# COLETA DE DADOS A PARTIR DOS CURRÍCULOS DA PLATAFORMA LATTES: PROCEDIMENTOS UTILIZADOS NO REPOSITÓRIO INSTITUCIONAL UNESP

**Silvana Aparecida Borsetti Gregorio Vidotti**
Docente do Departamento de Ciência da Informação e do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP), Campus de Marília,. Membro do Grupo Gestor da Política do Repositório Institucional UNESP.
vidotti@reitoria.unesp.br

**Fabrício Silva Assumpção**
Doutorando do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP), Campus de Marília, Membro da Equipe Técnica do Repositório Institucional UNESP.
fabricio@reitoria.unesp.br

**Juliano Benedito Ferreira**
Mestre pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP), Campus de Marília, Membro da Equipe Técnica do Repositório Institucional UNESP.
julianoferreira@reitoria.unesp.br

**Ana Paula Grisoto**
Mestranda do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP), Campus de Marília, Membro da Equipe Técnica do Repositório Institucional UNESP.
grisotoana@reitoria.unesp.br

**Renata Eleuterio da Silva**
Mestre pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP), Campus de Marília, São Paulo, Brasil. Membro da Equipe Técnica do Repositório Institucional UNESP.
renata_silva@marilia.unesp.br

**Vítor Silvério Rodrigues**
Analista de sistemas na Coordenadoria Geral de Bibliotecas (CGB) da Universidade Estadual Paulista (UNESP). Membro da Equipe Técnica do Repositório Institucional UNESP.
vitorsrodrigues@reitoria.unesp.br

**Oberdan Luiz May**
Analista de sistemas na Coordenadoria Geral de Bibliotecas (CGB) da Universidade Estadual Paulista (UNESP). Membro da Equipe Técnica do Repositório Institucional UNESP.
oberdan@reitoria.unesp.br

**Flávia Maria Bastos**
Doutora pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP), Campus de Marília,, . Coordenadora da Coordenadoria Geral de Bibliotecas (CGB) da UNESP e membro do Grupo Gestor da Política do Repositório Institucional UNESP.
fmbastos@reitoria.unesp.br

**Resumo:** O Repositório Institucional UNESP foi povoado principalmente a partir de dados coletados da *Web of Science*, da *Scopus* e da SciELO. No entanto, com essa forma de povoamento, o Repositório não contemplava a produção da Universidade de forma fidedigna, uma vez que essas bases de dados cobrem principalmente as publicações internacionais nas áreas de ciências biológicas e exatas. Partindo da necessidade de contemplar também as publicações não indexadas nessas bases de dados, foram desenvolvidos procedimentos para a utilização dos dados da Plataforma Lattes na criação de registros para inclusão no Repositório. A partir da experiência da UNESP, este trabalho tem por objetivo apresentar os procedimentos desenvolvidos, que estão agrupados em seis etapas: coleta dos currículos, conversão para um formato de importação aceito pelo *DSpace*, remoção dos registros duplicados, verificação dos dados e das licenças, organização dos registros nas coleções e importação no Repositório. Como considerações finais, destaca-se que os procedimentos utilizados, ainda que tenham suas limitações, permitem ao Repositório contemplar a produção da Universidade de maneira mais fidedigna.

**Palavras-chave**: Coleta de dados. Currículos da Plataforma Lattes. Repositório institucional.

# DATA COLLECTION FROM LATTES: PROCEDURES USED IN UNESP INSTITUTIONAL REPOSITORY

**Abstract:** The UNESP Institutional Repository was primarily populated by data collected from Web of Science, Scopus and SciELO. However, this data collection did not allow the Repository to reliably comprise the university's production, since the databases used include mostly international publications on biological and hard sciences. Facing the need of include in the Repository the publications not covered by these databases, we developed some procedures to use data from Platform Lattes (a Brazilian curricula database) in order to create items for import in Repository. In this paper we present these procedures in six steps: data collection, data conversion to a DSpace accepted format, deduplication, checking data and license, organizing records into collections, and importing records into Repository. As conclusions, we highlight that, even with their limitations, these procedures enable the Repository to reliably comprises the university's production.

**Keywords**: Data collection. Platform Lattes (Brazilian curricula database). Institutional repository.

## 1 INTRODUÇÃO

Os repositórios institucionais, aqui entendidos como serviços de informação científica em ambiente digital e interoperável dedicados ao gerenciamento da produção científica e/ou acadêmica de uma instituição (LEITE et al., 2012, p. 7), têm despertado o interesse das universidades, entre outros motivos, por seu potencial para o aumento da visibilidade principalmente das atividades de pesquisa desenvolvidas nessas instituições. Com isso, diversas universidades têm implantado seus repositórios institucionais.

Na Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), o Repositório Institucional[1] foi implantando em 2013 e povoado principalmente a partir de dados coletados da *Web of Science*, da *Scopus* e da SciELO (ASSUMPÇÃO et al., 2014). No entanto, essa forma de povoamento não permitia que o Repositório contemplasse a produção científica da Universidade de forma fidedigna, uma vez que as bases de dados utilizadas cobriam principalmente as publicações internacionais nas áreas de ciências biológicas e exatas.

Partindo da necessidade de contemplar também as publicações não indexadas nessas bases de dados, a Equipe Técnica do Repositório desenvolveu e aplicou um conjunto de procedimentos para a coleta de dados a partir da Plataforma Lattes e para o aproveitamento desses dados na criação de registros para inclusão no Repositório. Partindo da experiência realizada na UNESP, este artigo tem por objetivo apresentar esses procedimentos desenvolvidos pela Equipe. Para tanto, organiza-se em quatro principais partes: a contextualização acerca do Repositório Institucional UNESP, a apresentação dos procedimentos utilizados para a coleta e o aproveitamento dos dados da Plataforma Lattes, a síntese dos resultados alcançados até o momento e as considerações finais.

## 2 IMPLANTAÇÃO DO REPOSITÓRIO INSTITUCIONAL UNESP

Em 2013, com base na necessidade das três universidades estaduais paulistas (UNESP, USP e Unicamp) implantarem o Repositório da Produção Científica do CRUESP[2], foi dado início à implantação de um repositório institucional na UNESP. Entre as primeiras ações para a implantação desse repositório, esteve a criação do Grupo Gestor da Política do Repositório

[1] Disponível em: <http://repositorio.unesp.br>.
[2] Disponível em: <http://www.repositorio.cruesp.sp.gov.br>.

Institucional UNESP (UNIVERSIDADE ESTADUAL PAULISTA, 2013, p. 47) e a definição de uma Equipe Técnica. O Grupo Gestor é composto de representantes das pró-reitorias da Universidade (pesquisa, pós-graduação, graduação, extensão e administração), do Núcleo de Educação à Distância (NEaD), da Assessoria Especial de Planejamento Estratégico (APE) e da Coordenadoria Geral de Bibliotecas (CGB).

A UNESP conta com 34 unidades universitárias (faculdades, institutos e campi experimentais) localizadas em 24 cidades do estado de São Paulo. Considerando essa configuração, o Repositório foi organizado de modo a refletir principalmente a estrutura organizacional da Universidade. Em um primeiro nível, o Repositório foi organizado por tipo de produção: "Produção científica" (para documentos científicos como os artigos, os trabalhos publicados em anais de eventos, as teses, as dissertações, os livros, etc.) e "Produção técnica" (para documentos técnicos como as patentes, por exemplo). Na comunidade "Produção científica" foram criadas subcomunidades para as unidades universitárias e, dentro destas, subcomunidades para os departamentos e programas de pós-graduação; por fim, dentro dessas subcomunidades foram incluídas coleções voltadas aos tipos de documentos (artigos, dissertações, teses, livros, etc.).

O objetivo da Universidade era inaugurar o Repositório Institucional UNESP junto dos repositórios da USP[3], da Unicamp[4] e do CRUESP durante a 4ª Conferência Luso-Brasileira sobre Acesso Aberto (CONFOA), realizada em outubro de 2013 em São Paulo. Para isso, foi definida como meta inicial a inclusão, no Repositório, da produção institucional dos cinco anos anteriores (2008-2012) indexada na base de dados referencial *Web of Science*[5].

Levando em conta o prazo para a inauguração do Repositório, a quantidade de documentos abrangidos nesta meta (cerca de 16.400 documentos sendo principalmente artigos e trabalhos publicados em anais de eventos) e a indisponibilidade de recursos humanos para a inclusão de forma manual, a Equipe Técnica estabeleceu procedimentos que possibilitaram a inclusão de forma automática a partir do reaproveitamento dos dados já existentes na *Web of Science*.

Após o alcance da meta inicial e da inauguração do Repositório, a Equipe Técnica aperfeiçoou os procedimentos inicialmente utilizados e os aplicou no reuso dos dados da base

[3] Disponível em: <http://producao.usp.br>.
[4] Disponível em: <http://unicamp.sibi.usp.br>.
[5] Disponível em: <http://webofknowledge.com>.

de dado referencial *Scopus*[6], dos periódicos publicados no Portal SciELO Brasil[7], e do catálogo da Rede de Bibliotecas da UNESP (ASSUMPÇÃO et al., 2014; VIDOTTI et al., 2015).

A utilização desses procedimentos permitiu a inclusão de cerca de 70 mil itens no Repositório Institucional UNESP durante seu primeiro ano de existência. No entanto, a utilização desses procedimentos com a *Web of Science*, a *Scopus* e a SciELO não estava permitindo ao Repositório representar a produção científica da universidade de forma fidedigna, já que essas bases de dados cobrem, principalmente, as publicações internacionais nas áreas de ciências biológicas. Partindo da necessidade de incluir no Repositório também os artigos publicados em periódicos não indexados nessas bases de dados, a Equipe Técnica adaptou os procedimentos e os aplicou na Plataforma Lattes[8].

A Plataforma Lattes "representa a experiência do [Conselho Nacional de Desenvolvimento Científico e Tecnológico] CNPq na integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações" (CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO, 2015). Atualmente, manter um currículo na Plataforma Lattes é considerado uma exigência para os pesquisadores brasileiros:

> O Currículo Lattes se tornou um padrão nacional no registro da vida pregressa e atual dos estudantes e pesquisadores do país, e é hoje adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do País. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, se tornou elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia. (CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO, 2015).

Na seção seguinte são apresentados os procedimentos para a coleta dos currículos da Plataforma Lattes e para a utilização dos dados desses currículos no povoamento do Repositório Institucional UNESP.

[6] Disponível em: <http://scopus.com>.
[7] Disponível em: <http://scielo.br>.
[8] Disponível em: <http://lattes.cnpq.br>.

## 3 COLETA A UTILIZAÇÃO DOS DADOS DA PLATAFORMA LATTES

Os procedimentos para a coleta dos currículos da Plataforma Lattes e para a utilização dos dados desses currículos no povoamento do Repositório podem ser agrupados em seis etapas: coleta dos dados, conversão, remoção dos registros duplicados, verificação dos dados e das licenças, organização dos registros nas coleções e importação no Repositório. Essas etapas são descritas nos itens seguintes desta seção.

### 3.1 Coleta dos dados

Além de permitir a visualização dos currículos, a Plataforma Lattes permite o *download* deles no formato *Extensible Markup Language* (XML) (Linguagem de marcação extensível). Atualmente, esta opção está disponível no canto superior direito da página de cada currículo. Para reduzir o trabalho manual de *download* dos currículos um-a-um, foram compiladas listas com os endereços permanentes dos currículos dos docentes de cada unidade universitária da UNESP.

As listas foram incluídas em um programa desenvolvido pela Equipe Técnica. A partir dos identificadores dos currículos (códigos numéricos presentes ao final do endereço permanente), o programa acessou cada currículo e realizou o *download* do arquivo XML. Sendo necessário que o operador apenas digitasse o *captcha* (código de segurança para provar que o operador é um humano) para que o *download* fosse autorizado pela Plataforma Lattes.

Cada arquivo XML coletado pelo programa, continha apenas o currículo de um docente. Para facilitar as etapas seguintes, os currículos dos docentes de cada unidade universitária foram agrupados por meio de uma folha de estilo criada com a linguagem *Extensible Stylesheet Language for Transformation* (XSLT) (Linguagem extensível para folhas de estilo de transformação), dando origem a um único arquivo XML por unidade universitária.

## 3.2 Conversão

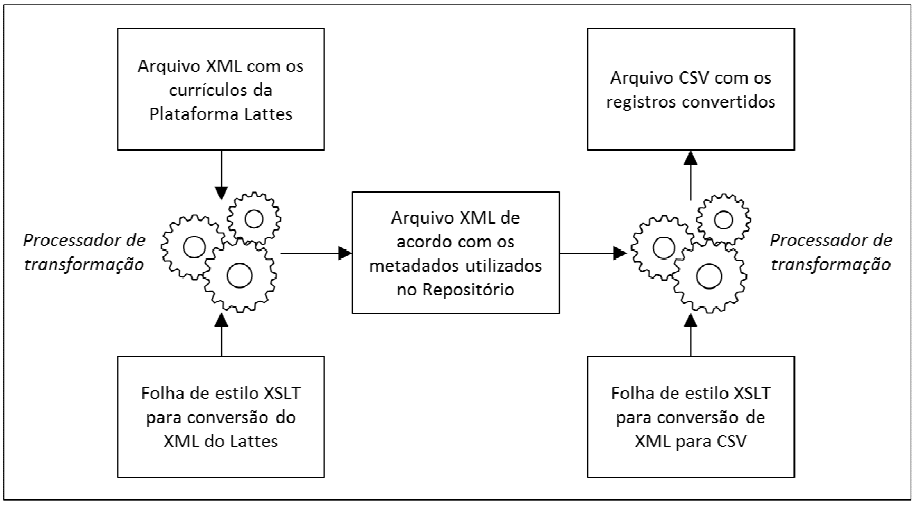
Após a coleta e a junção dos currículos, foi necessário converter os registros presentes dentro deles em registros em um formato de importação aceito pelo *DSpace* e de acordo os metadados utilizados no Repositório.

Entre as diferentes linguagens de programação que poderiam ser utilizadas nessa conversão, foi escolhida a XSLT. Essa escolha deu-se em razão do uso que a Equipe já fazia dessa linguagem para a conversão dos registros de outras fontes (*Web of Science*, SciELO e *Scopus*), como relatado por Assumpção et al. (2014).

A XSLT é uma das tecnologias relacionadas à XML desenvolvidas pelo *World Wide Web Consortium* (W3C) (Consórcio World Wide Web) (W3C, 2007). Essa linguagem contém um conjunto de elementos e de atributos para a criação de regras que, em folhas de estilo, são utilizadas principalmente (1) para converter documentos XML em documentos HTML para apresentação em navegadores e (2) para converter documentos XML criados com uma linguagem de marcação em documentos XML de acordo com outra linguagem de marcação ou em outros formatos, por exemplo, em um formato de texto simples (.txt).

Uma vez que com a XSLT é possível converter um documento XML em outro documento XML ou um documento em outro formato, o fluxo para a conversão dos registros presentes nos currículos foi estabelecido como apresentado na Figura 1.

**Figura 1** – Conversão dos registros coletados da Plataforma Lattes




**Fonte:** Elaborada pelos autores.

O arquivo XML contendo os currículos dos docentes foi incluído no processador de transformação junto da folha de estilo responsável por convertê-lo em um arquivo XML contendo os registros de acordo com os metadados utilizados no Repositório. O processador de transformação é o software responsável por ler e executar as regras da folha de estilo e, a partir delas, gerar um arquivo de saída. O processador de transformação utilizado pela Equipe Técnica foi o *Saxon HE*, disponível no software *Oxygen XML Editor*[9].

O resultado dessa transformação (um arquivo XML contendo os registros de acordo com os metadados utilizados no Repositório) foi, então, incluído no processador de transformação junto de uma segunda folha de estilo, responsável por converter o arquivo XML em um arquivo no formato *Comma-Separated Values* (CSV) (Valores separados por vírgula), aceito para importação no *DSpace*, software utilizado no Repositório. Os registros, em seus três estágios (formato XML da Plataforma Lattes, formato XML com os metadados corretos e formato CSV para importação no *DSpace*), são exemplificados nas Figuras 2, 3 e 4.

**Figura 2** – Registro presente no arquivo XML coletado da Plataforma Lattes

```
<ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="129" ORDEM-IMPORTANCIA="">
    <DADOS-BASICOS-DO-ARTIGO NATUREZA="COMPLETO"
    TITULO-DO-ARTIGO="Relexão e ações na formação e atuação do professor de Língua
    Portuguesa: o diálogo como condição de autoria"
    ANO-DO-ARTIGO="2008" PAIS-DE-PUBLICACAO="" IDIOMA="Português"
    MEIO-DE-DIVULGACAO="IMPRESSO" HOME-PAGE-DO-TRABALHO="" FLAG-RELEVANCIA="NAO"
    DOI="" TITULO-DO-ARTIGO-INGLES="" FLAG-DIVULGACAO-CIENTIFICA="NAO"/>

    <DETALHAMENTO-DO-ARTIGO
    TITULO-DO-PERIODICO-OU-REVISTA="Linguagem &amp;amp; Ensino (UCPel. Impresso)"
    ISSN="14151928" VOLUME="1" FASCICULO="" SERIE="" PAGINA-INICIAL="12"
    PAGINA-FINAL="20" LOCAL-DE-PUBLICACAO=""/>

    <AUTORES NOME-COMPLETO-DO-AUTOR="Regina Aparecida Ribeiro Siqueira"
    NOME-PARA-CITACAO="SIQUEIRA, R. A. R." ORDEM-DE-AUTORIA="2" NRO-ID-CNPQ=
    "66342508689184 33"/>

    <AUTORES NOME-COMPLETO-DO-AUTOR="Rozana Aparecida Lopes Messias"
    NOME-PARA-CITACAO="MESSIAS, R. A. L." ORDEM-DE-AUTORIA="1"/>

    <AREAS-DO-CONHECIMENTO/>

    <INFORMACOES-ADICIONAIS DESCRICAO-INFORMACOES-ADICIONAIS=""
    DESCRICAO-INFORMACOES-ADICIONAIS-INGLES=""/>
</ARTIGO-PUBLICADO>
```

**Fonte:** Elaborada pelos autores.

[9] Disponível em: <http://www.oxygenxml.com>.

**Figura 3** – Registro em XML de acordo com os metadados utilizados no Repositório

```
<dublin_core schema="dc">
   <dcvalue element="contributor" qualifier="author">Siqueira, Regina Aparecida Ribeiro</dcvalue>
   <dcvalue element="contributor" qualifier="author">Messias, Rozana Aparecida Lopes [UNESP]
   </dcvalue>
   <dcvalue element="contributor" qualifier="institution">Universidade Estadual Paulista (UNESP)
   </dcvalue>
   <dcvalue element="date" qualifier="issued">2008</dcvalue>
   <dcvalue element="description" qualifier="extent">12-20</dcvalue>
   <dcvalue element="identifier" qualifier="citation">Linguagem &amp; Ensino, v. 1, p. 12-20, 2008.
   </dcvalue>
   <dcvalue element="identifier" qualifier="issn">1415-1928</dcvalue>
   <dcvalue element="identifier" qualifier="file">ISSN1415-1928-2008-01--12-20.pdf</dcvalue>
   <dcvalue element="identifier" qualifier="lattes">0074027009123924</dcvalue>
   <dcvalue element="identifier" qualifier="lattes">6634250868918433</dcvalue>
   <dcvalue element="language" qualifier="iso">por</dcvalue>
   <dcvalue element="relation" qualifier="ispartof">Linguagem &amp; Ensino</dcvalue>
   <dcvalue element="source">Currículo Lattes</dcvalue>
   <dcvalue element="title" language="pt">Relexão e ações na formação e atuação do professor de
   Língua Portuguesa: o diálogo como condição de autoria</dcvalue>
   <dcvalue element="type">Artigo</dcvalue>
   <dcvalue element="campus">Faculdade de Ciências e Letras (FCLAS)</dcvalue>
   <dcvalue element="department">Educação</dcvalue>
</dublin_core>
```

**Fonte:** Elaborada pelos autores.

**Figura 4** – Registros em um arquivo CSV de acordo com o formato de importação do DSpace

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | id | collection | dc.contributor.au | dc.contributor | dc.date | abstract[pt] | sponsorship[] | extent[] | dc.identifier[] | dc.identifier.cit | dc.title[pt] |
| 2 | + | collection | Siqueira, Regina Aparecida Ribeiro\|\|Messias, Rozana Aparecida Lopes [UNESP] | Universidade Estadual Paulista (UNESP) | 2008 | | | 12-20 | | Linguagem & Ensino, v. 1, p. 12-20, 2008. | Relexão e ações na formação e atuação do professor de Língua Portuguesa: o diálogo como condição de autoria |
| 3 | + | collection | Messias, Rozana Aparecida Lopes [UNESP] | Universidade Estadual Paulista (UNESP) | 2008 | | | 335-352 | http://www.uepg.br/uniletras | Uniletras, v. 30, p. 335-352, 2008. | O Conhecimento pessoal prático nas aulas de Língua Portuguesa: gêneros orais em foco |
| 4 | + | collection | Messias, Rozana Aparecida Lopes [UNESP] | Universidade Estadual Paulista (UNESP) | 2010 | | | 199-220 | http://www.cedap.assis.unesp.br/patrimonio_e_mem | Patrimônio e Memória, v. 6, n. 1, p. 199-220, 2010. | Ensino de língua: pressupostos para a consideração de gêneros orais |

**Fonte:** Elaborada pelos autores.

O arquivo XML de acordo com os metadados utilizados no Repositório (Figura 3) já poderia ser importado no Repositório, pois está em conformidade com um formato de importação aceito pelo *DSpace*. No entanto, optou-se por convertê-lo em um arquivo no formato CSV (Figura 4) para facilitar as etapas seguintes (remoção dos registros duplicados,

verificação dos dados, etc.). Um arquivo CSV é um arquivo semelhante a uma planilha (contem linhas e colunas), sendo que cada coluna representa um metadado e cada linha representa um registro. Os arquivos no formato CSV podem ser visualizados e editados em programas como o *Microsoft Office Excel* e *LibreOffice Calc*, embora este último seja o mais recomendado para a edição de dados para importação no *DSpace*.

A folha de estilo utilizada na primeira conversão, além de transformar o arquivo XML com os currículos em um arquivo XML de acordo com os metadados utilizados no Repositório, permitiu a adequação dos dados e a seleção apenas dos registros de interesse para o Repositório.

Adequação dos dados envolveu, em outros: (1) a transformação dos valores, por exemplo, a transformação da palavra “Português” no código “por” da norma ISO 639-2; (2) a junção de valores, por exemplo, título do periódico, volume, número, ano, etc. foram juntados para compor a referência do documento; (3) e a correção do uso de maiúsculas, por exemplo, nos títulos dos periódicos.

Para a seleção dos registros de interesse foi estabelecido um filtro que permitiu a conversão apenas dos registros referentes aos artigos científicos. A decisão por converter apenas os artigos, descartando, assim, livros, capítulos de livros, trabalhos publicados em anais de eventos, etc., foi tomada considerando, principalmente:

- a importância dos artigos enquanto instrumentos consagrados para a comunicação científica; segundo Macias-Chapula (1998, p. 136) o “artigo de periódico com a sua lista de citações é, e provavelmente assim permanecerá, o meio universalmente aceito pelo qual a instituição científica registra e divulga os resultados de suas investigações”;
- a relevância dos artigos científicos nos repositórios institucionais,
- a disponibilidade dos artigos na Web; os trabalhos publicados em anais de eventos, por exemplo, nem sempre estão disponíveis na Web ou podem ser facilmente localizados, já os livros e seus capítulos, são publicados na maior parte das vezes apenas em formato impresso.

Após a conclusão das conversões, o arquivo no formato CSV resultante foi encaminhado para a etapa de remoção dos registros duplicados, descrita no item seguinte.

### 3.3 Remoção dos registros duplicados

De posse do arquivo CSV contendo os registros convertidos, a Equipe Técnica iniciou a remoção dos registros duplicados. Primeiramente foram removidos os registros duplicados dentro do próprio arquivo CSV. Essas duplicações aconteceram porque, em diversos casos, os artigos têm entre seus autores mais de um docente da UNESP, o que fez com que tais artigos estivessem presentes em mais de um currículo.

Após essa primeira eliminação dos registros duplicados, o arquivo CSV convertido foi comparado com um arquivo CSV contendo todos os registros existentes no Repositório. O objetivo dessa comparação foi remover do arquivo CSV convertido os registros dos artigos que já estavam no Repositório.

Para esses procedimentos foi utilizado um software desenvolvido pela Equipe Técnica para a comparação e a remoção de registros duplicados, sendo que para a identificação desses registros foi utilizado, primeiramente, o *Digital Object Identifier* (DOI) (Identificador de objeto digital) e, em seguida, o título e a data de publicação juntos.

### 3.4 Verificação dos dados e das licenças

O arquivo CSV contendo os registros não duplicados, resultante da etapa anterior, foi encaminhado para a etapa de verificação dos dados e das licenças.

Na verificação dos dados, os registros foram verificados um-a-um com o objetivo de:

- identificar, a partir das informações de afiliação, se realmente eram parte da produção institucional; os artigos em que a UNESP não constava em nenhuma das afiliações foram removidos;
- completar os dados que não puderam ser obtidos a partir dos currículos, por exemplo, as instituições dos autores, as agências de fomento, o resumo, o título em outro idioma; e
- corrigir possíveis erros, por exemplo, ordem dos autores e o endereço correto para o acesso online.

A verificação das licenças consistiu em verificar as permissões de acesso (acesso aberto ou acesso restrito) e de arquivamento (arquivamento da versão final em repositórios

institucionais permitido ou não). Para essa verificação foi utilizado o serviço SHERPA/RoMEO e as políticas dos publicadores das revistas. O SHERPA/RoMEO "é uma base de dados pesquisável de políticas de publicadores relacionadas ao autoarquivamento de artigos de periódicos na web em repositórios de acesso aberto" (SHERPA/RoMEO, 2011, tradução nossa). Nos casos em que o arquivamento da versão final era permitido, uma cópia digital do artigo no formato PDF foi salva e nomeada com um código identificador presente no arquivo CSV.

### 3.5 Organização dos registros nas coleções

Para que os registros convertidos e verificados pudessem ser importados no Repositório e inseridos nas coleções corretas, foi executado um programa que, a partir das informações de afiliação e de autoria, incluiu no campo *collection* de cada registro os códigos "Handle" das coleções às quais ele pertenceria. Nos casos em que o artigo seria adicionado em uma coleção e mapeado para outras, por exemplo, quando um artigo tinha entre seus autores docentes de diferentes departamentos da UNESP, os códigos das coleções eram separados dentro do campo *collection* por duas barras verticais (" || ").

O programa utilizado, desenvolvido pela Equipe Técnica, incluiu os códigos das coleções com base nas regras presentes em arquivos XML. Nesses arquivos XML, foram incluídas as formas variantes do nome da universidade, das unidades universitárias, dos departamentos e dos nomes dos autores.

### 3.6 Importação no Repositório

Uma vez que as ações para a verificação e a preparação dos dados foram concluídas, o arquivo CSV foi importado no Repositório utilizando os procedimentos para a importação de registros em lote no *DSpace* (IMPORTING..., 2015).

Como descrito no item 3.4, durante a verificação das licenças, os artigos cujo arquivamento da versão final em repositórios institucionais era permitido tiveram uma cópia digital no formato PDF salva e nomeada com um código identificador presente no arquivo CSV. Uma vez que o código identificador estava presente tanto no registro importado no Repositório quanto no arquivo PDF, foi possível utilizar um programa, desenvolvido pela

Equipe Técnica, para incluir automaticamente no *DSpace* cada arquivo PDF em seu respectivo registro, poupando, assim, o trabalho manual de *upload* desses arquivos um-a-um no Repositório.

## 4 RESULTADOS ALCANÇADOS

Segundo seu anuário estatístico de 2015 (UNIVERSIDADE ESTADUAL PAULISTA, 2015, p. 3), a UNESP conta com 3.880 docentes, distribuídos em 34 unidades universitárias localizadas em 24 cidades do estado de São Paulo. Algumas unidades contemplam apenas uma área de estudo, por exemplo, as faculdades de odontologia, enquanto outras contemplam diversas áreas, por exemplo, o Instituto de Biociências, Letras e Ciências Exatas (IBILCE).

Considerando essa configuração da Universidade e a disponibilidade de recursos humanos na Equipe Técnica, a coleta dos dados da Plataforma Lattes para o povoamento do Repositório foi pensada para ser executada em uma unidade universitária de cada vez. Além disso, foi definido que, inicialmente, seriam coletados apenas os dados referentes aos artigos publicados nos cinco anos anteriores (2010 a 2014).

Para a condução de um projeto piloto que possibilitasse a verificação e o aperfeiçoamento dos procedimentos esquematizados pela Equipe Técnica, foi utilizada a Faculdade de Filosofia e Ciências (FFC). Essa unidade contava com 180 docentes distribuídos em dez departamentos (Administração e supervisão escolar, Ciência da informação, Ciências políticas e econômicas, Didática, Educação especial, Filosofia, Fisioterapia e terapia Ocupacional, Fonoaudiologia, Psicologia da educação, e Sociologia e antropologia), cuja produção científica é predominantemente da área de ciências humanas.

Os 180 currículos foram coletados no final de janeiro de 2015 e deles puderam ser extraídos 1.701 registros referentes a artigos publicados no período de 2010 a 2014. Após a remoção das duplicações, restaram 1.150 registros, que foram encaminhados para a etapa de verificação dos dados e das licenças e de coleta dos arquivos digitais. A verificação foi realizada pela Equipe Técnica, composta por três bibliotecários, durante cerca de duas semanas e meia. Ao final da verificação, foram removidos os registros que não faziam parte da produção institucional ou eram duplicados mas não foram identificados na etapa de

remoção de registros duplicados. Os 959 registros restantes foram então organizados nas coleções e importados no Repositório junto dos 760 arquivos digitais que puderam ser coletados para o arquivamento.

Antes da importação dos registros obtidos a partir da Plataforma Lattes, as coleções da FFC no Repositório somavam 630 artigos. Após a importação, essa quantidade passou para 1.589, o que representa um aumento de 152%.

Considerando os resultados obtidos com o piloto realizado na FFC, os procedimentos utilizados foram considerados adequados e passaram a integrar o rol de procedimentos já estabelecidos para o povoamento do Repositório. Com isso, foi iniciada a coleta dos currículos dos docentes das demais unidades universitárias da UNESP. Até a data da redação deste trabalho (setembro de 2015) foram coletados, convertidos, verificados e importados os registros referentes aos currículos dos docentes de 6 unidades universitárias. A quantidade de artigos incluída no Repositório a partir da Plataforma Lattes para cada uma dessas unidades universitárias é apresentada na Tabela 1.

**Tabela 1** – Quantidade de artigos incluídos no Repositório a partir da Plataforma Lattes

| Unidade universitária | Quant. de docentes | Quant. de artigos antes da coleta do Lattes | Quant. de artigos coletados do Lattes | Quant. de artigos após a coleta do Lattes | Aumento da quantidade de artigos (%) |
|---|---|---|---|---|---|
| Faculdade de Filosofia e Ciências (FFC) | 180 | 630 | 959 | 1.589 | 152% |
| Instituto de Biociências, Letras e Ciências Exatas de São José do Rio Preto (IBILCE) | 251 | 2.439 | 662 | 3.101 | 26% |
| ***1.*** *Instituto de Química de Araraquara (IQ)* | 116 | 4.376 | 240 | 4.616 | 5,48% |
| Faculdade de Ciências e Letras de Araraquara (FCLAR) | 252 | 1.869 | 889 | 2.758 | 47,56% |
| Faculdade de Odontologia de Araraquara (FOAR) | 125 | 2.946 | 764 | 3.710 | 26% |
| Faculdade de Ciências e Letras de Assis (FCLAS) | 167 | 1.355 | 560 | 1.915 | 41,32% |

**Fonte:** Elaborada pelos autores.

## 5 CONSIDERAÇÕES FINAIS

Nessas considerações finais, dois dos entraves encontrados pela Equipe Técnica durante a realização dos procedimentos merecem destaque: (1) o mau preenchimento dos currículos e (2) a falta de clareza das revistas nacionais sobre as políticas de direitos autorais.

Apesar da importância da Plataforma Lattes enquanto reflexo da produção científica ser reconhecida pelos docentes, observa-se, com uma alta frequência, o preenchimento incorreto dos currículos nessa Plataforma, sendo dois dos erros mais frequentes a ordem incorreta dos autores e os links incorretos. Isso evidencia, entre outros, a necessidade de ações da Universidade para a conscientização e a capacitação para o preenchimento do currículo.

Ainda que a maior parte das revistas nacionais disponibilize seus artigos gratuitamente, nota-se pouca clareza de seus editores acerca das questões de direitos autorais e de acesso aberto. É possível encontrar, por exemplo, revistas sem qualquer menção aos direitos autorais ou com declarações contraditórias, tais como o uso de uma licença *Creative Commons* seguida pela frase “Reprodução proibida” ou “Todos os direitos reservados”.

Esses entraves, multiplicados, por exemplo, por um mil artigos, aumentam consideravelmente o tempo demandado pela Equipe Técnica na preparação dos registros para a importação no Repositório. No entanto, mesmo com esses entraves, o uso dos dados da Plataforma Lattes, assim como ocorre com o uso dos dados da *Web of Science*, da SciELO e da *Scopus*, isentará o docente do esforço de submeter sua produção no Repositório ou enviá-la para uma submissão mediada, sendo que essa isenção do docente é considerada um aspecto importante para o desenvolvimento do Repositório Institucional UNESP.

Embora os resultados alcançados não permitam uma generalização, pode-se destacar que, como estimado pela Equipe Técnica, a utilização dos currículos da Plataforma Lattes mostrou-se vantajosa para as áreas de ciências humanas, cuja produção científica nacional é pouco contemplada nas bases de dados que até então haviam sido utilizadas como fontes de dados. Com isso, entende-se que os procedimentos apresentados neste trabalho estão permitindo ao Repositório contemplar a produção da Universidade de maneira mais fidedigna, compensando a ênfase que até então havia sido dada às publicações das áreas de ciências exatas e biológicas.

Por fim, destaca-se que existem procedimentos e ferramentas para a integração dos repositórios às plataformas de currículos (por exemplo, Lattes e DeGóis) com o objetivo de

aproveitar os dados dessas plataformas no povoamento dos repositórios, no entanto, os procedimentos apresentados neste trabalho foram desenvolvidos considerando que a Universidade no momento não dispunha dos recursos necessários para viabilizar essa integração. Nesse sentido, espera-se que este trabalho contribua com as instituições que buscam, da forma mais automatizada possível e com os recursos disponíveis, ampliar sua visibilidade a partir da inclusão, em seus repositórios, da produção presente nos currículos da Plataforma Lattes, e, como forma de efetivar tal contribuição, a UNESP disponibiliza na plataforma *GitHub*[10] os programas desenvolvidos pela Equipe Técnica apresentados neste trabalho.

## REFERÊNCIAS


ASSUMPÇÃO, F. S. et al. A conversão de registros na implantação de repositórios institucionais: o caso do Repositório Institucional UNESP. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 18., 2014, Belo Horizonte. **Anais...** Belo Horizonte: UFMG, 2014. p. 1-16. Disponível em: <http://hdl.handle.net/11449/123645>. Acesso em: 16 set. 2015.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. **Sobre a Plataforma Lattes**. Brasília, [2015?] Disponível em: <http://www.cnpq.br/web/portal-lattes/sobre-a-plataforma>. Acesso em: 16 set. 2015.

IMPORTING Items via basic bibliographic formats (Endnote, BibTex, RIS, TSV, CSV) and online services (OAI, arXiv, PubMed, CrossRef, CiNii). In: DSPACE 5.x Documentation. DuraSpace, 2015. Disponível em: <https://wiki.duraspace.org/pages/viewpage.action?pageId=45548176>. Acesso em: 16 set. 2015.

LEITE, F. et al. **Boas práticas para a construção de repositórios institucionais da produção científica**. Brasília: Ibict, 2012. Disponível em: <http://livroaberto.ibict.br/handle/1/703>. Acesso em: 16 set. 2015.

MACIAS-CHAPULA, C. A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ciência da Informação**, v. 27, n. 2, p. 134-140, maio/ago.1998. Disponível em: <http://dx.doi.org/10.1590/S0100-19651998000200005>. Acesso em: 16 set. 2015.

SHERPA/RoMEO. **FAQ**: Publisher copyright policies & self-archiving. Nottingham, 2011. Disponível em: <http://www.sherpa.ac.uk/romeo/faq.php>. Acesso em: 21 set. 2015.

UNIVERSIDADE ESTADUAL PAULISTA. **Anuário estatístico 2015**. São Paulo: 2015. Disponível em: <https://ape.unesp.br/anuario/pdf/Anuario_2015.pdf>. Acesso em: 21 set. 2015.

UNIVERSIDADE ESTADUAL PAULISTA. Portaria n.º 88, de 28 de fevereiro de 2013. **Diário Oficial do Estado de São Paulo**, Executivo, São Paulo, 01 mar. 2013. Caderno 1, p. 47.


[10] O programas e folhas de estilo descritos neste trabalho estão disponíveis nos seguintes repositórios do GitHub: https://github.com/fsassumpcao/metadata-conversions-to-dspace, https://github.com/vitorsilverio/Item2CollectionRuler e https://github.com/jaideraf/DSpace-tools.

VIDOTTI, S. A. B. G. et al. Reutilização de metadados para o povoamento de um repositório institucional: procedimentos aplicados no Repositório Institucional UNESP. In: INTERNATIONAL CONFERENCE ON DUBLIN CORE & METADATA APPLICATIONS (DC-2015), 15., 2015, São Paulo. **Proceedings**…, 2015. p. 234-235. Disponível em: <http://hdl.handle.net/11449/127972>16 set. 2015.

W3C. **XSL Transformations (XSLT) Version 2.0**: W3C Recommendation 23 January 2007. Cambridge, 2007.