todas as patentes ligado à mesma invenção em vários países ou graus de inventividade (invenção ou aplicação).
l) Referências bibliográficas (relação entre as entidades *Patent* e *Publication*).
m) Referências patentárias (relação da entidade *Patent* consigo mesma).
n) Citações recebidas (relação da entidade *Patent* consigo mesma).

O modelo de dados é implementado em uma instância local da Plataforma LA Referencia, em uma infraestrutura desenvolvida dentro dos servidores hospedados fisicamente no datacenter do IBICT, em Brasília-DF. Os dados são então coletados, armazenados e organizados de acordo com o modelo estabelecido. Permite-se, desta forma, a geração de um modelo semântico de organização de todas as entidades abordadas no BrCris, com vocabulários semânticos específicos, tais como o BIBO, acrescidos de atributos gerados especificamente para o BrCris.

Após realização de coleta dos currículos via Extrator Lattes, identificou-se um conjunto declarado de 72.422 patentes cadastradas no Lattes (inclusive duplicadas), das quais 5.593 foram depositadas fora do Brasil, e onde 32.346 patentes brasileiras foram encontradas na base Espacenet (correspondência exata dos números de patente). Observa-se ainda que, este número de 32.346 registros representa apenas 4,5% dos 717.603 pedidos de patente brasileiras existentes na base Espacenet (ver Figura 2).

Para poder coletar a informação destas patentes, na base Espacenet, é necessário o número internacional da patente, que serve como identificador único (semelhante ao DOI adotado nas publicações). Mas devido à pouca consistência do registro da patente no ato declarativo da parte dos inventores no CV Lattes, é necessário recorrer a diversas estratégias para se encontrar a correspondência entre os números de patentes presentes no Currículo e os números de patentes relacionados nas bases internacionais.

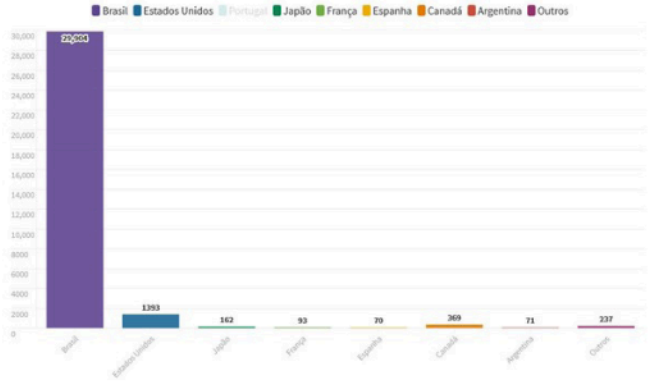
Figura 2 – Quantitativo de patentes



Fonte: Elaborado pelos autores.

Do conjunto de patentes únicas registradas nos currículos da Plataforma Lattes e que foram recuperadas no diretório da Espacenet, consta também um conjunto de patentes internacionais. Este conjunto de patentes pode proporcionar diversos indicadores para uma compreensão sobre a colaboração ou visibilidade internacional dos trabalhos técnicos desenvolvidos por brasileiros (ver Figura 3).

Figura 3 – Internacionalização das Patentes dos Currículos da Plataforma Lattes



Fonte: Elaborado pelos autores.

A coleta destas patentes, através da API OPS (*Open Patent Service*) permite *a priori* a realização de algumas análises, bem como, verificar a proporção das patentes incluídas numa determinada família. Com o conjunto de dados recuperados da Espacenet, diversas validações, bem como atualização dos registros contidos nos

currículos da Plataforma Lattes, podem ser realizadas. Por estar em um formato de dados padronizado, a criação de *scripts* para leitura e tratamento dos dados é facilitada (ver Figura 4).

Figura 4 – Fragmento de um registro extraído da Espacenet


```
},
{
    "@country": "BR",
    "@doc-number": "0200325",
    "@family-id": "28047928",
    "@kind": "B1",
    "@system": "ops.epo.org",
    "bibliographic-data": {
        "application-reference": {
            "@doc-id": "3482190",
            "document-id": [
                {
                    "@document-id-type": "docdb",
                    "country": {
                        "$": "BR"
                    },
                    "doc-number": {
                        "$": "0200325"
                    },
                    "kind": {
                        "$": "A"
                    }
                },
                {
                    "@document-id-type": "epodoc",
                    "date": {
                        "$": "20020125"
                    },
                    "doc-number": {
                        "$": "BR20020200325"
                    }
                },
```


Fonte: Elaborado pelos autores.

Dentre os dados recuperados da Espacenet, informações como a família da patente e de suas citações são de extrema importância nos processo de análise dos dados. Informações estas que não estão presentes nos currículos da Plataforma Lattes. Logo, ao se utilizar destes dados, diversos indicadores ou propostas de classificações poderiam ser realizadas *a posteriori*.

3 Aplicação de algoritmos de *matching* aproximado para o reconhecimento do identificador internacional das patentes

Devido ao baixo *casamento* entre os registros declarados no CV Lattes e na base Espacenet, buscou-se o desenvolvimento de estratégias para aumentar o número de associações.

3.1 Número internacional da patente

Um número de patente *completo* é da forma BR102014031740B1, onde as 2 primeiras letras se referem ao país de depósito, e as duas últimas ao *status jurídico* (*kind code*). É necessário notar que o *kind code* de uma patente pode evoluir com o tempo (mas ele pode ser omitido nas buscas posteriores).

Uma busca através da API da Espacenet necessitará um número exato, no mínimo da parte código de país com a parte numérica. Nos números que constam nos currículos Lattes, pode-se ter a esperança da exatidão da parte numérica do número da patente (i.e. 5561111), exceto para as patentes brasileiras, dado que o INPI acrescenta um dígito suplementar na parte numérica.

Contudo, a base *Google Patent*, bem como a base Espacenet (interface web) possuem um sistema que permite buscar as patentes fornecendo-se somente sua parte numérica. Estas bases não permitem, no entanto, a realização de buscas por frações do identificador, que não sejam exatamente a parte numérica. Por esta razão deve-se utilizar uma busca aproximada por outros atributos conhecidos da patente. Uma estratégia possível é a busca por título, com a desambiguação realizada por meio da medida de distância de *Levenshtein* entre o título procurado e a lista de títulos das patentes recuperadas. Ocorre que geralmente os títulos das patentes estão escritos em inglês, quanto os títulos das patentes brasileiras antigas podem estar, ou não, traduzidos para este idioma na base internacional. Então, antes de se comparar os títulos de patentes, é necessário se detectar o idioma destes, e em caso de necessidade, recorrer-se a uma tradução automatizada para que a comparação seja realizada sempre entre títulos no mesmo idioma.

Assim foi determinado que o número 5561111, se referia a patente US5561111A. Os testes preliminares desta estratégia são promissores e permitem esperar uma alta taxa de recuperação de *matching* dos números das patentes declarados no CV Lattes com os números presentes nas bases de dados patentárias mundiais, como os da base Espacenet.

Com a recuperação dos números de patentes exatos, é possível utilizar a Web-Service da base Espacenet (base do office Europeu de patentes), para construir indicadores patentários. Um teste foi executado com os currículos dos bolsistas de produtividade 1A. Foi utilizado o software Patent2Net (livre, gratuito e de código aberto). A seguir se tem um conjunto de análise dos resultados:

1) É possível analisar as estratégias de "conquista dos territórios", com invenções e inovações, com o depósito inicial (patente prioritário). Este dado pode ser também analisado na forma tabular para enfatizar as evoluções temporais (Figuras 5 e 6);
2) Deixa analisar as estratégias de "conquista dos territórios" com invenções e inovações, usando o depósito em vários países e fazendo um "zoom", em uma família (Figura 7);
3) Exibe-se a importância das instituições que suportam o depósito e manutenção das patentes. São os pesquisadores que de fato realizam o trabalho,

e raramente, hoje em dia, o trabalho individual basta. Observa-se as colaborações entre pesquisadores, como formação estratégica, no âmbito da gestão dos laboratórios (Figuras 9, 10 e 11);

4) Pode ser analisado quais são as referências e quais as citações obtidas pelas patentes (Figuras 12 e 13);
5) Obtêm-se a clusterização textual em torno de um tema buscado (Figura 14).

Figura 5 – País de depósito das patentes

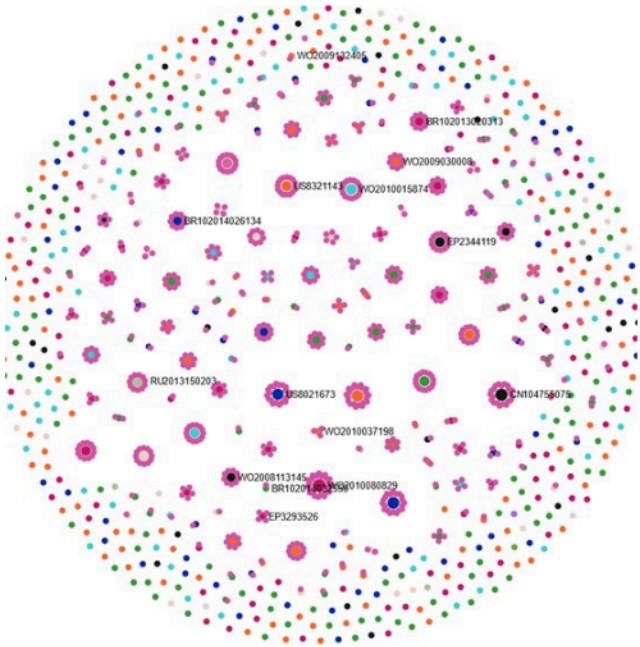
Fonte: Elaborado pelos autores.

Figura 6 – Evolução temporal das patentes por país

country	BR	WO	US	EP	JP	SU	FR	CN	DE	KR	CA	UY	AU	SK	UA	Totaux
year																
2018	86		1	1				1	2							91
2016	79	2		1				1		1						84
2017	62	2	4											1		69
2015	62		1	1	2			1			1					68
2019	59		1													60
2014	33	4	4	1						1		1				44
2020	36	1	1													38
2013	16	5	3													24
2011	5	9	5	1	1											21
2010	10	5	2													17
2012	6	4	4	1	2											16
2009	7	6	1										1			15
2021	15															15
2008	6	4	1													11
2006	8	1	1													10
2000	4		1				1							1		7
2007	3	3														6
2001	3		1				1									5
2004	3	2														5
1999	1					2	1									4
2002	1			1												2
1983					1											1
1998							1									1
1995									1							1
2005		1														1
1991			1													1
1979						1										1
1990						1										1
2003			1													1
1996			1													1
1992	1															1
1980						1										1
Totaux	470	49	34	7	6	5	4	3	3	2	1	1	1	1	1	631

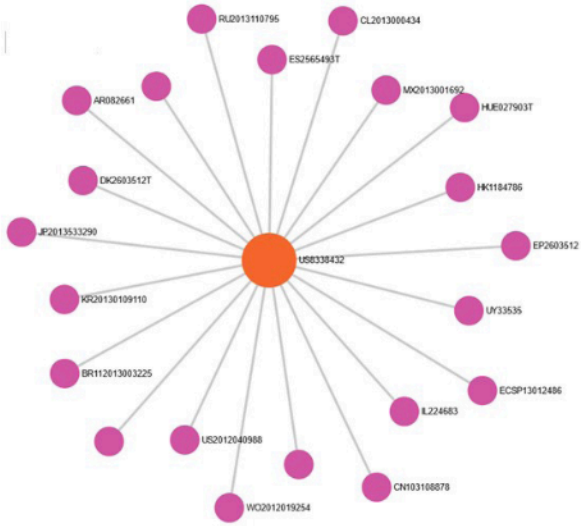
Fonte: Elaborado pelos autores.

Figura 7 – Extensão das patentes em famílias de patentes



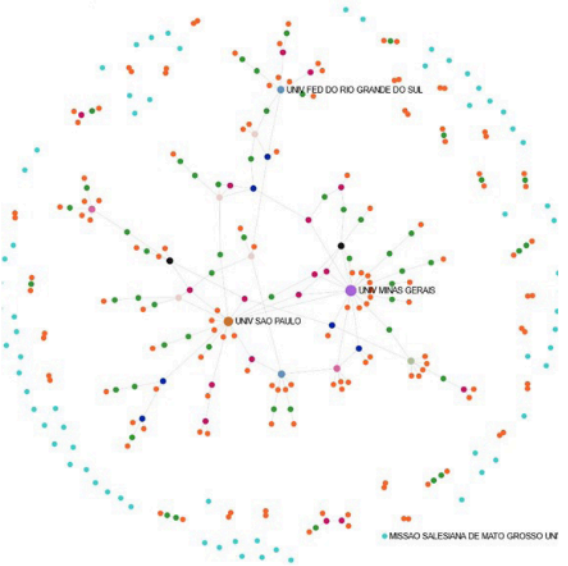
Fonte: Elaborado pelos autores.

Figura 8 – Representação da família da patente US8338432



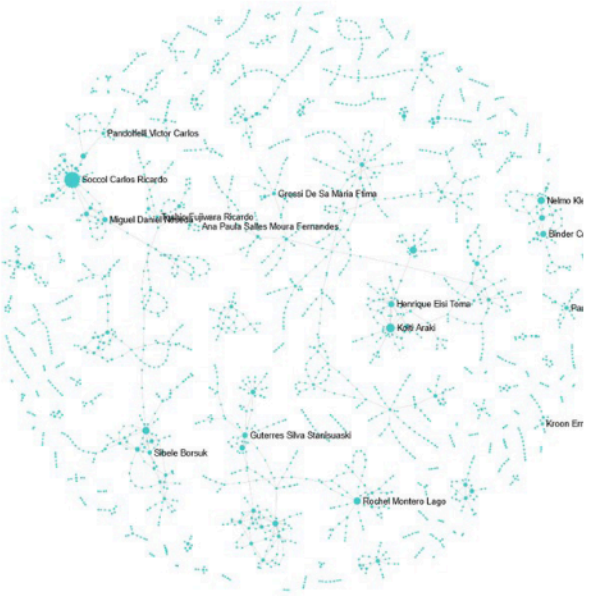
Fonte: Elaborado pelos autores.

Figura 9 – Cooperação ao nível institucional



Fonte: Elaborado pelos autores.

Figura 10 – Rede de colaborações de pessoas físicas



Fonte: Elaborado pelos autores.

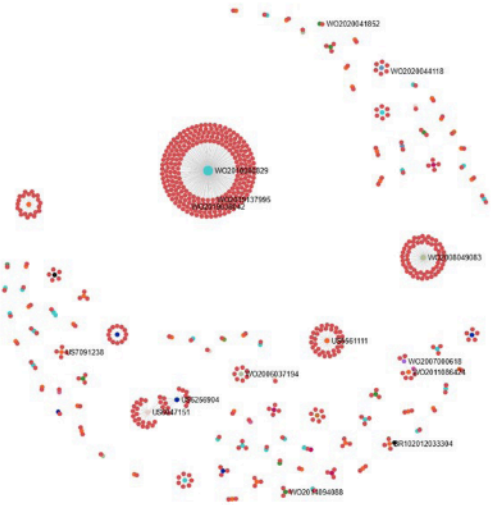
Figura 11 – Combinando-se os grafos das Figuras 9 e 10, obtém-se a rede de colaboração de pesquisadores e organizações

Fonte: Elaborado pelos autores.

Figura 12 – Rede de referências das patentes

Fonte: Elaborado pelos autores.

Figura 13 – Rede de citações recebidas pelas patentes



Fonte: Elaborado pelos autores.

Figura 14 – *Cluster analysis* sobre o conteúdo textual das patentes - patentes ligadas à farmacologia

Fonte: Elaborado pelos autores.

4 Resultados

Como resultado principal, busca-se a certificação de declarações de patentes nos Currículos dos pesquisadores, sendo o código informado, exato ao existente no Portal da Espacenet, ou aproximado, cujo casamento é obtido pela busca utilizando-se as estratégias já descritas.

Para além da certificação, tem-se as opções de visualização da informação agregada, nas classificações de grandes áreas e áreas dos pesquisadores descritos como inventores nos registros de patentes. Visualizações do tipo *nuvem de tags* são adequadas à verificação de frequência das áreas e palavras-chave envolvidas.

A evolução temporal, quantidade por ano, de patentes e a distribuição espacial dos pesquisadores/instituições, requerentes das patentes são também alvo para a construção de *dashboards* de visualização. Uma atenção particular deve ser dada a análise das classificações patentárias IPC e CPC, por ser obrigatórias em qualquer patente, por ser independente da língua (adequa-se para patente de qualquer país de origem) e, enfim, pela OMPI fornecer uma API reindexando automaticamente qualquer texto com essas classificações, fornecendo assim uma verdadeira passarela potencial entre ciência (publicações, teses) e tecnologia (áreas patentárias).

Ainda em relação às possíveis análises sobre a massa de dados agregada, há que se estabelecer índices de valor de patentes, observando-se, diferentes fatores: 1) se uma dada patente é classificada como *triádica*, o que significa que possui registros nos escritórios de patentes do Estados Unidos, Europa e Japão; 2) se pertence a uma estratégia territorial (regional tipo Mercosul, África, Ásia ou estritamente local tipo Brasil); 3) se possui citações realizadas por outras patentes; 4) se ela possui um forte embasamento teórico devido a extensas referências patentária e bibliográfica; 5) se faz parte de uma família de patentes que caracteriza uma *invenção*.

Por fim, tem-se que a base de dados estabelecida deve ser utilizada também para a pesquisa sobre patentes que não possuem proteção no Brasil, ou cuja proteção não está vigente por várias razões (uma sendo a falta de pagamento das anuidades), todas razões que tornam o conhecimento da patente de domínio público e que portanto poderiam ser exploradas tecnologicamente ou comercialmente sem obrigação de pagamento de *royalties*.

Há ainda a pesquisa sobre patentes que possam ser aplicadas à solução de problemas técnicos, transferência de tecnologia, com baixo custo na realização ou por finalidade. Esses casos caracterizam o que se denomina por *Inovação Frugal* (BHATTI, 2012). As patentes podem constituir uma verdadeira base de respostas técnicas.

5 Considerações finais

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 se pretende ser o alvo da proposta apresentada.

No caso da informação sobre as patentes gerada por pesquisadores brasileiros, tem-se na Plataforma Lattes um espaço rico de análise, que se complementa à certificação buscada na base Espacenet. Há que se observar que a informação sobre patentes existente no Currículo Lattes trata muitas vezes apenas da solicitação de patente, o que não caracteriza o registro propriamente dito. Ainda há o caráter au-

todeclaratório que não exime a ocorrência de não veracidade da informação prestada. Reforça-se portanto o valor da certificação pretendida.

Deste modo, a base sobre patentes de pesquisadores brasileiros criada pelo BrCris gerará serviços que vão desde a certificação, organização e visualização dos dados agregados, até a construção de índices de valor de patentes e a criação de subsídios à inovação frugal.

6 Referências

BHATTI, Y. A. **What is frugal, what is innovation?** Towards a theory of frugal innovation. London: Imperial College London, 2012. (Working Paper)

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR (CAPES). **Portal da Plataforma Sucupira.** Brasília: CAPES, 2012.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR (CAPES). **Portal de Dados Abertos da Capes.** Brasília: CAPES, s/d..

CATIVELLI, A. S.; PINTO, A. L.; SANCHEZ, M. L. L. Patent value index: measuring brazilian green patents based on family size, grant, and backward citation. **Iberoamerican Journal of Science Measurement and Communication**, v. 1, n. 1, p. 4, 2021.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPq). **Plataforma Lattes.** Brasília: CNPQ, s/d.

EUROPEAN PATENT OFFICE (EPO). **Espacenet.** München: EPO, s/d.

EUROPEAN CURRENT RESEARCH INFORMATION SYSTEMS (EUROCRIS). **Common European Research Information Format (CERIF).** The Hague (Holanda): EUROCRIS, s/d.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT). **Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)**. Brasília: IBICT, s/d.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT). **Plataforma de Instituições de Ciência, Tecnologia e Inovação (PICTI)**. Brasília: IBICT, s/d.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT). **Portal Brasileiro de Dados e Publicações Científicas (Oasisbr)**. Brasília: IBICT, s/d.

LIRASYS. **VIVO Ontology**: versão 1.11. Phoenix , s/d.

OPEN ACCESS INFRASTRUCTURE FOR RESEARCH IN EUROPE (OPENAIRE). **Openaire guidelines for CRIS managers**. Oakville (Canadá): OPENAIRE, s/d.

OPEN ACCESS INFRASTRUCTURE FOR RESEARCH IN EUROPE (OPENAIRE). **Portal Openaire**. Oakville (Canadá): OPENAIRE, s/d.

WIKIDATA REPOSITORY. **Portal Wikidata**. Berlin: WIKIDATA, 2012.