UNIVERSIDADE FEDERAL DO PARANÁ

ALUÍSIO AUGUSTO SILVA GONÇALVES

UMA REDE DE CITAÇÕES PARA A BIBLIOTECA DIGITAL DE TESES E DISSERTAÇÕES DA UNIVERSIDADE FEDERAL DO PARANÁ

CURITIBA

2017

ALUÍSIO AUGUSTO SILVA GONÇALVES

UMA REDE DE CITAÇÕES PARA A BIBLIOTECA DIGITAL DE TESES E DISSERTAÇÕES DA UNIVERSIDADE FEDERAL DO PARANÁ

Trabalho apresentado como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação no curso de Ciência da Computação, Setor de Ciências Exatas da Universidade Federal do Paraná. Orientador: Prof. Dr. Marcos Didonet Del Fabro

CURITIBA

2017

À G. G. H. Obrigado por perguntar.

AGRADECIMENTOS

Gostaria de agradecer àqueles que me acompanharam durante a graduação e o desenvolvimento deste trabalho, em especial:

- aos meus pais, Celso e Cláudia, sem o suporte dos quais este trabalho não existiria;
- ao meu orientador, Prof. Dr. Marcos Didonet Del Fabro, pela paciência e incentivo;
- à Elisabete Ferreira, que me apresentou ao tema deste trabalho e cujas observações foram fundamentais para a forma que este tomou;
- ao pessoal da Biblioteca Central da UFPR e à Karolayne Costa Rodrigues de Lima, da Biblioteca de Artes, Comunicação e Design da UFPR, por me expor ao mundo da biblioteconomia e da bibliometria;
- aos meus colegas Andressa Schaff Steffens, Carolina Aparecida de Lara Moraes, e Tulio Roberto Dias, com quem enfrentei a jornada de quatro anos que nos levou até aqui.

“What did you dream?
It’s all right, we told you what to dream.
(Roger Waters, *Welcome to the Machine*)

RESUMO

Este trabalho apresenta uma ferramenta para geração de redes de citações a partir de documentos armazenados em um repositório digital, tendo como fim possibilitar a descoberta dos trabalhos mais influentes na Biblioteca Digital de Teses e Dissertações da Universidade Federal do Paraná. Foram avaliadas ferramentas com objetivos similares, mas as mesmas ou operam sobre bases seletivas de documentos ou não suportam a limitação de buscas à documentos produzidos por autores afiliados à uma instituição específica.

A avaliação da ferramenta sobre um subconjunto de 7.275 documentos da Biblioteca de Teses e Dissertações da UFPR revelou um percentual baixo de referências bibliográficas internas à esta biblioteca. Espera-se que melhorias nos parâmetros para extração de referências e ligação entre referências e documentos citados revelem um panorama mais preciso da comunicação científica na Universidade.

Palavras-chave: Bibliometria. Índice de citações. Rede de referências. Repositório digital.

ABSTRACT

This work presents a tool for generating citation networks from documents stored in a digital repository, developed in order to discover the most influential works in the Electronic Theses and Dissertations Library of the Federal University of Paraná. Tools with similar objectives were evaluated, but they either operate with a curated, and therefore limited, *corpus*, or they do not support limiting queries to documents produced by authors affiliated with a given entity.

The evaluation of the tool on a subset of 7,275 documents from the UFPR Theses and Dissertations repository revealed a low percentage of bibliographic references internal to this repository. It is expected that tweaks to the parameters of the reference extraction and linking processes will reveal a more accurate picture of the scientific communication at the University.

Keywords: Bibliometrics. Citation index. Digital repository. Record linkage.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ABNT ..... Associação Brasileira de Normas Técnicas
ACM ..... Association for Computing Machinery
API ..... Application Programming Interface
BDEC ..... Biblioteca Digital de Eventos Científicos
BDP ..... Biblioteca Digital de Periódicos
BDTD ..... Biblioteca Digital de Teses e Dissertações
C3SL ..... Centro de Computação Científica e Software Livre
CERMINE ..... Content Extractor and Miner
CitEc ..... Citations in Economics
CNB ..... Citation Network Builder
CRF ..... Conditional Random Field
CSV ..... Valores Separados por Vírgulas
ERP ..... Sistema Integrado de Gestão Empresarial
ETL ..... Extract, Transform, Load
GNU ..... GNU's not Unix
HMM ..... Hidden Markov Model
HTML ..... Hypertext Markup Language
JATS ..... Jornal Article Tag Suite
JCR ..... Journal Citation Reports
JDBC ..... Java Database Connectivity
JIF ..... Journal Impact Factor
JPA ..... Java Persistence API
JVM ..... Máquina Virtual Java
MIT ..... Massachusetts Institute of Technology
OAI ..... Open Archives Initiative
OAI-ORE ..... Open Archives Initiative Object Reuse and Exchange
OAI-PMH ..... Open Archives Initiative Protocol for Metadata Harvesting
PDF ..... Portable Document Format
RDF ..... Resource Description Framework
RePEc ..... Research Papers in Economics
SCI ..... Science Citation Index
SciELO ..... Scientific Electronic Library Online
SGBD ..... Sistema Gerenciador de Banco de Dados
SiBi ..... Sistema de Bibliotecas
SPI ..... Interface de Provedor de Serviço
SVM ..... Support Vector Machine

TDIDF ..... Term Frequency × Inverse Document Frequency
UFPR ..... Universidade Federal do Paraná
URL ..... Uniform Resource Locator
UTF ..... Formato de Transformação Unicode
XML ..... Extensible Markup Language
YAML ..... YAML Ain't Markup Language

SUMÁRIO

1 INTRODUÇÃO

O número de vezes que um determinado trabalho ou autor é citado por outros é um indicador de seu impacto no campo da pesquisa científica, tanto diretamente como componente de métricas, por exemplo, o índice h (HIRSCH, 2005) e o fator de impacto de periódicos (GARFIELD, 2006). Para obter este número, é preciso identificar quais as obras que fazem referência ao trabalho sendo analisado a partir de suas referências bibliográficas. Ao fazer esta identificação para todos os documentos de um conjunto, obtém-se um índice de citações deste conjunto, o que permite, para cada documento, identificar os trabalhos que vieram a citá-lo posteriormente. Pode-se ainda construir o índice de citações na forma de um grafo denominado Rede de Citações, o qual permite navegar em ambos os sentidos da relação de referência.

Aplicada à produção científica de uma instituição de pesquisa, uma rede de citações permite identificar quais trabalhos e autores têm maior influência sobre a pesquisa interna da instituição, o que pode apoiar decisões estratégicas por parte desta (FU; YUAN, 2010; HUANG, 2012). Porém, obter esta rede, que está limitada à produção de uma instituição em particular, é um processo complexo, primariamente por dois motivos. Alguns índices de citação disponíveis não proveem mecanismos para a busca apenas de documentos produzidos por autores afiliados a uma instituição. Outros se limitam a um acervo restrito de documentos, seja por questões comerciais, de curadoria, ou por associação a um repositório digital específico. Deste modo, a produção da instituição pode não ser coberta por um único índice, ou em parte até mesmo por nenhum (FU; YUAN, 2010).

Um modo de contornar a falta de um filtro de produções científicas por instituição é promovendo o uso de um repositório digital institucional, que armazena em um único local cópias das publicações de seus pesquisadores que foram originalmente publicadas em outros meios, como conferências e periódicos. Porém, esta ação sozinha não resolve os problemas quanto à indexação destes trabalhos por índices de citação, uma vez que, até por ser uma fonte secundária de publicação, o repositório institucional não será indexado por índices existentes, e mesmo que seja, volta-se ao problema de não poder ser selecionado apenas as produções da instituição. Assim, faz-se necessária não apenas a adoção de um repositório institucional, mas também a construção de uma rede de citações específica para este repositório.

Este trabalho originou-se a partir da necessidade da Universidade Federal do Paraná (UFPR) de identificar quais produções científicas influenciam sua produção acadêmica. Para tal, desenvolveu-se uma ferramenta que, a partir dos documentos na Biblioteca Digital de Teses e Dissertações da universidade, relaciona estes documentos por meio de suas referências bibliográficas e, deste modo, constrói uma rede de citações para o repositório digital da instituição. Prezou-se por arquitetar esta ferramente de modo que a mesma pudesse ser facilmente adaptada para uso em outros conjuntos de documentos acadêmico-científicos do Sistema de

Bibliotecas da UFPR.

A seguir, explica-se os conceitos relativos aos repositórios digitais e às redes de citações, bem como uma visão geral dos trabalhos relacionados. Depois, o capítulo 3 apresenta a arquitetura e a implementação da ferramenta mencionada anteriormente. No capítulo 4 é descrito o caso utilizado para avaliação da ferramenta, os resultados obtidos, e as principais dificuldades encontradas para a obtenção de resultados satisfatórios. Por fim, o capítulo 5 relata as conclusões e identifica trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados os conceitos gerais envolvidos na construção de uma rede de citações, além de definições relevantes ao estudo de caso e uma visão de trabalhos relacionados.

2.1 REPOSITÓRIOS DIGITAIS

Em sua concepção mais simples, um repositório digital é um sistema de acesso, preservação e gerenciamento de recursos em meio digital, sejam estas representações digitalizadas a partir de materiais físicos (como livros ou edições de um jornal) ou artefatos com origem digital (*born-digital*), tais como planilhas eletrônicas, *e-books* e arquivos multimídia (JANTZ; GIARLO, 2005).

Uma possível classificação de repositórios digitais utiliza a abrangência de seus conteúdos como critério, podendo ser:

- *Repositórios institucionais*: sistemas para captura e preservação da produção científica e intelectual dos membros de uma instituição (CROW, 2002; LYNCH, 2003);
- *Repositórios temáticos*: armazenam documentos relacionados a um assunto, tema ou área específica, sem distinguir sua origem (CROW, 2002).

Outro critério pelo qual se pode categorizar os repositórios digitais é o tipo de conteúdo que se propõem a armazenar. Na área acadêmico-científica, podemos encontrar:

- *Bases de teses e dissertações*: preservam os trabalhos de cursos de pós-graduação em instituições de ensino superior (YIOTIS, 2003);
- *Servidores de e-prints*: dão acesso à cópias auto-arquivadas de artigos científicos (HARNAD, 2001);
- *Bases de dados científicos*: disponibilizam os dados utilizados em pesquisas científicas (BEAGRIE; LAVOIE; WOOLLARD, 2010).

Para Lagoze e Van de Sompel (2003), como o objetivo maior dos repositórios digitais é a disponibilização de documentos ao seu público-alvo, é crucial a interoperabilidade entre repositórios e outros sistemas. Assim, torna-se possível não apenas novas formas de busca e descoberta de conteúdos, mas também a criação de novos serviços que agreguem valor a estes repositórios, o que aumenta a visibilidade do repositório, dos arquivos nele contidos e de seus autores.

2.1.1 Interoperabilidade através dos protocolos OAI

A Open Archives Initiative (OAI) foi formada em 2000 com a missão de desenvolver e promover padrões de interoperabilidade entre repositórios de conteúdo digital, incluindo artigos científicos (LAGOZE; VAN DE SOMPEL, 2001). Entre as especificações desenvolvidas pela iniciativa estão o Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) e o Open Archives Initiative Object Reuse and Exchange (OAI-ORE).

O protocolo OAI-PMH tem como objetivo permitir a coleta seletiva de metadados[1] de um repositório, isto é, a obtenção dos metadados de todos os itens presentes no repositório ou daqueles que obedecem a critérios temporais (data de criação, exclusão ou modificação) ou hierárquicos (coleções) (OAI, 2015).

A comunicação entre o repositório (que produz os metadados expostos via OAI-PMH) e o *harvester* (que colhe estes metadados) se dá através do protocolo HTTP, facilitando a integração com servidores *web* já existentes. As respostas às requisições realizadas são efetuadas através de documentos XML, que são definidos através de um esquema XML para fins de validação (LAGOZE; VAN DE SOMPEL, 2003). Não há restrições para os esquemas de metadados suportados pelo repositório, porém os metadados devem estar disponíveis no formato Dublin Core de modo a permitir a interoperabilidade direta entre serviços (OAI, 2015).

Como o propósito do OAI-PMH é o compartilhamento de metadados, o mesmo não expõe nenhuma funcionalidade para obter o conteúdo dos itens do repositório. Para este fim pode-se fazer uso do protocolo OAI-ORE, que permite a descoberta dos diferentes recursos contidos em um item e das relações entre itens. O Modelo de Dados ORE, baseado em Resource Description Framework (RDF), introduz duas entidades: a *agregação*, que é uma entidade conceitual que serve como uma coleção de recursos RDF; e o *mapa de recursos*, que representa informações como a agregação que o mapa descreve, os recursos agregados, relações e propriedades desses recursos, assim como metadados sobre o próprio mapa (OAI, 2008)

Em conjunto, os protocolos OAI-PMH e OAI-ORE permitem a replicação total ou parcial do conteúdo de um repositório digital, seja em seu formato original ou resultante de alguma transformação.

O ANEXO A apresenta um exemplo dos protocolos OAI-PMH e OAI-ORE, fazendo uma solicitação a um repositório DSpace via OAI-PMH requisitando metadados no formato OAI-ORE.

2.2 CITAÇÃO E REFERÊNCIA BIBLIOGRÁFICA

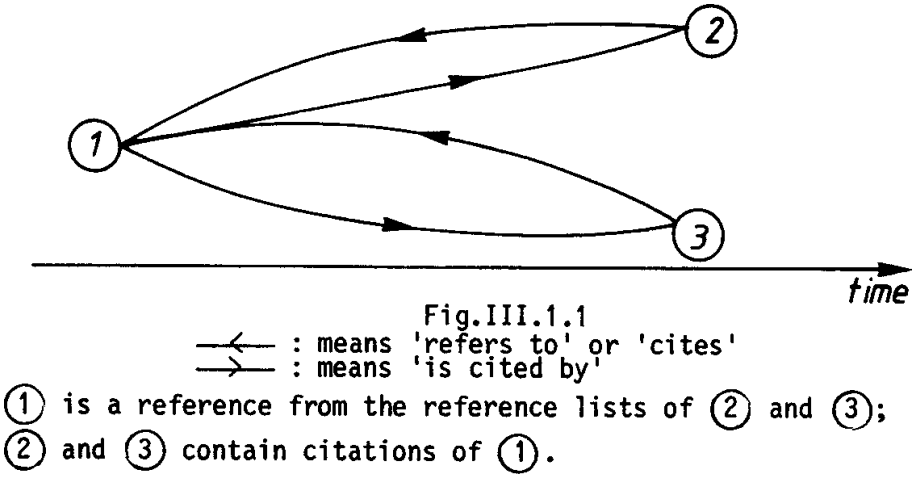
Para Sandison (1989), uma citação representa a indicação pelo autor de um documento (o *documento citante*) de que um outro trabalho (o *documento citado*) é relevante para o assunto sendo tratado em alguma parte de seu documento.

[1] Conjunto de informações associadas a um conteúdo que o identificam e o descrevem (FERREIRA, E., 2016).

Spinak (1996, p.51) observa que os termos *citação* e *referência*, dependendo do autor e da área, podem ser considerados sinônimos ou podem ter conotações distintas. Uma diferenciação comum é a proposta por De Solla Price (1986, p. 284), que utiliza os termos para distinguir os dois sentidos da relação de citação, de modo que, dados dois artigos A e B onde A aparece na lista de referências bibliográficas de B, então B refere-se à A e A é citado por B. Este exemplo está diagramado na FIGURA 1, onde os artigos A e B são numerados ① e ②, respectivamente.

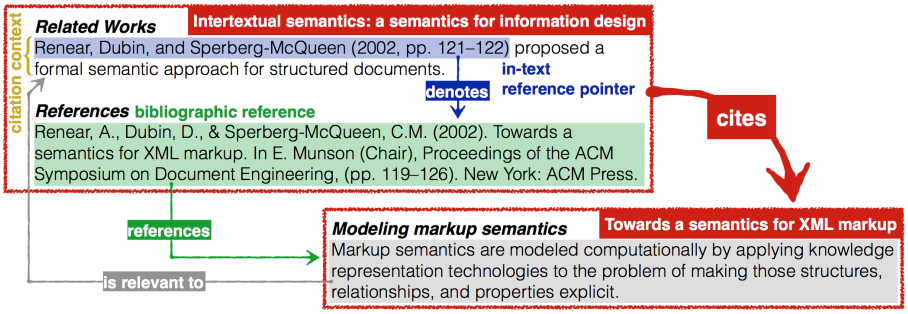
Outra abordagem, descrita por Peroni (2014, p. 172), faz uma distinção mais fundamental entre os termos: a citação é a declaração ou conteúdo que relaciona o artigo citante ao artigo citado, enquanto que a referência é o texto presente no artigo citante que identifica o artigo citado. A FIGURA 2 apresenta os diferentes componentes relacionados a uma citação neste modelo.

FIGURA 1 – Relação entre citação e referência segundo De Solla Price (1986)



Fonte: Egghe e Rousseau (1990, p. 204)

FIGURA 2 – Relação entre citação e referência segundo Peroni (2014)



Fonte: Di Iorio et al. (2014, p. 3)

A citação é uma ferramenta fundamental para a comunicação científica, utilizada para dar crédito aos autores de ideias e técnicas sobre as quais um trabalho se sustenta (PERONI et al., 2015, p. 255) e reconhecer trabalhos relacionados (GARFIELD, 1965, p. 189). Por outro lado, o número de citações que um trabalho recebe é, segundo Garfield (1979, p. 23–24), uma das medidas mais objetivas da relevância do trabalho para a pesquisa científica corrente.

2.3 ANÁLISE DE CITAÇÕES

O estudo das relações entre documentos, determinadas a partir de suas referências é o assunto da área da bibliometria denominada *análise de citações* (EGGHE; ROUSSEAU, 1990, p. 203), que abrange desde a criação de métricas de impacto de artigos, autores e periódicos até a construção de grafos de citação.

2.3.1 Métricas de impacto

Garfield (1977) destaca que a frequência com que um trabalho é citado é uma medida de atividade e comunicação científica, e não de sua significância. Assim, outras métricas devem ser utilizadas em conjunto para a obtenção de um indicador útil, porém, cada uma dessas métricas individualmente pode servir de ponto de partida para um processo de tomada de decisão.

Várias métricas surgiram com base na contagem de referências para quantificar o impacto de artigos, autores e periódicos científicos. Por exemplo, o Journal Impact Factor (JIF), criado por Eugene Garfield e Irving H. Sher em 1969 e muito utilizado para avaliação da qualidade das publicações de periódicos e pesquisadores (SAHA; SAINT; CHRISTAKIS, 2003), é calculado levando em consideração o número de referências em um dado ano para todas as publicações nos dois anos anteriores (GARFIELD, 2006):

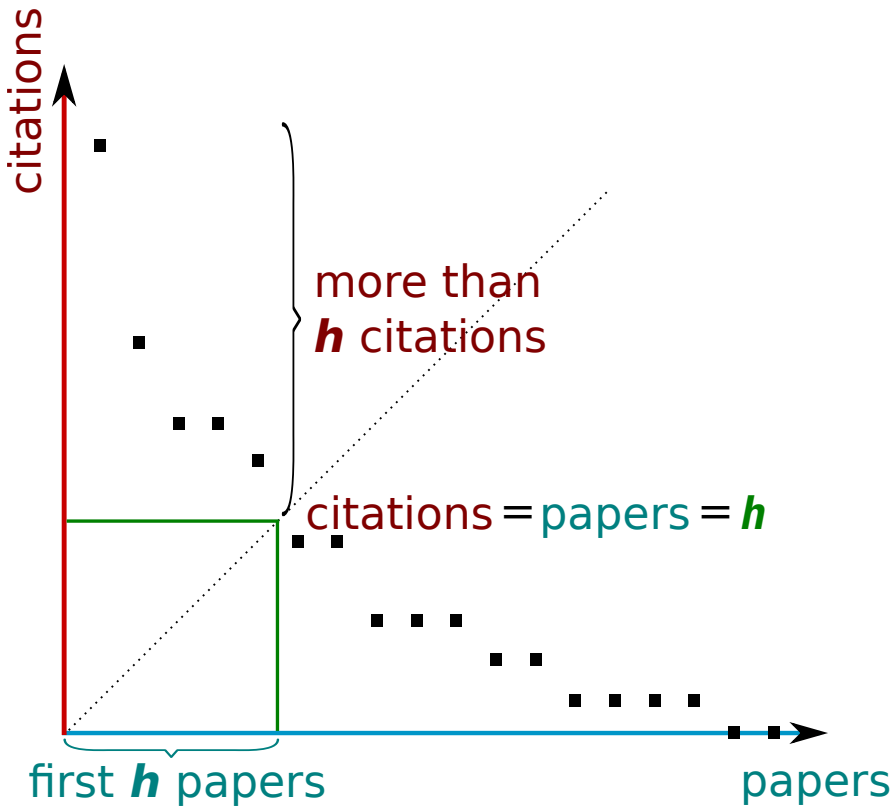
$$\frac{|\{r \in R(a)\ \forall a \in A(Y_{atual}) \quad \setminus \quad T(r) \in A(Y_{atual}-2) \cup A(Y_{atual}-1)\}|}{|A(Y_{atual}-2) \cup A(Y_{atual}-1)|} \tag{1}$$

onde

$$\begin{aligned} R(a) &= \text{referências contidas no artigo } a \\ A(y) &= \text{artigos publicados no periódico no ano } y \\ T(r) &= \text{artigo referenciado por } r \\ Y_{atual} &= \text{ano corrente} \end{aligned}$$

Entre as métricas voltadas aos autores está o *índice-h* (HIRSCH, 2005), cujo valor é calculado ordenando-se as publicações do autor por ordem decrescente de citações feitas à cada publicação e identificando o maior h tal que o h-ésimo documento tem h ou mais citações (FIGURA 3). O índice-h também pode ser adaptado para avaliação quantitativa de instituições com base nas publicações de autores afiliados àquela instituição (HUANG, 2012; PRATHAP, 2006).

FIGURA 3 – Cálculo do índice-h baseando em uma série de publicações e suas citações



FONTE: https://commons.wikimedia.org/wiki/File:H-index-en.svg.

2.3.2 Redes de citações

Um uso comum da análise de citações é o estudo da estrutura da pesquisa e comunicação científica através de redes de citações, conceituadas por Ruas e Marta Ferreira (2016, p. 159) como "uma rede [ou grafo direcionado] na qual os atores (nós) são os autores e os laços [arestas] as citações realizadas", sendo que a direção das arestas "indica se um determinado autor fez uma citação ou foi citado".

De Solla Price (1965), considerado um pioneiro na área de análise de redes de citações, estudou 200 artigos sobre os *raios N*, uma área relativamente isolada da física devido à subsequente revelação de sua natureza como um erro experimental. Foi observado que estes artigos tendiam a fazer mais referências à uma porção de trabalhos mais recentes, que Price denominou de "frente de pesquisa" e conjecturou ser um fenômeno de ocorrência decrescente entre as ciências naturais, as ciências sociais, e as humanidades. Embora problemas tenham sido encontrados com o conjunto de artigos utilizado por Price, sua avaliação sobre a estrutura de redes de citações foi parcialmente validada por estudos posteriores (BALDI; HARGENS, L., 1995; BALDI; HARGENS, L. L., 1997).

Outras aplicações das redes de citações na observação da comunicação científica

incluem o mapeamento das contribuições que levaram a um dado avanço científico, para a qual Garfield (1979, p. 81) provê vários exemplos, e a identificação de comunidades informais de pesquisadores e de áreas interdisciplinares, sugerida por Schwartzman (1984).

Meadows (1999, p.63) sugere, para além da simples análise dos trabalhos referenciados, a análise de co-citações: de fato, ao associar pares de artigos pela frequência com que são referenciados em conjunto por outros trabalhos, discerne-se grupos de artigos que tratam de um mesmo assunto. Small e Griffith (1974) validam esta observação através de um experimento com artigos das mais diversas áreas do conhecimento.

Para construção de uma rede de citações, é necessária a identificação dos documentos na rede e das citações feitas a cada um deles. Esta informação é obtida por meio de um *índice de citações*, originalmente proposto por Garfield (1955) como um complemento aos tradicionais índices alfabéticos e por assunto. Atualmente, existem vários índices de citações disponíveis em meio eletrônico:

- **Web of Science** (1964) – Conjunto de índices de citações em periódicos separados por área do conhecimento, além de bases de citações em conferências e livros, e bases de citações em periódicos regionais. Inclusão seletiva de periódicos (TESTA, 2016).
- **CiteSeer** (1997) / **CiteSeerX** (2008) – Primeiro índice autônomo de citações, faz buscas pela Web por artigos para indexação, não se limitando assim a qualquer base.
- **ACM Digital Library** (1997) – Repositório de publicações da Association for Computing Machinery (ACM) e de editores selecionados.
- **IEEE Xplore** (2000) – Repositório de publicações do IEEE.
- **Citations in Economics (CitEc)** (2001) – Índice autônomo de citações para o repositório Research Papers in Economics (RePEc), inicialmente baseado no CiteSeer.
- **Scopus** (2004) – Base de resumos e citações de literatura revisada por pares.
- **Google Scholar** (2004) – Mecanismo de busca que inclui fontes não-tradicionais, como *preprints*, anais de conferências e repositórios institucionais (GILES, J., 2005).
- **SciELO Analytics** (2015) – Serviço de indicadores e estatísticas de acesso para os artigos disponíveis na biblioteca Scientific Electronic Library Online (SciELO).

2.3.3 Extração de citações

O processo de análise de citações requer, em primeiro lugar, a obtenção e divisão semântica das referências bibliográficas. Em alguns casos, como o dos artigos científicos publicados pela SciELO e disponíveis no formato de marcação XML Jornal Article Tag Suite (JATS) (PACKER et al., 2014), essa informação já está disponível no próprio documento.

Geralmente, no entanto, os documentos a serem analisados estão nos formatos PDF ou HTML, sem metadados ou descrição semântica que permita o fácil acesso às suas partes constituintes. Assim, faz-se necessário o uso de ferramentas para localizar, extrair e segmentar as referências bibliográficas nestes documentos.

Em C. Lee Giles, Bollacker e Lawrence (1998), trabalho considerado um dos pioneiros no desenvolvimento de índices de citações autônomos, isto é, compilados sem intervenção humana, a segmentação de referências bibliográficas é apontada como uma das dificuldades no desenvolvimento deste tipo de sistema, agravada ainda pela existência de vários padrões de formatação de referências e pela ambiguidade na separação entre os campos.

Lipinski et al. (2013) avalia algumas das ferramentas disponíveis para extração de metadados estruturados em documentos PDF científicos, identificando como métodos principais utilizados para essa tarefa: análise estilística do documento; técnicas de aprendizado de máquina como Support Vector Machines (SVMs), *hidden Markov model*s (HMMs) e Conditional Random Fields (CRFs) aplicadas ao texto; e busca em bases de dados (para a identificação dos campos em uma referência).

2.3.4 Combinação de citações

Uma vez obtidas as referências, é preciso identificar o documento ao qual cada uma se refere. Hitchcock et al. (1997) e Lawrence, C. Lee Giles e Bollacker (1999) investigam a aplicação de métricas de distância entre palavras e frases para realizar essa associação. Caplan (2001) relata esforços no uso de URLs persistentes para ligação de referências à documentos.

Mais recentemente, técnicas de mineração de dados e *big data* vêm sendo aplicadas à combinação (ou correspondência) de citações. Schimidt (2012) propõe o uso de chaves derivadas de dados bibliográficos para combinação de citações. Fedoryszak, Tkaczyk e Bolikowski (2013) utilizam SVMs e CRFs em conjunto com o framework Apache Hadoop[2] para combinação de citações em larga escala.

2.4 EXTRAÇÃO, TRANSFORMAÇÃO E CARREGAMENTO DE DADOS (ETL)

Vassiliadis (2009) aponta que a consolidação de dados advindos de diversas fontes em uma única base de dados é uma tarefa que possui diversos problemas. Inicialmente, cada uma das fontes consultadas pode ter seus dados organizados de diferentes modos, e é preciso adaptá-los a um formato comum. A qualidade desses dados pode variar, o que influencia nos processos executados posteriormente sobre estes dados. Por fim, os dados podem ser atualizados em sua origem, sendo necessário realimentar a base de dados final com as atualizações.

O conjunto de processos que obtém dados de fontes distintas e os insere em uma única base, observando e auxiliando na resolução dos problemas citados, é conhecido como *Extract,*

[2] http://hadoop.apache.org/

Transform, Load (ETL) (VASSILIADIS, 2009). Esta forma de integração de dados extraí os dados de um sistema-fonte, os converte em um formato que pode ser utilizado para análises sobre esses dados, e armazena os dados transformados em um sistema de destino (SAS, 2017). No decorrer desses processos são utilizadas técnicas como *data cleansing* para melhorar a qualidade dos dados obtidos (JAGUŠT; STOJANOVSKI; BARANOVIĆ, 2014).

2.5 *RECORD LINKAGE*

O processo de *record linkage*, também conhecido como resolução de entidades ou de-duplicação, almeja identificar registros que correspondem a um mesmo indivíduo, objeto ou evento (FELLEGI; SUNTER, 1969). Assume-se que tais registros são similares, mas não necessariamente idênticos, pois do contrário seria trivial a identificação de duplicatas (CAVALIERI, 2014).

O primeiro passo para a deduplicação é a limpeza e padronização dos dados, de modo a remover ou corrigir dados inconsistentes (que possam levar a resultados errôneos nas etapas posteriores), e a eliminar diferenças entre representações de um mesmo valor, tornando mais fácil a comparação de registros (CHRISTEN; CHURCHES; HEGLAND, 2004). Os procedimentos exatos variam de acordo com as necessidades de cada conjunto de dados, mas costumam incluir técnicas como normalização de *strings* (envolvendo conversão para todas letras maiúsculas ou minúsculas, remoção de diacríticos e caracteres não-alfanuméricos, e expansão de abreviaturas) e remoção de formatação em valores como números de telefone e datas (RAHM; DO, 2000).

Os dados em cada par de registros são então comparados, campo a campo. Novamente, os algoritmos utilizados variam, mas todos resultam em um valor que indica o quão similares aqueles campos são. Por exemplo:

- números podem ser comparados por sua diferença absoluta (que pode ser vista então como uma medida de erro numérico);
- listas e conjuntos podem ser avaliados pelo número total de elementos que não são comuns à ambos os registros;
- *strings* podem ser comparadas através de medidas de distância de edição como *q-gramas* (SUTINEN; TARHIO, 1995) e a distância de Levenshtein (LEVENSHTEIN, 1966), que conta as operações sobre caracteres necessárias para transformar uma das *strings* na outra.

Em conjuntos de registros de tamanho substancial, é interessante reduzir o número de comparações realizadas. Para tal, utiliza-se a técnica de *blocagem* ou *clusterização*, que gera chaves a partir dos registros sendo processados e faz com que apenas registros com a mesma chave sejam comparados; assim, é necessário que o algoritmo de geração de chaves garanta que registros de fato similares sempre resultem na mesma chave (CHRISTEN; CHURCHES; HEGLAND, 2004).

2.6 TRABALHOS RELACIONADOS

2.6.1 Science Citation Index

Nove anos após a proposta de Garfield (1955) para a criação de um índice de citações para produção científica, é lançado o Science Citation Index (SCI), indexando um conjunto seleto de periódicos em diversas áreas. Com o tempo, foram agregados serviços como o Journal Citation Reports (JCR) e métricas como o Journal Impact Factor (JIF) (GARFIELD, 2007). Hoje o SCI faz parte, junto com outros índices específicos sobre áreas e tipos de literatura científica, da *Web of Science Core Collection*, indexando um total de 18 mil periódicos, 80 mil livros, e 180 mil anais de conferências (CLARIVATE, 2017).

Uma análise informal revela o uso do SCI em grande parte dos artigos científicos encontrados através de uma busca pelas palavras-chave *"citation index"* tanto no Google Scholar quanto nas bases de dados disponíveis para a UFPR[3]. Devido à cobertura limitada (e em declínio) deste índice de citações, existem preocupações quanto ao seu uso como base para indicadores de produção científica (LARSEN; VON INS, 2010).

2.6.2 CiteSeer

C. Lee Giles, Bollacker e Lawrence (1998) introduzem o *CiteSeer*, primeiro índice de citações com operação totalmente automática, dramaticamente reduzindo os custos de operação associados a um sistema desse tipo. Ao invés de limitar-se a uma base pré-selecionada de artigos, o CiteSeer busca em toda a *web* por documentos científicos no formato PostScript. Uma limitação desta abordagem é que não são cobertos documentos não disponíveis no meio digital ou cujo acesso é de alguma forma restrito.

A aquisição de documentos no CiteSeer é feita através de pesquisas em mecanismos de busca por páginas contendo palavras-chave como "publications", "papers" e "postscript", que são então vasculhadas em busca de *links* para arquivos PostScript. É feita então uma busca nos arquivos obtidos por uma seção de bibliografia ou referências, para verificar que o documento é um trabalho de pesquisa. Destes arquivos identificados como produção científica são extraídos metadados como título e autores, partes como resumo e introdução, a lista de referências bibliográficas, e o contexto das citações presentes no texto do documento.

O processo de ligação das citações extraídas a outros documentos presentes na base de conhecimento do CiteSeer é simples. Primeiro, ocorre a normalização do texto da citação com a conversão de letras para minúsculas, remoção de caracteres especiais e marcadores de citação, e expansão de abreviaturas comuns. Após normalizadas, as citações são agrupadas através da contagem de palavras e pares de palavras em comum.

[3] Acesso disponível para a comunidade da UFPR em: `http://search.ebscohost.com/login.aspx?authtype=ip,guest&custid=s6842739&groupid=main&profile=eds`

Além da indexação de citações, o CiteSeer também possui um sistema de recomendação de documentos, relacionando-os por meio de métricas como term frequency – inverse document frequency (TDIDF) e referências bibliográficas em comum.

Li et al. (2006) apresentam o *CiteSeer^x*, uma nova arquitetura para o CiteSeer como foco em escalabilidade e extensibilidade. Mudanças incluem a utilização de um *web crawler* próprio para localização de documentos ao invés de buscas em mecanismos externos, e a adoção do software ParsCit (COUNCILL; GILES, C. L.; KAN, 2008) para extração de metadados e referências dos documentos obtidos, com suporte a arquivos PDF.

2.6.3 CitEc

O Citations in Economics (CitEc), desenvolvido por Barrueco e Krichel (2005), é um índice de citações para o repositório decentralizado Research Papers in Economics (RePEc), que contém *working papers* e artigos publicados em periódicos científicos na área da Economia. Sua arquitetura é ilustrada pela FIGURA 4.

O CitEc utiliza uma base de conhecimento construída a partir dos metadados dos documentos presentes no RePEc para a ligação de referências a outros trabalhos presentes naquele repositório. Estes metadados, assim como os documentos a serem processados, são obtidos por meio de um protocolo específico ao RePEc desenvolvido para comunicação entre os acervos participantes.

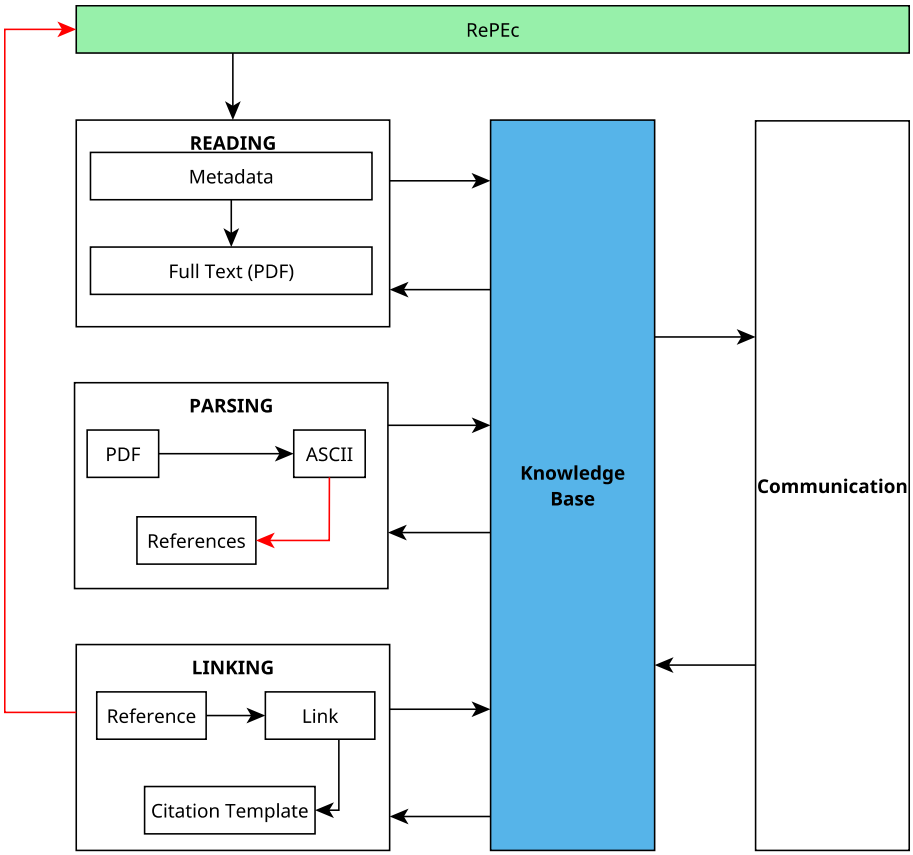
A extração de referências no CitEc se dá através do uso do CiteSeer, mencionado anteriormente. A ligação de referências compara uma versão normalizada do título da referência bibliográfica (com conversão para letras minúsculas e remoção de espaços múltiplos e artigos) com os títulos presentes na base de conhecimento, utilizando como métrica de comparação a distância de Levenshtein. Os registros considerados próximos são então comparados pelo ano de publicação, sendo considerados uma combinação positiva caso o mesmo seja igual ao ano de publicação listado na referência.

2.6.4 Institutional Citation Index

O *Institutional Citation Index* foi proposto por Fu e Yuan (2010) como uma alternativa ao uso do Science Citation Index como base para métricas de produção científica em uma instituição de pesquisa. Para tanto, seriam utilizados módulos integrados a um repositório digital já existente, conforme apresentado na FIGURA 5.

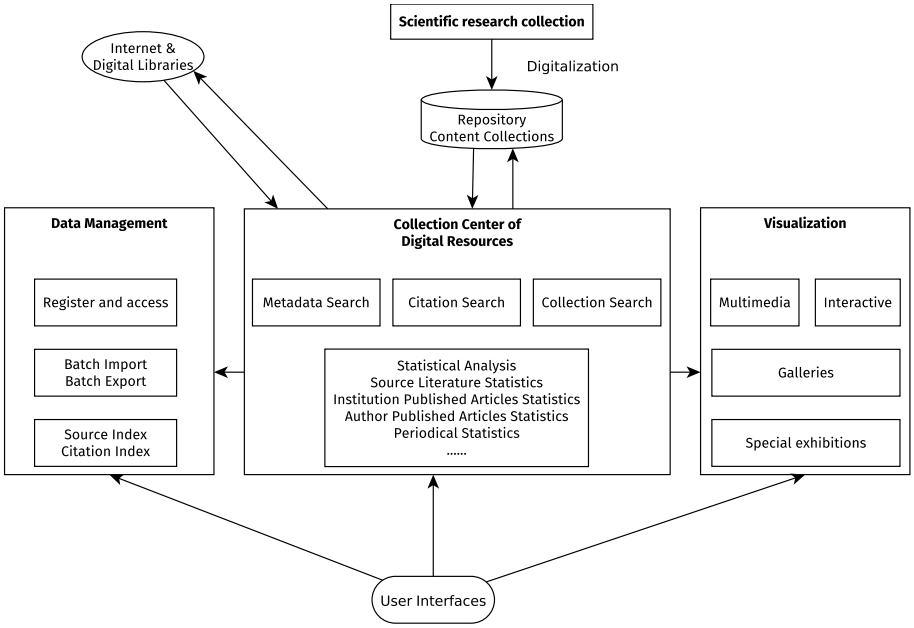
O *Institutional Citation Index* compartilha os mesmos objetivos do presente trabalho: a inclusão de toda a produção científica de interesse, independentemente de onde foi originalmente publicada; o suporte a outros tipos de literatura científica (e não somente artigos em periódicos e conferências); e a construção de redes de citação focadas na produção de uma única instituição.

FIGURA 4 – Arquitetura do CitEc



FONTE: Adaptada de Barrueco e Krichel (2005, p. 5).

FIGURA 5 – Módulos componentes do *Institutional Citation Index*



FONTE: Adaptada de Fu e Yuan (2010, p. 1209).

3 REDE DE CITAÇÕES PARA A BIBLIOTECA DIGITAL DE TESES E DISSERTAÇÕES DA UFPR

Devido às limitações identificadas anteriormente dos índices de citações existentes, faz-se necessário o desenvolvimento de um índice específico para a Biblioteca Digital de Teses e Dissertações (BDTD) da Universidade Federal do Paraná. A princípio, considerou-se utilizar o CiteSeerX para construção da rede de citações, posteriormente integrando-o ao Acervo Digital da UFPR para visualização da rede e disponibilização de um índice de citações sobre sobre a Biblioteca Digital de Teses e Dissertações (BDTD). Entretanto, após estudo do código-fonte do *software* e de seu manual de operações (WU et al., 2015), optou-se por desenvolver um sistema novo, com escalabilidade limitada porém adequada às demandas do projeto, e que permitisse a avaliação de diferentes implementações dos processos envolvidos na construção da rede de citações. Assim, deu-se início ao desenvolvimento do Citation Network Builder (CNB).

3.1 VISÃO GERAL

Inspirado pela arquitetura do índice de citações CitEc (BARRUECO; KRICHEL, 2005), o CNB é composto por módulos que correspondem às três fases do processo de construção de uma rede de citações, sendo que cada fase depende do sucesso das fases anteriores:

1. **Aquisição de documentos** e registros bibliográficos associados;
2. **Extração de referências bibliográficas** a partir dos documentos obtidos em formato PDF;
3. **Ligação das referências** aos registros bibliográficos obtidos na primeira etapa.

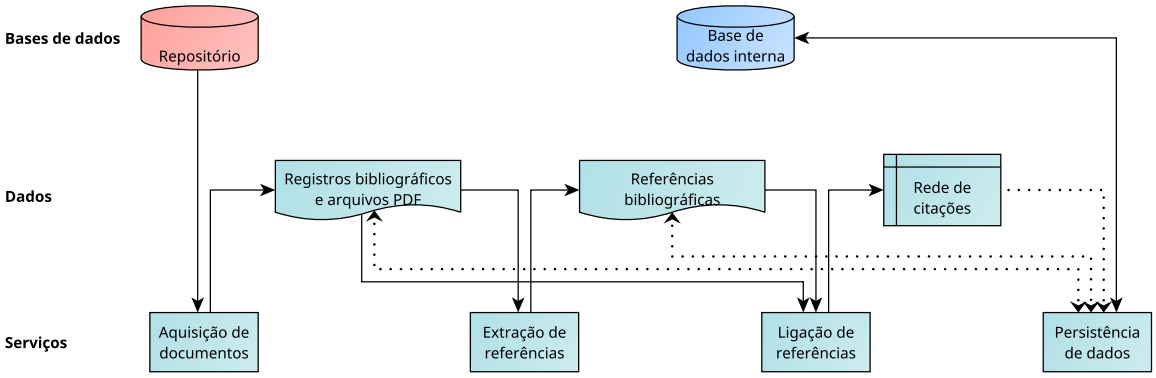
Os dois primeiros processos podem ser vistos como processos ETL, e o último é uma aplicação de *record linkage*.

Durante o desenvolvimento do CNB, determinou-se como característica diferencial do *software* o suporte à diversas implementações dos processos destas três fases. Esta característica permitiria a fácil adaptação da ferramenta para operação sobre outras bases de documentos, como a Biblioteca Digital de Periódicos (BDP) e a Biblioteca Digital de Eventos Científicos (BDEC) da UFPR, e também a avaliação e utilização de diferentes estratégias de extração e ligação de referências, adaptadas aos documentos presentes em cada repositório. Esta característica é obtida através da separação clara entre as partes do sistema responsáveis por cada fase, o que evita que a *implementação* de uma parte influencie as outras.

Inspirado pelo conceito de interface de provedor de serviço (SPI) da linguagem de programação Java (ORACLE, 2017), concebeu-se uma arquitetura, apresentada na FIGURA 6, em que cada uma das fases delineadas, bem como quaisquer componentes igualmente substituíveis

do sistema, é representada por um *serviço* com entradas, saídas e responsabilidades bem-definidas. Esta delimitação permite não apenas a substituição da implementação de qualquer serviço sem prejuízo à operação do resto do sistema, como também a livre modificação das partes restantes do sistema, por exemplo, para diferentes paradigmas de interface com o usuário.

FIGURA 6 – Fluxo de dados entre serviços no *Citation Network Builder*



FONTE: O autor (2017).

3.2 IMPLEMENTAÇÃO

Para implementação do CNB foi utilizada a linguagem de programação Kotlin[1] tendo como plataforma-alvo a máquina virtual Java (JVM). Para reduzir o acoplamento entre a interface com o usuário e os serviços listados na seção 3.1, foi utilizada a técnica de injeção de dependências com auxílio do *framework* Spring[2]. Ainda por meio do Spring, foi abstraído o acesso ao banco de dados que armazena as informações resultantes de cada fase.

O gerenciamento do processo de compilação do *software* se dá através do Apache Maven[3]. Para garantir o desacoplamento entre partes distintas do *software* e permitir a não-implantação de partes não utilizadas, o sistema foi dividido em módulos do Maven que são compilados independentemente um do outro exceto por relações de dependência. No futuro, estes módulos poderão ser mapeados transparentemente para módulos do Java 9.

Para representação da camada de dados da FIGURA 6, existem três interfaces definidas no módulo `cnb2-api` (que contém definições de interfaces utilizadas por todo o sistema):

- `Work`: representa um registro bibliográfico obtido do repositório durante a fase de aquisição. Campos: identificador no repositório, título, autores, e ano de publicação;

[1] `https://kotlinlang.org`
[2] `https://spring.io`
[3] `https://maven.apache.org`

- `Resource`: representa, como um recurso *web*, um arquivo Portable Document Format (PDF) vinculado a um registro no repositório. Contém o registro-pai do recurso, a URL do recurso e um indicador informando se já foram extraídas as referências presentes no recurso;

- `Reference`: representa uma referência bibliográfica extraída de um recurso. Seus campos incluem os metadados resultantes da segmentação da referência (os mesmos metadados presentes em `Work`), o registro ao qual o recurso de onde foi extraída pertence, e o registro para o qual a referência aponta, identificado durante a fase de ligação.

Similarmente, os serviços que compõem a arquitetura do sistema (apresentados na FIGURA 7) também são definidos por interfaces no módulo `cnb2-api`. A passagem de dados entre estes serviços ocorre através de fluxos reativos (*reactive streams*[4]) providos pela biblioteca RxJava[5], que agem como sequências assíncronas de objetos. Este modelo de comunicação permite o desenvolvimento futuro de serviços que operam concorrentemente, vindo a agilizar o processamento sobre repositórios em larga escala.

- `WorkSource`: serviço de aquisição de documentos. Produz como saída um fluxo de `Records`, classe que pode encapsular um `Work` ou indicar que um determinado registro foi excluído do repositório (quando esta informação está disponível).

- `ReferenceExtractor`: serviço de extração de referências bibliográficas. Recebe um `Work` e retorna um fluxo de `References` já associadas à este `Work`.

- `PdfTextExtractor`: serviço de extração de texto a partir de arquivos PDF, utilizado em conjunto com extratores de referências que trabalham apenas com texto puro. Recebe um fluxo de dados de um arquivo PDF e o transforma em uma *string* com o conteúdo textual do documento.

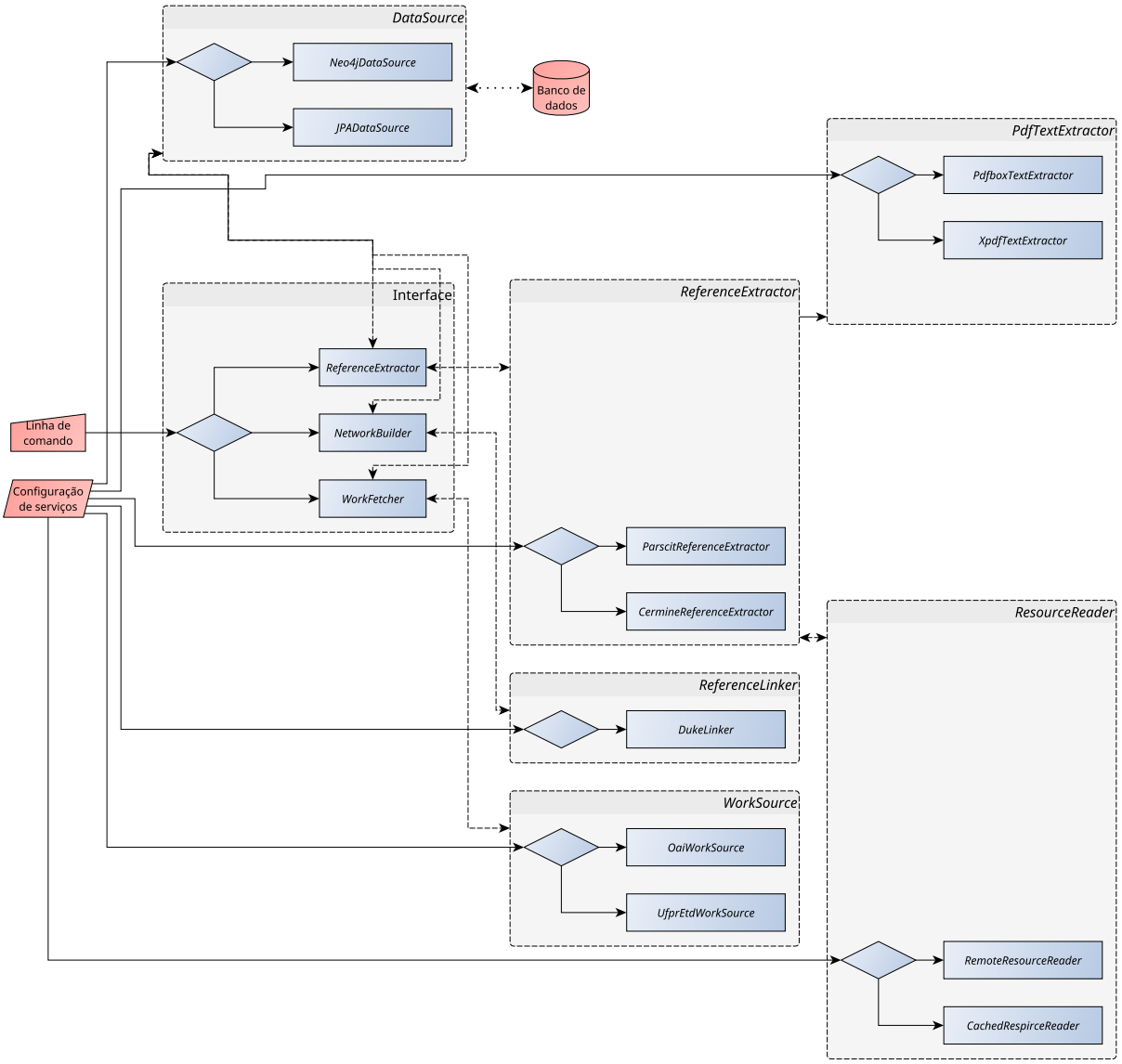
- `ReferenceLinker`: serviço de ligação de referências. Recebe fluxos de registros e referências obtidos com os serviços anteriores e produz um fluxo de *combinações*, que incluem os pares de registros que foram ligados, a referência através da qual foi feita a ligação, e opcionalmente um nível de confiança para aquela combinação.

O módulo `cnb2-services-common`, que dá suporte à implementação desses serviços, define dois serviços relacionados à persistência de dados:

- `Dao`: serviço de acesso ao banco de dados interno. Define métodos para busca, contagem e armazenamento de instâncias dos modelos de dados. Utilizado pelos comandos definidos a seguir.

[4] `http://reactivex.io`
[5] `https://github.com/ReactiveX/RxJava`

FIGURA 7 – Relacionamentos entre serviços e implementações de serviços do *Citation Network Builder* e o ambiente externo



FONTE: O autor (2017).

- `ModelFactory`: serviço de criação de instâncias dos modelos de dados. Pode criar novas instâncias ou modificar instâncias existentes de forma transparente ao usuário. Utilizado pelos serviços de aquisição de metadados e extração de referências.

A interface de linha de comando com o usuário e a integração entre dados e serviços encontra-se no módulo `cnb2`. Os *comandos* implementados aqui orquestram o fluxo de dados entre o banco de dados interno, que contém as instâncias persistidas dos modelos de dados gerados nas fases anteriores, e os serviços responsáveis pela execução propriamente dita de cada fase.

Antes de prosseguir com a descrição dos módulos que implementam cada serviço, lista-se a seguir módulos auxiliares desenvolvidos em conjunto com o CNB mas que são ortogonais

ao objetivo do sistema:

- `oai-harvester`: biblioteca para colheita de metadados via OAI-PMH.
- `rop`: suporte à abordagem de tratamento de erros *Railway Oriented Programming* (WLASCHIN, 2014), baseada na implementação da linguagem Rust[6].

3.2.1 Persistência de dados

O módulo `cnb2-datastore-jpa` fornece uma implementação utilizando a Java Persistence API (JPA) para conexão a sistemas de gerenciamento de banco de dados SGBDs relacionais através do Java Database Connectivity (JDBC). Cada modelo de dados é mapeado para uma tabela e cada campo para uma coluna; tabelas adicionais são criadas para as listas de autores de registros e referências, devido à sua multiplicidade.

Já o módulo `cnb2-datastore-neo4j` utiliza o banco de dados de grafos Neo4j[7]. Registros, recursos e referências são todos representados por nós com propriedades atreladas em um grafo, com arestas denotando os relacionamentos entre estes elementos. Embora fosse possível representar as referências bibliográficas por meio de arestas, como o registro sendo referenciado é desconhecido pelo menos até a conclusão da fase de ligação de referências, as arestas-referência teriam um nó de origem mas não necessariamente um de destino, o que é uma configuração não suportada pelo Neo4j.

3.2.2 Aquisição de documentos e metadados

O módulo `cnb2-worksource-oai` implementa a colheita de registros do repositório por OAI-PMH e OAI-ORE[8], através do módulo `oai-harvester`.

O módulo `cnb2-worksource-ufpr-etd` complementa a colheita por OAI com detalhes específicos à Biblioteca de Teses e Dissertações da UFPR, como configurações especiais de conexão ao Acervo Digital[9], seleção do conteúdo relevante dentro do acervo, e limpeza dos metadados para retirar a universidade da lista de autores e remover datas de nascimento e falecimento dos nomes dos autores.

A FIGURA 8 representa, na linguagem de serialização de dados YAML[10], as instâncias criadas a partir da resposta OAI-PMH no apêndice A.

[6] `https://doc.rust-lang.org/std/result/`
[7] `https://neo4j.com`
[8] Vide subseção 2.1.1.
[9] Vide capítulo 4.
[10] `http://yaml.org`

FIGURA 8 – Modelo de instâncias de dados após aquisição de metadados

```
!Work &work1
  external id: "oai:dspace.c3sl.ufpr.br:1884/18388"
  title: "Avaliação da atividade reprodutiva da ictiofauna
  ↪  capturada na pesca artesanal de arrasto camaroeiro
  ↪  pela comunidade de Itapema do Norte, Itapoa, litoral
  ↪  norte de Santa Catarina"
  authors:
    - "Juliana Ventura de Pina"
  publication year: 2009

!Resource &res1
  part of: *work1
  url: "http://acervodigital.ufpr.br/bitstream/1884/18388/1/
  ↪  DISSERTACAO_JULIANA%20VENTURA%20DE
  ↪  %20PINA.pdf"
  processed: false
```

FONTE: O autor (2017).

3.2.3 Extração de referências

Foram desenvolvidos dois módulos implementando o serviço `ReferenceExtractor`, ambos detalhados nas subseções subsequentes: um utiliza o *software* ParsCit, o mesmo utilizado pelo CiteSeerX; o outro utiliza o *framework* CERMINE.

Para suportar extratores de referências que não trabalhem com arquivos PDF, mas apenas com texto puro, existe a classe abstrata `TextOnlyReferenceExtractor`, que utiliza o serviço `PdfTextExtractor` para extrair o texto do documento antes de repassá-lo ao extrator de referências. Duas implementações deste serviço foram escritas: uma utilizando a biblioteca Apache PDFBox®[11], e outra para uso do conjunto de ferramentas de linha de comando do projeto Xpdf[12].

3.2.3.1 CERMINE

O *Content Extractor and Miner* (CERMINE)[13] é um sistema de extração de metadados e dados estruturados a partir de artigos científicos de origem digital (TKACZYK et al., 2015). Ao aplicar técnicas de aprendizado de máquina ao texto e ao *layout* do documento, é possível identificar não apenas metadados básicos como título, autores, e afiliação, mas também conectar os autores a suas afiliações e informações de contato. O CERMINE pode ainda extrair informações sobre o periódico no qual um artigo foi publicado, além de extrair e segmentar referências bibliográficas.

[11] https://pdfbox.apache.org
[12] http://www.xpdfreader.com
[13] http://cermine.ceon.pl

Assim como o presente trabalho, o CERMINE tem como característica a flexibilidade, permitindo implementações diversas para cada tarefa executada pelo sistema. Atualmente é utilizada a biblioteca iText para extração de caracteres e informação de posicionamento de texto no documento PDF, seguida pela divisão das páginas em blocos de texto e identificação das partes do artigo através de SVM. As referências bibliográficas são identificadas por meio da aplicação da técnica de clusterização K-médias sobre os blocos de texto identificados na segunda etapa. Por fim, são usados CRFs para segmentação do texto das referências encontradas.

3.2.3.2 ParsCit

ParsCit[14] é uma ferramenta escrita em Perl para extração e segmentação de referências bibliográficas, desenvolvida para uso no projeto CiteSeerX (COUNCILL; GILES, C. L.; KAN, 2008). Tomando o caminho inverso ao do CERMINE, o ParsCit aceita como entrada apenas texto puro na codificação UTF-8, ignorando a formatação do documento para evitar perda de generalidade. Assim como o CERMINE, são utilizados modelos CRF para segmentação de referências.

O apêndice B apresenta um excerto da saída XML do ParsCit para o presente trabalho, a qual é consumida pelo módulo `cnb2-refparsers-parscit` para geração dos modelos das referências. Os anexos D e E tabelam as listas de referências bibliográficas de duas dissertações, Salvi (2009) e Berleze (1988) respectivamente, e os resultados da segmentação efetuada pelo ParsCit sobre as referências extraídas.

3.2.4 Ligação de referências

O serviço de ligação de referências é responsável por identificar a qual documento na base de registros bibliográficos cada uma das referências extraídas se refere, caso esta combinação exista. Este é o único serviço que possui apenas uma implementação: o módulo `cnb2-reflinker-duke`, que utiliza a biblioteca de *record linkage* Duke, explicada a seguir.

3.2.4.1 Duke

A biblioteca Duke[15] foi desenvolvida por Lars Marius Garshol para resolução de entidades por meio de inferência Bayesiana, com o objetivo de auxiliar com o processo de extração de dados a partir de um sistema ERP (GARSHOL, 2011). O funcionamento básico da biblioteca é como segue: o usuário informa como obter os conjuntos de dados de entrada, e se a operação desejada é o *record linkage* entre conjuntos de registros de tipos distintos ou a deduplicação de registros em uma mesma base de dados. São definidas então propriedades sobre os conjuntos de dados de entrada, as quais são utilizadas posteriormente para comparação entre registros. Estas propriedades podem ser configuradas para pré-processar e comparar seus conteúdos de maneiras específicas, por exemplo: uma propriedade que contém números de telefone pode

[14] `http://parscit.comp.nus.edu.sg/`
[15] `https://github.com/larsga/Duke`

FIGURA 9 – Análise do Duke sobre combinação positiva de referência e registro bibliográfico

```
---TEXT
'classificacao automatica de erros de aprendizes humanos no
 ↪  processo de inducao analitica. dissertacao de mestrado,
 ↪  universidade federal do' ~ 'classificacao automatica de
 ↪  erros de aprendizes humanos do processo de inducao
 ↪  analitica': 0.9770114942528736
(prob 0.8818205839608932)
Result: 0.5 -> 0.8818205839608932

---AUTHORS
'gustavo cesar bazzo' ~ 'gustavo cesar bazzo': 1.0 (prob 0.7)
Result: 0.8818205839608932 -> 0.9456836352595892

---YEAR
'2013' ~ '2011': 0.75 (prob 0.5281250000000001)
Result: 0.9456836352595892 -> 0.9511864150739742

Overall: 0.9511864150739742
```

FONTE: O autor (2017).

remover caracteres não-numéricos, e uma propriedade de texto pode realizar comparações por fonemas, utilizando o algoritmo Soundex.

Para cada campo presente nos registros a serem comparados, são atribuídos pelo usuário dois valores, os quais indicam a probabilidade dos registros serem similares com base apenas na comparação daquele campo. Uma probabilidade maior que 0.5 aponta para registros similares, e uma probabilidade menor indica registros distintos. A probabilidade 0.5 significa que aquele campo não contribui para o cálculo da similaridade entre os registros, por exemplo: ao comparar registros de corporações, nomes diferentes não necessariamente significam companhias diferentes (considerando-se divisões e nomes regionais), porém nomes similares apontam para (mas não garantem) uma mesma companhia. Probabilidades mais próximas dos extremos (0.0 e 1.0) sinalizam um maior impacto na comparação.

As probabilidades calculadas para cada campo são então agregadas por meio de inferência Bayesiana, resultando na probabilidade dos registros serem sobre a mesma entidade. Este valor é, por fim, comparado com limites definidos pelo usuário para determinar se o par de registros é considerado uma combinação ou não.

As figuras 9, 10 e 11 apresentam o registro do processo de comparação feito pelo Duke em diferentes situações.

FIGURA 10 – Análise do Duke sobre combinação avaliada erroneamente

```
---TEXT
'cultura' ~ 'cultura esportiva : um possivel legado dos jogos
 ↪  olimpicos e paralimpicos rio 2016?': 1.0 (prob 0.9)
Result: 0.5 -> 0.9

---AUTHORS
'sao paulo' ~ 'ana paula prestes de souza': 0.5 (prob 0.55)
Result: 0.9 -> 0.9166666666666667

---YEAR
'2001' ~ '2015': 0.75 (prob 0.5281250000000001)
Result: 0.9166666666666667 -> 0.9248756218905473

Overall: 0.9248756218905473
```

FONTE: O autor (2017).

FIGURA 11 – Análise do Duke sobre combinação positiva com baixa probabilidade

```
---TEXT
'o espelho e a miragem : ecletismo, moradia e modernidade
 ↪  na curitiba do inicio do seculo.' ~ 'o espelho e a
 ↪  miragem': 1.0 (prob 0.9)
Result: 0.5 -> 0.9

---AUTHORS
'marcelo s sute' ~ 'marcelo saldanha sutil':
 ↪  0.6153846153846154 (prob 0.5757396449704142)
Result: 0.9 -> 0.9243191893603547

---YEAR
'1996' ~ '2011': 0.0 (prob 0.4)
Result: 0.9243191893603547 -> 0.8906178489702519

Overall: 0.8906178489702519
```

FONTE: O autor (2017).

3.2.5 Seleção de serviços

O CNB utiliza o Spring Boot[16] para configuração da execução do sistema. O comando a ser executado é obtido a partir da opção `command` da linha de comando, e outras configurações (como os dados de acesso ao banco de dados, para implementações do serviço de persistência de dados que utilizam Spring Data) podem ser especificadas no arquivo `application.yml`, exemplificado no apêndice A.

A escolha e configuração dos serviços a serem utilizados é feita por meio do arquivo `cnb-services.xml`, que segue o mesmo formato de um arquivo de configuração *JavaBeans* (`beans.xml`)[17]. Um exemplo deste arquivo encontra-se no apêndice B.

[16] https://projects.spring.io/spring-boot/

[17] https://docs.spring.io/spring/docs/4.3.x/spring-framework-reference/html/beans.html#beans-factory-metadata

4 AVALIAÇÃO

Para validação do sistema desenvolvido, aplicou-se o CNB para construção de uma rede de citações sobre os documentos na BDTD da UFPR (parte do Acervo Digital da UFPR).

O Acervo Digital da UFPR[1], hospedado pelo Centro de Computação Científica e Software Livre (C3SL), é um repositório digital institucional que abriga boa parte da produção acadêmica da UFPR, como monografias, teses, dissertações e trabalhos de especialização. O Acervo Digital engloba ainda repositórios de recursos e práticas educacionais abertas (REA/PEA), vídeos da TV UFPR, relatórios e documentos administrativos de interesse público, e está aberto à preservação de outros conteúdos criados pela universidade.

O Acervo Digital executa o *software* de repositório digital DSpace[2], desenvolvido inicialmente pelo Massachusetts Institute of Technology (MIT) e pelo Hewlett-Packard Labs "como um repositório para a pesquisa digital e material educativo produzido por membros de uma organização ou universidade de pesquisa" (SMITH et al., 2003, tradução nossa). O ambiente de execução inclui ainda o servidor de aplicação Apache Tomcat[3], o sistema gerenciador de banco de dados (SGBD) PostgreSQL[4], e o sistema operacional Debian GNU/Linux[5].

As teses e dissertações produzidas nos cursos de pós-graduação da UFPR formam a primeira coleção a fazer parte do Acervo Digital (SUNYE et al., 2009). Atualmente, estão disponíveis 17.760 documentos desses tipos advindos de 82 programas de pós-graduação.

4.1 CONFIGURAÇÃO DO CNB

Para construção da rede de citações, foram utilizadas as seguintes implementações de serviços:

- **Banco de dados**: JPA conectado à um SGBD PostgreSQL.
- **Aquisição de documentos**: UFPR/ETD.
- **Extração de referências**: ParsCit, acompanhado do extrator de texto Xpdf.
- **Ligação de referências**: Duke.

[1] http://acervodigital.ufpr.br
[2] http://dspace.org
[3] https://tomcat.apache.org
[4] https://www.postgresql.org
[5] https://www.debian.org

4.2 EXECUÇÃO E RESULTADOS

A partir da execução do CNB sobre a Biblioteca Digital de Teses e Dissertações (BDTD) da UFPR em novembro de 2017, foram obtidos registros bibliográficos de 7.275 documentos, dos quais foram extraídas um total de 657.253 referências bibliográficas utilizando o ParsCit em sua configuração padrão, sem treinamento específico para estes textos.

Para comparação entre referências e registros bibliográficos, utilizou-se a configuração listada no ANEXO C, com os parâmetros detalhados no QUADRO 1 para cada campo. Devido a problemas de desempenho para o carregamento dos dados durante a fase de ligação de referências, optou-se por extrair os dados relevantes do banco de dados em formato CSV e executar as operações desta fase através da ferramenta de linha de comando do Duke, utilizando a mesma configuração caso fosse feito o processamento por meio do CNB.

QUADRO 1 – Parâmetros de comparação entre campos de registros bibliográficos e referências

Campo	**Comparador**	**Justificativa**	**Probabilidade (similar)**	**Probabilidade (dissimilar)**
Título	Bigramas	Leva em consideração não apenas as palavras em comum mas também sua ordem.	0,3	0,9
Lista de autores	Bigramas	Leva em consideração não apenas as palavras em comum mas também sua ordem.	0,2	0,7
Ano de publicação	Distância de Levenshtein	Não penaliza erros de digitação e casos com anos de publicação distintos (por exemplo, publicação eletrônica em 2014 e impressa em 2015).	0,4	0,55

FONTE: O autor (2017).

Deste processo resultaram 457 ligações entre referências bibliográficas e documentos na BDTD, 390 das quais foram avaliadas manualmente como sendo corretas dadas as informações disponíveis na fase de ligação, representando uma precisão de 85.34%. O QUADRO 2 lista os trabalhos com maior número de referências a partir de outros documentos dentro da BDTD; 268 trabalhos distintos foram identificados como alvo de referências. A rede de citações assim gerada pode ser visualizada nas figuras 12 e 13, criadas com a plataforma de visualização de grafos Gephi[6].

[6] https://gephi.org

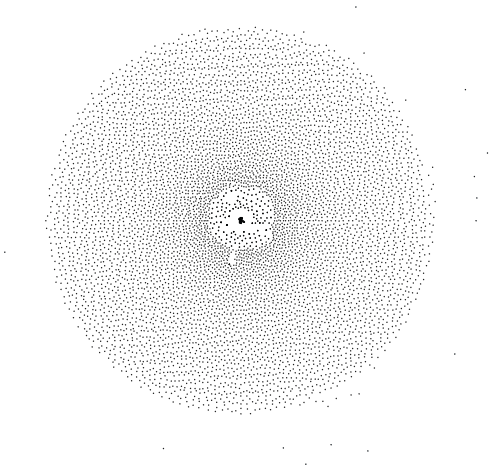
QUADRO 2 – Teses e dissertações mais citadas na Biblioteca Digital de Teses e Dissertações da UFPR

Identificador	Título	Tipo e área	Citado por
1884/6600	Intelectuais paranaenses e as concepções de Universidade: 1892–1950	Tese (Doutorado em Educação)	10
1884/24735	Estudo das representações de monarca nas Crônicas de Fernão Lopes (séculos XIV e XV)	Tese (Doutorado em História)	9
1884/19430	Das justiças e dos litígios: a ação judiciária da Câmara de Curitiba no século XVIII (1731–1752)	Tese (Doutorado em História)	6
1884/21040	Gerenciamento do projeto na ótica do gerenciamento da comunicação: manual para escritórios de arquitetura	Dissertação (Mestrado em Construção Civil)	6
1884/27804	A modernização da sociedade no discurso do empresariado paranaense: Curitiba 1890–1925	Dissertação (Mestrado em História)	5
1884/24578	Uma jornada civilizadora: imigração, conflito social e segurança pública na Província do Paraná – 1867 a 1882	Dissertação (Mestrado em História)	5
1884/25450	O espelho e a miragem: ecletismo, moradia e modernidade na Curitiba do início do século	Dissertação (Mestrado em História)	5
1884/27164	Paroquia de São Pedro do Rio Grande: estudo de história demográfica 1737–1850	Tese (Doutorado em História)	5
1884/18196	O aprendizado e a prática da rabeca no fandango caiçara: estudo de caso com os rabequistas da família Pereira da comunidade do Ariri	Dissertação (Mestrado em Música)	4
1884/24650	Política tributária do Paraná na Primeira República (1890–1930)	Dissertação (Mestrado em História)	4

FONTE: O autor (2017).

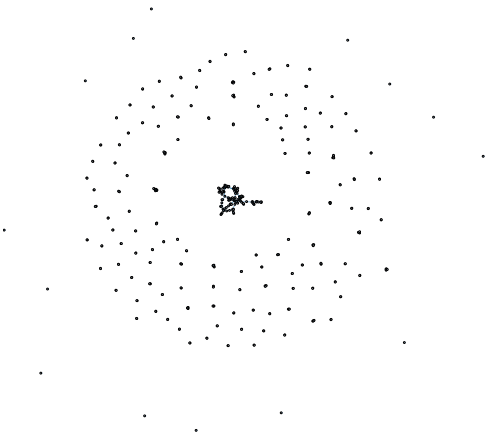
Deve-se notar que este é um resultado parcial: por motivos ainda desconhecidos, mas sob investigação, apenas pouco menos de metade dos documentos presentes na BDTD estão disponíveis para colheita via OAI-PMH, de modo que os documentos faltantes não foram utilizados para esta avaliação.

FIGURA 12 – Documentos e relacionamentos por referências bibliográficas na Biblioteca Digital de Teses e Dissertações da UFPR



FONTE: O autor (2017).

FIGURA 13 – Visualização de rede de citações gerada sobre parte da Biblioteca Digital de Teses e Dissertações da UFPR



FONTE: O autor (2017).

4.3 DIFICULDADES

A maior complicação encontrada durante os testes foi a inadequação da segmentação de referências feita pelo ParsCit aos formatos utilizados na UFPR, seja ao não distinguir o título dos autores, ou ao não encontrar o ponto correto de separação entre referências (embora este possa também ser atribuído ao extrator de texto e não apenas ao extrator de referências). O segundo caso é particularmente problemático, pois gera referências incompletas (que não podem ser combinadas) e referências múltiplas (que poderiam ser ligadas mais de uma vez), conforme ilustrado pelo exemplo nas figuras 14 e 15.

FIGURA 14 – Exemplo de referências bibliográficas

ALBURQUERQUE, FRANCISCO; ZAPATA, TANIA.A importância da estratégia de desenvolvimento local/territorial no Brasil, *in:* DOWBOR Ladislau; POCHMANN Marcio. Políticas para o desenvolvimento local. Perseu Abramo: São Paulo, 2010. Disponível em: <http://www.sct.rs.gov.br/upload/1353522830_A%20import%C3%A2ncia%20da%20estrat%C3%A9gia%20de%20desenvolvimento%20local%20territorial%20no%20Brasil%20-%20Alburquerque%20&%20Zapatapdf>Acesso em: 02 fev. 2017.

ANDRIGUETTO FILHO, JOSÉ MILTON; MARCHIORO, NILSON DE PAULA XAVIER. **Diagnóstico e problemática para a pesquisa**. in: RAYNAUT, C. et al. Desenvolvimento e meio ambiente: em busca da interdisciplinaridade: pesquisas urbanas e rurais. Ed UFPR: Curitiba, 2002.

FONTE: Silva (2017, p. 119).

A qualidade da extração realizada pelo ParsCit depende consideravelmente da qualidade do processo de extração de texto. O apêndice D traz as referências extraídas a partir de uma documento criado em meio digital, e cujo texto é lido corretamente pelo *kit* de ferramentas Xpdf. O apêndice E, por sua vez, apresenta o resultado da extração a partir de um documento

FIGURA 15 – Instâncias de referências extraídas, demonstrando a separação incorreta de referências bibliográficas

```
!Reference &ref57500:
  citing_work: *work846
  title: "TANIA.A importância da estratégia de
   ↪ desenvolvimento local/territorial no Brasil, in:
   ↪ DOWBOR Ladislau; POCHMANN Marcio. Políticas
   ↪ para o desenvolvimento local. Perseu Abramo: São
   ↪ Paulo,"
  year: 2010
  authors:
    - "FRANCISCO ALBURQUERQUE"
    - "ZAPATA"

!Reference &ref57501:
  citing_work: *work846
  title: "sil%20-%20Alburquerque%20&%20Zapatapdf>Acesso
   ↪ em: 02 fev."
  year: 2017
  authors:
    - "ANDRIGUETTO FILHO"
    - "JOSÉ MILTON"
    - "NILSON DE PAULA XAVIER MARCHIORO"

!Reference &ref57502:
  citing_work: *work846
  title: "Desenvolvimento e meio ambiente: em busca da
   ↪ interdisciplinaridade: pesquisas urbanas e rurais. Ed
   ↪ UFPR:"
  year: 2002
  authors: []
```

FONTE: O autor (2017).

datilografado e posteriormente digitalizado. As ferramentas Xpdf, em sua configuração padrão, não reconhecem corretamente a ordem do texto, levando à baixa qualidade das referências extraídas.

Outro problema encontrado ocorre quando um trabalho acadêmico possui título similar à outra obra que pode ser referenciada; por exemplo, dissertações de Letras que discorrem sobre um dado livro e são intituladas no formato "O Objeto de Estudo: Algum Comentário", ou trabalhos que dão origem a um artigo de nome idêntico ou semelhante. A primeira situação resulta em um falso positivo quando a referência não é segmentada corretamente e faltam os dados dos autores citados. Já a segunda é impossível de se resolver apenas com os metadados utilizados neste trabalho; porém, para referências bibliográficas formatadas no padrão ABNT NBR 6023, seria possível separar entre diferentes tipos de obras referenciadas

através da existência de certos marcadores, como indicação de tipo (obrigatória para trabalhos acadêmicos), números de volume ou edição (para artigos em periódicos), a expressão “In:” (para conferências), entre outros.

Em relação ao repositório digital, algumas práticas identificadas sobre os metadados dos registros bibliográficos tornaram mais difícil a ligação de referências. Em vários casos, o subtítulo do trabalho não está presente nos metadados, reduzindo a margem de confiança em combinações corretas onde a referência inclui o subtítulo, e assim podendo levar à falsos negativos no processo. Em outros casos, a data de publicação do registro é, de fato, a data na qual o registro foi adicionado ao repositório, resultando em um campo que não combina com as referências quando se trata de registros antigos disponibilizados recentemente. Por fim, o campo de metadados que denota os autores de um trabalho por vezes inclui o orientador ou até mesmo a própria instituição, este último introduzindo vários falsos positivos durante testes que comparavam os textos das referências bibliográficas como um todo (isto é, sem serem segmentados).

5 CONCLUSÃO

O presente trabalho teve como objetivo prover os meios para a construção de uma rede de citações para a Biblioteca Digital de Teses e Dissertações da Universidade Federal do Paraná, através da qual é possível descobrir as influências e caminhos de comunicação na produção acadêmica da Universidade. A existência desta rede de citações, que pode no futuro vir a ser ampliada à outras bibliotecas digitais da UFPR, abre caminho para a criação de indicadores que auxiliem a tomada de decisões por parte dos programas de pós-graduação, dos setores e da Pró-reitoria de Pesquisa e Pós-graduação no que tange à produção acadêmico-científica da UFPR.

O desenvolvimento de uma nova ferramenta para o alcance deste objetivo se fez necessário devido às deficiências nos sistemas já existentes, envolvendo o corpo restrito de obras indexadas e o suporte à geração de subgrafos a partir das rede de citações produzidas, necessário para o cálculo de indicadores relativos a uma única instituição. Apesar disso, os trabalhos anteriores na área foram insumo para o planejamento da arquitetura do *software* aqui desenvolvido, e serão fonte de inspiração para os próximos passos a serem dados.

A avaliação da ferramenta se deu através do processamento de um conjunto de 7.275 teses e dissertações, das quais foram extraídas 657.253 referências bibliográficas. Dessas, 457 foram identificadas como referências a outros trabalhos presentes na mesma base, e 390 desses foram validados manualmente, resultando em uma taxa de precisão de 85,34%.

As possibilidades de desenvolvimentos futuros se concentram em volta de quatro tópicos: criação de novos serviços para integração com outros repositórios digitais, abordagens alternativas para extração de referências e novos métodos de *record linkage*; seleção, ajuste e avaliação dos serviços e algoritmos já implementados para adaptação a repositórios digitais específicos; migração para novas plataformas a fim de suportar necessidades incomuns, como processamento de documentos em larga escala; e o cálculo de indicadores sobre a rede de citações.

Como exemplo das possibilidades do uso da rede de citações gerada por este trabalho, embora esta rede não possa ser diretamente utilizada para identificar as influências externas ao repositório sobre sua produção, é possível aplicar o processo de deduplicação nas referências extraídas e sintetizar registros para os trabalhos referenciados a partir de grupos de referências similares. Indo além, pode-se mapear esta rede sintética de documentos para uma rede de autores, a qual permitiria a identificação dos indivíduos mais influentes sobre a produção sendo analisada.

REFERÊNCIAS

BALDI, Stéphane; HARGENS, L. Reassessing the N-rays reference network: The role of self citations and negative citations. **Scientometrics**, Kluwer, Dordrecht, v. 34, n. 2, p. 239–253, 1995. DOI: 10.1007/BF02020422.

BALDI, Stéphane; HARGENS, Lowell L. Re-examining Price's Conjectures on the Structure of Reference Networks: Results from the Special Relativity, Spatial Diffusion Modeling and Role Analysis Literatures. **Social Studies of Science**, SAGE, Thousand Oaks, v. 27, n. 4, p. 669–687, 1 ago. 1997. DOI: 10.1177/030631297027004004. Disponível em: <http://www.jstor.org/stable/285561>. Acesso em: 8 dez. 2017.

BARRUECO, José Manuel; KRICHEL, Thomas. Building an autonomous citation index for grey literature: the Economics working papers case. In: INTERNATIONAL CONFERENCE ON GREY LITERATURE, 6., 2004, New York. **Proceedings**. Amsterdam: TextRelease, 2005. HDL: 10760/5879. Acesso em: 18 ago. 2016.

BEAGRIE, Neil; LAVOIE, Brian; WOOLLARD, Matthew. **Keeping Research Data Safe 2**: Final Report. Salisbury, UK, 2010. 88 p. Disponível em: <http://repository.essex.ac.uk/2147/>. Acesso em: 7 dez. 2017.

BERLEZE, Sérgio Luiz Meister. **Efeitos pelicular e de proximidade em condutores nao-magnéticos**. 1988. Dissertação (Mestrado em Física) – Universidade Federal do Paraná, Curitiba, 1988. HDL: 1884/36657.

CAPLAN, Priscilla. Reference Linking for Journal Articles: Promise, Progress and Perils. **portal:** Libraries and the Academy, Johns Hopkins University Press, Baltimore, v. 1, n. 3, p. 351–356, 2001. DOI: 10.1353/pla.2001.0036.

CAVALIERI, Osvaldo Marcio. **Um método complementar ao processo de sanitização de registros duplicados em bases de dados CADSUS-multiplataforma**. 2014. Dissertação (Mestrado em Informática) – Universidade Federal do Paraná, Curitiba, 2014. HDL: 1884/36297.

CHRISTEN, Peter; CHURCHES, Tim; HEGLAND, Markus. Febrl: A Parallel Open Source Data Linkage System. In: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 8., 2004, Sydney. **Advances in Knowledge Discovery and Data Mining**. Berlin: Springer, 2004. p. 638–647. DOI: 10.1007/978-3-540-24775-3_75.

CLARIVATE ANALYTICS. **Web of Science Databases**. 2017. Disponível em: <https://clarivate.com/products/web-of-science/databases/>. Acesso em: 10 dez. 2017.

COUNCILL, Isaac G.; GILES, C. Lee; KAN, Min-Yen. ParsCit: An open-source CRF reference string parsing package. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 6., 2008, Marrakech. **Proceedings**. [S.l.: s.n.], 2008. Disponível em: <http://lrec-conf.org/proceedings/lrec2008/pdf/166_paper.pdf>. Acesso em: 4 ago. 2016.

CROW, Raym. The Case for Institutional Repositories: A SPARC Position Paper. **ARL**, Association of Research Libraries, Washington, D.C., n. 223, p. 1–4, 2002. Disponível em: <https://sparcopen.org/wp-content/uploads/2016/01/instrepo.pdf>. Acesso em: 24 nov. 2017.

DE SOLLA PRICE, Derek J. **Little Science, Big Science...and Beyond**. New York: Columbia, 1986. 301 p.

_____. Networks of Scientific Papers. **Science**, American Association for the Advancement of Science, Washington, D.C., v. 149, n. 3683, p. 510–515, 30 jul. 1965. DOI: 10.1126/science.149.3683.510.

DI IORIO, Angelo et al. Describing bibliographic references in RDF. In: WORKSHOP ON SEMANTIC PUBLISHING, 4., 2014, Anissaras. **Proceedings**. Aachen: CEUR-WS, 2014. Disponível em: <http://ceur-ws.org/Vol-1155#paper-05>. Acesso em: 29 ago. 2017.

EGGHE, Leo; ROUSSEAU, Ronald. **Introduction to Informetrics**: Quantitative Methods in Library, Documentation and Information Science. Amsterdam: Elsevier, 1990. 450 p.

FEDORYSZAK, Mateusz; TKACZYK, Dominika; BOLIKOWSKI, Łukasz. Large Scale Citation Matching Using Apache Hadoop. In: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF DIGITAL LIBRARIES, 17., 2013, Valletta, Malta. **Research and Advanced Technology for Digital Libraries**. Berlin: Springer, 2013. p. 362–365. DOI: 10.1007/978-3-642-40501-3_37. Disponível em: <https://arxiv.org/abs/1303.6906v1>. Acesso em: 18 set. 2016.

FELLEGI, Ivan P.; SUNTER, Alan B. A Theory for Record Linkage. **Journal of the American Statistical Association**, Taylor & Francis, Boston, v. 64, n. 328, p. 1183–1210, 1969. DOI: 10.1080/01621459.1969.10501049.

FERREIRA, Elisabete. **Um Método de Coleta e Classificação de Metadados de Produção Científica em Repositórios Digitais Institucionais**. 2016. Dissertação (Mestrado em Informática) – Universidade Federal do Paraná, Curitiba, 2016. HDL: 1884/44782.

FU, Qiao; YUAN, Run. Institutional Citation Index in the application of scientific research achievements digital repository. In: IEEE INTERNATIONAL CONFERENCE ON PROGRESS IN INFORMATICS AND COMPUTING, 1., 2010, Shangai. **Proceedings**. New York: IEEE, 2010. p. 1207–1210. DOI: 10.1109/PIC.2010.5687972.

GARFIELD, Eugene. Can Citation Indexing be Automated? In: STATISTICAL ASSOCIATION METHODS FOR MECHANIZED DOCUMENTATION, 1964, Washington. **Symposium Proceedings**. Washington D.C.: National Bureau of Standards, 1965. p. 189–192. Google Books: r56ZrbfTdkYC. Acesso em: 3 nov. 2017.

_____. Citation frequency as a measure of research activity and performance. In: ESSAYS of an Information Scientist: 1962–1973. Philadelphia: ISI Press, 1977. v. 1, p. 406–408. Disponível em: <http://garfield.library.upenn.edu/essays/V1p406y1962-73.pdf>. Acesso em: 7 dez. 2017.

_____. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. **Science**, New York, v. 122, n. 3159, p. 108–111, 15 jul. 1955. DOI: 10.1126/science.122.3159.108.

_____. **Citation Indexing**: Its Theory and Application in Science, Technology, and Humanities. New York: Wiley, 1979. 274 p. Disponível em: <http://www.garfield.library.upenn.edu/ci/title.pdf>. Acesso em: 19 ago. 2016.

_____. The evolution of the Science Citation Index. **International Microbiology**, SEM, Madrid, v. 10, n. 1, p. 65–69, 2007. DOI: 10.2436/20.1501.01.10. Disponível em: <http://garfield.library.upenn.edu/papers/barcelona2007a.pdf>. Acesso em: 10 dez. 2017.

_____. The History and Meaning of the Journal Impact Factor. **JAMA**, Chicago, v. 295, n. 1, p. 90–93, 2006. DOI: 10.1001/jama.295.1.90.

GARSHOL, Lars Marius. **Bayesian identity resolution**. 11 fev. 2011. Disponível em: <http://www.garshol.priv.no/blog/217.html>. Acesso em: 5 dez. 2017.

GILES, C. Lee; BOLLACKER, Kurt D.; LAWRENCE, Steve. CiteSeer: An Automatic Citation Indexing System. In: ACM CONFERENCE ON DIGITAL LIBRARIES, 3., 1998, Pittsburgh, PA, USA. **DL '98**. New York: ACM, 1998. p. 89–98. DOI: 10.1145/276675.276685.

GILES, Jim. Science in the web age: Start your engines. **Nature**, Macmillan, London, n. 438, p. 554–555, 1 dez. 2005. DOI: 10.1038/438554a.

HARNAD, Stevan. The self-archiving initiative: Freeing the refereed research literature online. **Nature**, Macmillan, London, n. 410, p. 1024–1024, 26 abr. 2001. DOI: 10.1038/35074210.

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, Washington, D.C., v. 102, n. 46, p. 16569–16572, 7 nov. 2005. DOI: 10.1073/pnas.0507655102.

HITCHCOCK, S. et al. Citation Linking: Improving Access to Online Journals. In: ACM CONFERENCE ON DIGITAL LIBRARIES, 2., 1997, Philadelphia, PA, USA. **DL '97**. New York: ACM, 1997. p. 115–122. DOI: 10.1145/263690.263804.

HUANG, Mu-Hsuan. Exploring the hindex at the institutional level: A practical application in world university rankings. **Online Information Review**, Emerald, Bingley, v. 34, n. 4, p. 534–547, 2012. DOI: 10.1108/14684521211254059.

JAGUŠT, Tomislav; STOJANOVSKI, Jadranka; BARANOVIĆ, Mirta. Implementing the additional knowledge in the Croatian Scientific Bibliography. In: INTERNATIONAL CONVENTION ON INFORMATION AND COMMUNICATION TECHNOLOGY, ELECTRONICS AND MICROELECTRONICS, 37., 2014, Opatija. **Proceedings**. New York: IEEE, 2014. p. 1469–1472. DOI: 10.1109/MIPRO.2014.6859798.

JANTZ, Ronald; GIARLO, Michael J. Digital Preservation: Architecture and Technology for Trusted Digital Repositories. **Microform & Imaging Review**, De Gruyter, Berlin, v. 34, n. 3, p. 135–147, 23 set. 2005. DOI: 10.1515/MFIR.2005.135.

LAGOZE, Carl; VAN DE SOMPEL, Herbert. The making of the Open Archives Initiative Protocol for Metadata Harvesting. **Library Hi Tech**, Emerald, Bingley, v. 21, n. 2, p. 118–128, 2003. DOI: 10.1108/07378830310479776.

_____. The Open Archives Initiative: Building a low-barrier interoperability framework. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 1., 2001, Roanoke, VA, USA. **Proceedings**. New York: ACM, 2001. p. 54–62. DOI: 10.1145/379437.379449.

LARSEN, Peder Olesen; VON INS, Markus. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. **Scientometrics**, Springer, Dordrecht, v. 84, n. 3, p. 575–603, 2010. DOI: 10.1007/s11192-010-0202-z.

LAWRENCE, Steve; GILES, C. Lee; BOLLACKER, Kurt D. Autonomous Citation Matching. In: ANNUAL CONFERENCE ON AUTONOMOUS AGENTS, 3., 1999, Seattle, WA, USA. **Proceedings**. New York: ACM, 1999. p. 392–393. DOI: 10.1145/301136.301255.

LEVENSHTEIN, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. **Soviet Physics Doklady**, Nauka/Interperiodica, v. 10, n. 8, p. 707–710, 1966.

LI, Huajing et al. CiteSeer^x: an Architecture and Web Service Design for an Academic Document Search Engine. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 15., 2006, Edinburgh. **Proceedings**. New York: ACM, 2006. p. 883–884. DOI: 10.1145/1135777.1135926.

LIPINSKI, Mario et al. Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 13., 2013, Indianapolis, IN, USA. **Proceedings**. New York: ACM, 2013. p. 385–386. DOI: 10.1145/2467696.2467753. Disponível em: <http://docear.org/papers/Evaluation_of_Header_Metadata_Extraction_Approaches_and_Tools_for_Scientific_PDF_Documents.pdf>. Acesso em: 4 ago. 2016.

LYNCH, Clifford A. Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age. **portal:** Libraries and the Academy, Johns Hopkins University Press, Baltimore, v. 3, n. 2, p. 327–336, 2003. DOI: 10.1353/pla.2003.0039.

MEADOWS, A. J. **A Comunicação Científica**. Brasília: Briquet de Lemos, 1999. 261 p.

OPEN ARCHIVES INITIATIVE. **ORE User Guide**: Primer. 17 out. 2008. Disponível em: <http://www.openarchives.org/ore/1.0/primer>. Acesso em: 8 nov. 2017.

_____. **The Open Archives Initiative Protocol for Metadata Harvesting**. Version 2.0. 8 jan. 2015. Disponível em: <http://openarchives.org/OAI/openarchivesprotocol.html>. Acesso em: 8 nov. 2017.

ORACLE CORPORATION. **ServiceLoader**. Java SE 9 & JDK 9. 2017. Disponível em: <https://docs.oracle.com/javase/9/docs/api/java/util/ServiceLoader.html>. Acesso em: 13 dez. 2017.

PACKER, Abel L. et al. **XML, por quê?** 4 abr. 2014. Disponível em: <http://blog.scielo.org/blog/2014/04/04/xml-porque/>. Acesso em: 28 nov. 2017.

PERONI, Silvio. The Semantic Publishing and Referencing Ontologies. In: _____. **Semantic Web Technologies and Legal Scholarly Publishing. Law, Governance and Technology Series**. Cham: Springer, 2014. p. 121–193. (Law, Governance and Technology Series, 15). DOI: 10.1007/978-3-319-04777-5_5.

PERONI, Silvio et al. Setting our bibliographic references free: Towards open citation data. **Journal of Documentation**, Emerald, Bingley, v. 71, n. 2, p. 153–277, 2015. DOI: 10.1108/JD-12-2013-0166.

PRATHAP, Gangan. Hirsch-type indices for ranking institutions' scientific research output. **Current Science**, Current Science Association, Bengaluru, v. 91, n. 11, p. 1439, 10 dez. 2006. Disponível em: <http://www.currentscience.ac.in/Downloads/article_id_091_11_1439_1439_0.pdf>. Acesso em: 10 dez. 2017.

RAHM, Erhard; DO, Hong Hai. Data Cleaning: Problems and Current Approaches. **Bulletin of the Technical Committee on Data Engineering**, IEEE Computer Society, Washington D.C., v. 23, n. 4, p. 3–13, 2000. Disponível em: <https://dbs.uni-leipzig.de/file/TBDE2000.pdf>. Acesso em: 13 dez. 2017.

RUAS, Wilimar; FERREIRA, Marta. Análise de citações e ARS: Rede de referências em educação científica. **Revista ACB:** Biblioteconomia em Santa Catarina, Florianópolis, v. 21, n. 1, p. 156–166, 2016. Disponível em: <https://revistaacb.emnuvens.com.br/racb/article/view/1061>. Acesso em: 8 ago. 2016.

SAHA, Somnath; SAINT, Sanjay; CHRISTAKIS, Dimitri A. Impact factor: a valid measure of journal quality? **Journal of the Medical Library Association**, v. 91, n. 1, p. 42–46, 2003. PMID: 12572533.

SALVI, Jorge Luis. **Relacionamentos temporais entre redes de petri e planejamento automático**. 2009. Dissertação (Mestrado em Informática) – Universidade Federal do Paraná, Curitiba, 2009. HDL: 1884/21188.

SANDISON, Alexander. Documentation note: thinking about citation analysis. **Journal of Documentation**, Emerald, Bingley, v. 45, n. 1, p. 59–64, 1989. DOI: 10.1108/eb026839.

SAS INSTITUTE. **What Is ETL?** 24 mai. 2017. Disponível em: <https://www.sas.com/en_us/insights/data-management/what-is-etl.html>. Acesso em: 21 dez. 2017.

SCHIMIDT, Marion. Development and Evaluation of a Match Key for Linking References to Cited Articles. In: INTERNATIONAL CONFERENCE ON SCIENCE AND TECHNOLOGY INDICATORS, 17., 2012, Montréal, Canada. **Proceedings**. Montréal: Science-Metrix e OST, 2012. p. 707–718. Disponível em: <http://2012.sticonference.org/Proceedings/vol2/Schmidt_Development_707.pdf>. Acesso em: 29 nov. 2017.

SCHWARTZMAN, Simon. A Ciência da Ciência. **Ciência Hoje**, Sociedade Brasileira para o Progresso da Ciência, Rio de Janeiro, v. 2, n. 11, p. 54–59, 1984. Disponível em: <http://www.schwartzman.org.br/simon/ciencia2.htm>. Acesso em: 2 nov. 2017.

SILVA, Marcus Aurélio Soares da. **O microempreendedor individual (MEI) no litoral do Paraná**: uma análise da formalização sob a perspectiva do trabalho (2008–2016). 2017. Dissertação (Mestrado em Desenvolvimento Territorial Sustentável) – Universidade Federal do Paraná, Matinhos, 2017. HDL: 1884/48939.

SMALL, Henry; GRIFFITH, Belver C. The Structure of Scientific Literatures I: Identifying and Graphing Specialties. **Science Studies**, SAGE, Thousand Oaks, v. 4, n. 1, p. 17–40, 1974. DOI: 10.1177/030631277400400102. Disponível em: <http://www.jstor.org/stable/284536>. Acesso em: 8 dez. 2017.

SMITH, MacKenzie et al. DSpace: An Open Source Dynamic Digital Repository. **D-Lib Magazine**, CNRI, Reston, VA, USA, v. 9, n. 1, 2003. DOI: 10.1045/january2003-smith.

SPINAK, Ernesto. **Diccionario Enciclopédico de Bibliometría, Centometría e Informetría**. Caracas: UNESCO, 1996. 261 p. Disponível em: <http://unesdoc.unesco.org/images/0024/002433/243329S.pdf>. Acesso em: 8 dez. 2017.

SUNYE, Marcos et al. A experiência da UFPR na construção de repositórios digitais: a implantação integrada das ferramentas DSpace e Open Journal Systems. In: SAYÃO, Luis et al. (Org.). **Implantação e gestão de repositórions institucionais**: políticas, memória, acesso livre e preservação. Salvador: EDUFBA, 2009. p. 107–122. Disponível em: <https://repositorio.ufba.br/ri/handle/ufba/473>. Acesso em: 24 nov. 2017.

SUTINEN, Erkki; TARHIO, Jorma. On using q-gram locations in approximate string matching. In: EUROPEAN SYMPOSIUM ON ALGORITHMS, 3., 1995, Corfu, GR. **Algorithms — ESA '95**. Berlin: Springer, 1995. p. 327–340. DOI: 10.1007/3-540-60313-1_153. Disponível em: <http://www.cs.hut.fi/u/tarhio/papers/esa.pdf>. Acesso em: 13 dez. 2017.

TESTA, James. **Journal Selection Process**. 18 jul. 2016. Disponível em: <https://clarivate.com/essays/journal-selection-process/>. Acesso em: 11 dez. 2017.

TKACZYK, Dominika et al. CERMINE: automatic extraction of structured metadata from scientific literature. **International Journal on Document Analysis and Recognition**, Springer, Berlin, v. 18, n. 4, p. 317–335, 2015. DOI: 10.1007/s10032-015-0249-8.

VASSILIADIS, Panos. A Survey of Extract–Transform–Load Technology. **International Journal of Data Warehousing and Mining**, IGI Publishing, Hershey, v. 5, n. 3, p. 1–27, 2009. DOI: 10.4018/jdwm.2009070101.

WLASCHIN, Scott. **Railway Oriented Programming**: A functional approach to error handling. 2014. Disponível em: <https://fsharpforfunandprofit.com/rop/>. Acesso em: 26 set. 2017.

WU, Jian et al. **CiteSeerX Manual**. 19 ago. 2015. 63 p. Disponível em: <https://github.com/SeerLabs/CiteSeerX/blob/615d1809672ab05aa3528dade3e9054d1678f6d8/doc/cxm.pdf>.

YIOTIS, Kristin. Electronic theses and dissertation (ETD) repositories: What are they? Where do they come from? How do they work? **D-Lib Magazine**, Emerald, Bingley, v. 9, n. 1, p. 101–115, 2003. DOI: 10.1108/10650750810875458.

APÊNDICES

APÊNDICE A – EXEMPLO DE ARQUIVO DE CONFIGURAÇÃO DO SPRING BOOT

```
# Configuração de banco de dados Neo4j
spring.data.neo4j:
   embedded.enabled: false
   username: neo4j
   password: secret
# Configuração de banco de dados JPA (PostgreSQL)
spring.datasource:
   url: "jdbc:postgresql://localhost/cnb"
   username: cnb
   password: 180a27b6-988c-4cd0-bea1-ab7fd89285d3
# Caminho p/ arquivo de configuração de logging (Log4j2)
logging.config: log4j2.xml
```

APÊNDICE B – EXEMPLO DE ARQUIVO DE CONFIGURAÇÃO DE SERVIÇOS

```
<beans xmlns="http://www.springframework.org/schema/beans">

  <bean id="WorkFetcher" class="br.ufpr.inf.aasg13.cnb.WorkFetcher">
    <constructor-arg index="0" ref="cnb.worksources.ufpretd"/>
  </bean>

  <bean id="ReferenceExtractor"
  ↪   class="br.ufpr.inf.aasg13.cnb.ReferenceExtractor">
    <constructor-arg index="0" ref="remoteResourceReader"/>
  </bean>

  <bean id="cnb.pdfextractors.xpdf"
  ↪   class="br.ufpr.inf.aasg13.cnb.pdfextractors.xpdf.XpdfTextExtractor">
    <constructor-arg value="pdftotext"/>
  </bean>

  <bean id="cnb.referenceextractors.parscit" class="br.ufpr.inf.aasg13.cnb
  ↪   .referenceextractors.parscit.ParscitReferenceExtractor">
    <constructor-arg index="2"
    ↪   value="/home/bcc/aasg13/cnb/deps/parscit.sh"/>
  </bean>

  <bean id="cnb.referencelinkers.duke"
  ↪   class="br.ufpr.inf.aasg13.cnb.referencelinkers.duke.DukeLinker">
    <constructor-arg value="duke.xml"/>
  </bean>

</beans>
```

APÊNDICE C – CONFIGURAÇÃO PARA *RECORD LINKAGE*

```
<duke>
  <object class="no.priv.garshol.duke.comparators.QGramComparator"
  ↪   name="BigramComparator">
    <param name="q" value="2"/>
    <param name="tokenizer" value="POSITIONAL"/>
  </object>
  <object class="no.priv.garshol.duke.cleaners.RegexpCleaner"
  ↪   name="IsoYearCleaner">
    <param name="regexp" value="^(\d+)"/>
  </object>

  <database class="no.priv.garshol.duke.databases.LuceneDatabase">
    <param name="path" value="index"/>
    <param name="fuzzy-search" value="false"/>
    <param name="min-relevance" value="0.9"/>
    <param name="boost-mode" value="INDEX"/>
  </database>

  <schema>
    <threshold>0.8</threshold>
    <maybe-threshold>0.5</maybe-threshold>

    <property type="id">
      <name>ID</name>
    </property>
    <property>
      <name>TEXT</name>
      <comparator>BigramComparator</comparator>
      <low>0.3</low>
      <high>0.9</high>
    </property>
    <property>
      <name>AUTHORS</name>
      <comparator>BigramComparator</comparator>
      <low>0.2</low>
      <high>0.7</high>
```

```
    </property>
    <property>
      <name>YEAR</name>
      <comparator>no.priv.garshol.duke.comparators.Levenshtein</comparator>
      <low>0.4</low>
      <high>0.55</high>
    </property>
  </schema>

  <group>
    <csv>
      <param name="input-file" value="cnb-20171101-works.csv"/>
      <param name="encoding" value="utf-8"/>
      <column name="id" property="ID"/>
      <column name="title" property="TEXT"
      ↪  cleaner="no.priv.garshol.duke.cleaners.LowerCaseNormalizeCleaner"/>
      <column name="year" property="YEAR" cleaner="IsoYearCleaner"/>
      <column name="authors" property="AUTHORS"
      ↪  cleaner="no.priv.garshol.duke.cleaners.LowerCaseNormalizeCleaner"/>
    </csv>
  </group>

  <group>
    <csv>
      <param name="input-file" value="cnb-20171101-references.csv"/>
      <param name="encoding" value="utf-8"/>
      <column name="id" property="ID"/>
      <column name="title" property="TEXT"
      ↪  cleaner="no.priv.garshol.duke.cleaners.LowerCaseNormalizeCleaner"/>
      <column name="year" property="YEAR" cleaner="IsoYearCleaner"/>
      <column name="authors" property="AUTHORS"
      ↪  cleaner="no.priv.garshol.duke.cleaners.LowerCaseNormalizeCleaner"/>
    </csv>
  </group>
</duke>
```

APÊNDICE D – REFERÊNCIAS BIBLIOGRÁFICAS EM SALVI (2009) EXTRAÍDAS E SEGMENTADAS ATRAVÉS DO PARSCIT

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
[1] SILVA, F. Rede de Planos: Uma proposta para a Solu¸c˜ao de Problemas de Planejamento em Intelig^encia Artificial usando Redes de Petri. Tese (Doutorado) — Centro Federal de Educa¸c˜ao Tecnol´ogica do Paran´a - CEFET-PR, Curitiba, PR, Brasil, fev. 2005.			
	Rede de Planos: Uma proposta para a Solucao de Problemas de Planejamento em Inteligencia Artificial usando Redes de Petri. Tese (Doutorado)	F SILVA	2005
[2] HICKMOTT, S. L. et al. Planning via petri net unfolding. In: VELOSO, M. M. (Ed.). IJCAI. [s.n.], 2007. p. 1904–1911. Dispon´ıvel em: <http://dblp.uni-trier.de/ db/conf/ijcai/ijcai2007.html>.			
	Planning via petri net unfolding. In:	S L HICKMOTT	2007
[3] PETRY, F. C. Planejamento Aplicado ‘a Verifica¸c˜ao de Bloqueios em Redes de Petri. Disserta¸c˜ao (Mestrado) — Universidade Federal do Paran´a - UFPR, Curitiba, PR, Brasil, ago. 2008.			
	Planejamento Aplicado a Verificacao de Bloqueios em Redes de Petri. Dissertacao (Mestrado) — Universidade Federal do Parana - UFPR,	F C PETRY	2008
[4] EDELKAMP, S.; JABBAR, S. Action planning for directed model checking of petri nets. Electronic Notes in Theoretical Computer Science, v. 149, n. 2, p. 3–18, 2006.			

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
	Action planning for directed model checking of petri nets.	S EDELKAMP; S JABBAR	2006
[5] SMITH, D. E. Planning Formalisms Commentary. set. 2007. ICAPS-07 Planning Formalisms Session Commentary. Dispon´ıvel em: <http: //ti.arc.nasa.gov/ people/ de2smith/ publications.html>.			
	Planning Formalisms Commentary. set.	D E SMITH	2007
[6] CUSHING, W. et al. When is temporal planning really temporal? In: VELOSO, M. M. (Ed.). IJCAI. [s.n.], 2007. p. 1852–1859. Dispon´ıvel em: <http://dblp.unitrier.de/db/conf/ijcai/ijcai2007.html>.			
	When is temporal planning really temporal? In:	W CUSHING	2007
[7] COLES, A. I. et al. Planning with problems requiring temporal coordination. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI 08). [S.l.: s.n.], 2008.			
	Planning with problems requiring temporal coordination. In:	A I COLES	2008
[8] GEREVINI, A. Automated planning in temporal domains: Some recent advances and current research topics. In: TIME. IEEE Computer Society, 2007. p. 3–4. ISBN 978-0- 7695-2836-6. Dispon´ıvel em: <http://dblp.uni-trier.de/db/conf/time/time2007.html>.			
	Automated planning in temporal domains: Some recent advances and current research topics. In:	A GEREVINI	2007
[9] NILSSON, N. J.; FIKES, R. E. STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. Artifical Intelligence, v. 2, n. 3-4, p. 189–208, 1971.			

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
	STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving.	N J NILSSON; R E FIKES	1971
[10] VILA, L. A survey on temporal reasoning in artificial intelligence. AI Communications, v. 7, n. 1, p. 4–28, 1994. Dispon´ıvel em: <citeseer.ist.psu.edu/vila94survey.html>.			
	A survey on temporal reasoning in artificial intelligence.	L VILA	1994
[11] PENBERTHY, J. S.; WELD, D. S. Temporal planning with continuous change. In: Proceedings of the 12th National Conference on Artificial Intelligence. Seattle, WA, USA: AAAI, 1994. v. 2, p. 1010–1015. Dispon´ıvel em: <http://dblp.unitrier.de/db/conf/aaai/aaai94-2.html>.			
	Temporal planning with continuous change. In:	D S WELD; J S PENBERTHY	1994
[12] YOUNES, H. L. S.; SIMMONS, R. G. VHPOP: Versatile Heuristic Partial Order Planner. J. Artif. Intell. Res. (JAIR), v. 20, p. 405–430, 2003. Dispon´ıvel em: <http://dblp.uni-trier.de/db/journals/jair/jair20.html>.			
	VHPOP: Versatile Heuristic Partial Order Planner.	H L S YOUNES; R G SIMMONS	2003
[13] DO, M. B.; KAMBHAMPATI, S. Sapa: A multi-objective metric temporal planner. J. Artif. Intell. Res. (JAIR), v. 20, p. 155–194, 2003. Dispon´ıvel em: <http://dblp.unitrier.de/db/journals/jair/jair20.html>.			
	A multi-objective metric temporal planner.	M B DO; S Sapa KAMBHAMPATI	2003
[14] HALSEY, K.; LONG, D.; FOX, M. CRIKEY - A Planner Looking at the Integration of Scheduling and Planning. In: Proceedings of the Workshop on Integration Scheduling Into Planning at 13th International Conference on Automated Planning and Scheduling (ICAPS'03). [S.l.: s.n.], 2004. p. 46–52.			

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
	CRIKEY - A Planner Looking at the Integration of Scheduling and Planning. In:	D LONG; K HALSEY; M FOX	2004
[15] MCDERMOTT, D. et al. PDDL - The Planning Domain Definition Language. New Haven, CN, USA, 1998.			
	PDDL - The Planning Domain Definition Language.	D MCDERMOTT	1998
[16] RINTANEN, J. Complexity of concurrent temporal planning. In: BODDY, M.; FOX, M.; THIEBAUX, S. (Ed.). ´ Proceedings of the Seventeenth International Conference on Automated Planning and Scheduling (ICAPS'07). Providence, Rhode Island, USA: [s.n.], 2007. v. 17, p. 280–287.			
	Complexity of concurrent temporal planning. In:	J RINTANEN	
			2007
[17] MURATA, T. Petri nets: Properties, analysis and applications. In: Proceedings of the IEEE. Los Alamitos, CA, USA: IEEE, 1989. v. 7, n. 4, p. 541–580.			
	Petri nets: Properties, analysis and applications. In:	T MURATA	1989
[18] PETERSON, J. L. Petri nets. ACM Comput. Surv., v. 9, n. 3, p. 223–252, 1977. Dispon´ıvel em: <http://dblp.uni-trier.de/db/journals/csur/csur9.html>.			
	Petri nets.	J L PETERSON	1977
[19] ALLEN, J. F. Planning as temporal reasoning. In: ALLEN, J. F.; FIKES, R.; SANDEWALL, E. (Ed.). KR'91: Principles of Knowledge Representation and Reasoning. San Mateo, California: Morgan Kaufmann, 1991. p. 3–14. Dispon´ıvel em: <citeseer.ist.psu.edu/allen91planning.html>.			
	Planning as temporal reasoning. In:	J F ALLEN	1991
[20] LAVALLE, S. M. Planning Algorithms. Cambridge, U.K.: Cambridge University Press, 2006. Available at http://planning.cs.uiuc.edu/.			

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
	Planning Algorithms.	S M LAVALLE	2006
[21] SMITH, D.; FRANK, J.; JONSSON, A. Bridging the gap between planning and ´ scheduling. Knowledge Engineering Review, v. 15, n. 1, p. 47–83, 2000.			
	Bridging the gap between planning and scheduling.	A JONSSON; D SMITH; J FRANK	2000
[22] RUSSELL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. 2nd edition. ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2003. 1132 p. ISBN 0-13-790395-2.			
	Intelligence: A Modern Approach. 2nd edition.		2003
[23] PEDNAULT, E. P. D. ADL: Exploring the Middle Ground Between STRIPS and the Situation Calculus. In: KR. [s.n.], 1989. p. 324–332. Dispon´ıvel em: <http://dblp.uni-trier.de/db/conf/kr/kr89.html>.			
	ADL: Exploring the Middle Ground Between STRIPS and the Situation Calculus. In:	E P D PEDNAULT	1989
[24] FOX, M.; LONG, D. PDDL2.1: An Extension to PDDL for Expressing Temporal Planning Domains. Journal of Artificial Intelligence Research (JAIR), v. 20, p. 61–124, 2003. Dispon´ıvel em: <http://dblp.uni-trier.de/db/journals/jair/jair20.html>.			
	PDDL2.1: An Extension to PDDL for Expressing Temporal Planning Domains.	D LONG; M FOX	2003
[25] VELOSO, M. M. Learning By Analogical Reasoning in General Problem Solving. Tese (PhD) — Carnegie Mellon University, Pittsburgh, PA, U.S.A., 1992.			

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
	Learning By Analogical Reasoning in General Problem Solving. Tese (PhD) —	M M VELOSO	1992
[26] CUSHING, W. et al. Evaluating temporal planning domains. In: BODDY, M.; FOX, M.; THIEBAUX, S. (Ed.). ´ Proceedings of the Seventeenth International Conference on Automated Planning and Scheduling (ICAPS'07). Providence, Rhode Island, USA: [s.n.], 2007. v. 17.			
	Evaluating temporal planning domains. In:	W CUSHING	
			2007
[27] WULLINGER, P. Transformation of Temporally Expressive Into Temporally Simple Planning Problems. Disserta¸c˜ao (Mestrado) — University of Bamberg, Bamberg, Bavaria, Alemanha, set. 2007.			
	Transformation of Temporally Expressive Into Temporally Simple Planning Problems. Dissertacao (Mestrado)	P WULLINGER	2007
[28] GHALLAB, M.; NAU, D.; TRAVERSO, P. Automated Planning: Theory and Practice. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2004. 663 p. ISBN 1-55860-856-7.			
	Automated Planning: Theory and Practice.	D NAU; M GHALLAB; P TRAVERSO	2004
[29] ALLEN, J. F. Maintaining knowledge about temporal intervals. Communications of the ACM, v. 26, p. 832–843, 1983.			
	Maintaining knowledge about temporal intervals.	J F ALLEN	1983
[30] LEVER, J.; RICHARDS, B. A planning architecture using temporal constraint solving. London, Greater London, Inglaterra, jan. 1993.			

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
	A planning architecture using temporal constraint solving.	B RICHARDS; J LEVER	1993

[31] VILAIN, M. B.; KAUTZ, H. A. Constraint propagation algorithms for temporal reasoning. In: AAAI. [s.n.], 1986. p. 377–382. Dispon´ıvel em: <http://dblp.uni-trier.de/db/conf/aaai/aaai86-1.html>.

	Constraint propagation algorithms for temporal reasoning. In:	H A KAUTZ; M B VILAIN	1986

[32] ZAIDI, A. K. On temporal logic programming using petri nets. IEEE Transactions on Systems, Man, and Cybernetics, Part A, v. 29, n. 3, p. 245–254, maio 1999. Dispon´ıvel em: <http://dblp.uni-trier.de/db/journals/tsmc/tsmca29.html>.

	On temporal logic programming using petri nets.	A K ZAIDI	1999

[33] DECHTER, R.; MEIRI, I.; PEARL, J. Temporal constraint networks. Artificial Intelligence, v. 49, n. 1-3, p. 61–95, 1991.

	Temporal constraint networks.	I MEIRI; J PEARL; R DECHTER	1991

[34] MEIRI, I. Combining qualitative and quantitative constraints in temporal reasoning. In: DEAN, T.; MCKEOWN, K. (Ed.). Proceedings of the Ninth National Conference on Artificial Intelligence. Menlo Park, California: AAAI Press, 1991. p. 260–267. Dispon´ıvel em: <citeseer.ist.psu.edu/meiri95combining.html>.

	Combining qualitative and quantitative constraints in temporal reasoning.	I MEIRI	1991

[35] BARBER, F. Reasoning on interval and point-based disjunctive metric constraints in temporal contexts. Journal of Artificial Intelligence Research, v. 12, p. 35–86, 2000. Dispon´ıvel em: <citeseer.ist.psu.edu/barber00reasoning.html>.

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
	Reasoning on interval and point-based disjunctive metric constraints in temporal contexts.	F BARBER	2000
[36] SMITH, D. E.; WELD, D. S. Temporal planning with mutual exclusion reasoning. In: DEAN, T. (Ed.). IJCAI. Morgan Kaufmann, 1999. p. 326–337. ISBN 1-55860-613-0. Dispon´ıvel em: <http://dblp.uni-trier.de/db/conf/ijcai/ijcai99.html>.			
	Temporal planning with mutual exclusion reasoning. In:	D E SMITH; D S WELD	1999
[37] COLES, A. et al. Managing coordination in temporal planning using plannerscheduler interaction. Preprint submitted to Elsevier Science. nov. 2007. Dispon´ıvel em: <http://personal.cis.strath.ac.uk/ ac/52426/crikeyaij>.			
	Managing coordination in temporal planning using plannerscheduler interaction. Preprint submitted to Elsevier Science.	A COLES	2007
[38] PETRI, C. A. Kommunikation mit Automaten. Tese (Doutorado) — Bonn: Institut fur Instrumentelle Mathematik, Schriften des IIM Nr. 2, 1962. Sec ¨ ond Edition:, New York: Griffiss Air Force Base, Technical Report RADC-TR-65–377, Vol.1, 1966, Pages: Suppl. 1, English translation: Communication with Automata.			
	Kommunikation mit Automaten. Tese (Doutorado) — Bonn: Institut fur Instrumentelle Mathematik,	C A PETRI	1962

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
[39] BALBO, G. et al. Petri nets 2000: Introductory tutorial. In: NIELSEN, M.; SIMPSON, D. (Ed.). 21st International Conference on Application and Theory of Petri Nets (ICATPN). Aarhus, Denmark: Springer, 2000. (Lecture Notes in Computer Science, v. 1825), p. 26–30. ISBN 3-540-67693-7. DOS SLIDES. Dispon´ıvel em: <http://www. informatik. uni-hamburg. de/ TGI/PetriNets/ introductions/ pn2000 introtut.pdf>.			
	Petri nets 2000: Introductory tutorial. In:	G BALBO	2000
[40] BUCHHOLZ, P. Petri nets. Course of International Center for Computational Logic - Technischen Universit¨at Dresden. 2002.			
	Petri nets.	P BUCHHOLZ	2002
[41] ZURAWSKI, R.; ZHOU, M. Petri nets and industrial applications: A tutorial. Industrial Electronics, IEEE Transactions on, v. 41, n. 6, p. 567–583, dez. 1994. Dispon´ıvel em: <http://dx.doi.org/10.1109/41.334574>.			
	Petri nets and industrial applications: A tutorial. Industrial Electronics,	M ZHOU; R ZURAWSKI	1994
[42] CARDOSO, J.; VALETTE, R. Redes de Petri. Florian´opolis, SC, Brasil: UFSC, 1997. 212 p.			
		J CARDOSO; R Redes de Petri VALETTE	1997
[43] J´uNIOR", N. M. Redes de Petri Temporais: M´etodo de An´alise Baseado em Tempo Global. Disserta¸c˜ao (Mestrado) — Universidade Federal do Paran´a - UFPR, Curitiba, PR, Brasil, fev. 2008.			
	Redes de Petri Temporais: Metodo de Analise Baseado em Tempo Global. Dissertacao (Mestrado) — Universidade Federal do Parana - UFPR,	N M JuNIOR"	2008

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
[44] RAMCHANDANI, C. Analysis of Asynchronous Concurrent Systems by Timed Petri Nets. Tese (Doutorado) — Cambridge, Mass.: MIT, Dept. Electrical Engineering, fev. 1974. Project MAC TR-120.			
	Analysis of Asynchronous Concurrent Systems by Timed Petri Nets. Tese (Doutorado) —	C RAMCHANDANI	
[45] SIFAKIS, J. Use of petri nets for performance evaluation. In: BEILNER, H.; GELENBE, E. (Ed.). Performance. Bad Godesberg, Bonn, Germany: North-Holland, 1977. p. 75–93. ISBN 0-444-85058-9.			
	Use of petri nets for performance evaluation. In:	J SIFAKIS	1977
[46] SIFAKIS, J. Performance evaluation of systems using nets. In: BRAUER, W. (Ed.). Proceedings of the Advanced Course on General Net Theory of Processes and Systems. London, UK: Springer-Verlag, 1980. (Lecture Notes in Computer Science, v. 84), p. 307–319. ISBN 3-540-10001-6. Dispon´ıvel em: <http://dblp.unitrier.de/db/conf/ac/nt.html>.			
	Performance evaluation of systems using nets. In:	J SIFAKIS	1980
[47] MERLIN, P. M. A Study of Recoverability of Computer Systems. Tese (Doutorado) — University of California, Dep. Comput. Sci, Irvine, CA, USA, 1974.			
	A Study of Recoverability of Computer Systems. Tese (Doutorado)	P M MERLIN	1974
[48] MELZER, S.; ROMER, S. Deadlock checking using net unfoldings. In: GRUMBERG, ¨ O. (Ed.). Computer Aided Verification, 9th International Conference, CAV '97. Haifa, Israel: Springer, 1997. (Lecture Notes in Computer Science, v. 1254), p. 352–363. ISBN 3-540-63166-6. Dispon´ıvel em: <http://dblp.uni-trier.de/db/conf/cav/cav97.html>.			
	Deadlock checking using net unfoldings. In:	S MELZER; S ROMER	1997

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
[49] CORBETT, J. C. Evaluating deadlock detection methods for concurrent software. IEEE Trans. Software Eng., v. 22, n. 3, p. 161–180, 1996. Dispon´ıvel em: <http://dblp.uni-trier.de/db/journals/tse/tse22.html>.			
	Evaluating deadlock detection methods for concurrent software.	J C CORBETT	1996
[50] HOFFMANN, J. The Metric-FF Planning System: Translating "Ignoring Delete Lists" to Numeric State Variables. J. Artif. Intell. Res. (JAIR), v. 20, p. 291–341, 2003. Dispon´ıvel em: <http://dblp.uni-trier.de/db/journals/jair/jair20.html>.			
	The Metric-FF Planning System: Translating "Ignoring Delete Lists" to Numeric State Variables.	J HOFFMANN	2003
[51] HELJANKO, K.; SIMONS, P. mcsmodels 1.4: a deadlock and reachability checker using net unfoldings. 1999. Helsinki University of Technology, Laboratory for Theoretical Computer Science, Espoo, Finland. Software.			
	mcsmodels 1.4: a deadlock and reachability checker using net unfoldings.	K HELJANKO; P SIMONS	1999
[52] ESPARZA, J.; ROMER, S.; VOGLER, W. An improvement of McMillan's unfolding ¨ algorithm. Form. Methods Syst. Des., Kluwer Academic Publishers, Hingham, MA, USA, v. 20, n. 3, p. 285–310, 2002. ISSN 0925-9856.			
	An improvement of McMillan's unfolding algorithm. Form. Methods Syst.	J ESPARZA; S ROMER; W VOGLER	2002

APÊNDICE E – REFERÊNCIAS BIBLIOGRÁFICAS EM BERLEZE (1988) EXTRAÍDAS E SEGMENTADAS ATRAVÉS DO PARSCIT

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
1 ARNOLD, A.H.M. Proximity effect in soliâ and hollow round conductors. Jou r . I E E , 82:537-45, 1938.			
2 BECCARI, C. & RONCA, C. L'effetto pellicolare in conduttori a nastro. L 1Elettrotecnica, 56(10):607-13, 1969.			
3 BELEVITCH, V. ; GUERRET, P.; LIÉNARD, J.C. Le skin-effect dans un ruban. Rev. H F , !>(5) :109-15, 1962.			
4 BROWN, H.E. Grandes sistemas elétricos - métodos matri ciais. São Paulo, Editora LTC/EFEI, 1977. s.p.			
5 BURKE, P.E. & ALDEN, R.T.H. Current density probes. IEEE Trans., P A S - 8 8 (2):181-5, Feb.1969.			
6 CAHEN, F. La repartition des courants dans les conducteurs massifs. Techniques de L 'Ingénieur, D-130:l-10, 1951.			
7 CASIMIR, H.B.G. & UBBINK, J. The skin effect. Philips Technical Review, 28(9):271-83 , 1967.			
8 CHEMICAL RUBBER COMPANY. Handbook of chemistry and physics. 51.st ed. Ohio, 1951.			
9 DALEY, J.L. Current distribution in a rectangular conductor. Tran s . A I E E , 58:687-91, 1939.			
10 DWIGHT, H.B. Skin effect in tubular and flat conductors. Trans AIEE, 37(2):1379-403, 1918.			
11 EDELMANN, H. J.C. Maxwell's geometric mean distances. Siemens Forsch*-u. Entwickl.-B e r . , 10(3):133-8, 1981.			1981
12 GOLDING, E.W. & WIDDIS, F.C. Electrical measurements and measuring instruments. 5.th ed. London, Pitman Paperbacks, 1963. Cap. 5.			
13 GRANEAU, P. Alternating and transient conduction currents in straight conductors of any cross-section. In t . J. Electronics, 3ŷ:41-59, 1965.	Alternating and transient conduction currents in straight conductors of any cross-section.	Pitman Paperbacks London	1963
		Electronics	1965

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
14 GROSS, H.G. Die Berechnung der Stromverteilung in zylindrischen Leitern mit rechteckigem und elliptischem Querschnitt. Arch. Elektrotech. , 2M:241-68, 1940.			
15 GUILLOT, M. La production de champs magnétiques intenses transitoires: les effets secondaires deviennent prépondérants. Revue Générale de L'Ëlectricité, 9:37-48, o c t . 1987.			
	La production de champs magnétiques intenses transitoires: les effets secondaires deviennent prépondérants.	Elektrotech	1940
16 HIGGINS, T.J. The origins and developments of the concepts of inductance, skin effect and proximity effect. American Journal of Physics, 9/6) î337-46, Dec.1941.			
	The origins and developments of the concepts of inductance, skin effect and proximity effect.		
17 ISELBORN, K.W. & WEIß, P. A numerical method for computation of current density distribution in conductors of circular cross-section. Fourth International Symposium on High Voltage Engineering, 13(4):l-4, 1983.			
	A numerical method for computation of current density distribution in conductors of circular cross-section.	Dec 1941 ISELBORN; K W; P WEIß	
18 KENNELLY, A.E. & AFFEL, H.A. Skin effect resistance measurements of conductors. Proc. IRE, £:523-55, D e c . 1916.			
19 LONDON, F. & LONDON, H. P r o c . R o y . S o c . A, 14 9 :71, 1935.			
20 MALEWSKI, R. Measurements of transient skin effect within nonlinear conductors. IEEE Tra n s ., P A S - 9 1 (5):1881-6, S e p t . / O c t . 1972.			

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
	measurements of conductors.	A E KENNELLY; H A AFFEL	1983
21 MANNEBACK, C. An integral equation for skin effect in parallel conductors. J o u r . M a t h . Phys ., 1:123-46, 1922.			
			1922
22 MATHEWS, J. & WALKER, R.L. Mathematical methods of phys i c s . 2.nd ed. California, Addison Wesley, 1973. 501 p.			
	Mathematical methods of California,		1973
23 MATVEEV, A.N. Electricity and m a gnet i s m . Moscow, Mir Publishers, 1986.			
	Moscow	1986	
24 MAXWELL, J.C. A treatise on electricity and ma gnetism. N.Y., Dover Publication Inc., 1954. 2 v., s.p.			
25 M E I , K. & BLADEL, J. Low-frequency scattering by rectangular cylinders. IEEE Trans. on Antennas and Propagation,ˆ:52-6, Jan.1963.			
		rectangular cylinders	
26 PARIS, D.T. & HURD, F.K. Teoria eletromagnétlca bâsica. Rio de Janeiro, Editora Guanabara Dois, 1983. 514 p.			
27 PIPPARD, A.B. P r o c . Ro y . S o c . A, 216:547, 1953.			
28 POINCELOT, P. Précis d *Ëlectromagnétisme théorique. Paris, Ed. Dunod, 1963. 456 p.			
29 PRESS, A. Resistance and reactance of massed rectangular conductors. Physical Rev i e w , 8(4):417-22, 1916.			
	Resistance and reactance of massed rectangular conductors.	Dunod	1963
	Physical Rev i e w ,	F W GROVER	1916
30 ROSA, E.B. & GROVER, F.W. Formulas and tables for the calculation of mutual and self-inductance. National Bureau of Standard, 8(l):l-223, 1912.			
	Physical Rev i e w ,	F W GROVER	1916

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
31 SCHAFFER, G. & BANDERET, P. L'effet Kelvin dans les barres pour grande intensité de courant. Revue Brown Boveri, 52(8):623-8, 1969.			
	L' effet Kelvin dans les barres pour grande intensité de courant.	G SCHAFFER; P BANDERET	1912
		Revue Brown Boveri	1969
32 SCHWENKHAGEN, H. Untersuchungen über Stromverdrangung in rechteckigen Querschnitten. A r c h . Elektrotech., 17:537-89, 1926-1927.			
	Untersuchungen über Stromverdrangung in rechteckigen Querschnitten. 17:537-89, 1926-1927. A r c h	H; H W PICKERING; SHIH Elektrotech	
33 SHIH, H. & PICKERING, H.W. Three-dimensional modeling of the potential and current distributions in an electrolytic cell. Jou r n . Electrochem. S o c . , 1 3 4 (3): 551-8, Mar.1987.			
34 SILVESTER, P. AC resistance and reactance of isolated rectangular conductors. IEEE Trans., PAS-8 6 (6) :770-4, June 1967.			
	AC resistance and reactance of isolated rectangular conductors.	S o c	1987
		IEEE Trans	1967
35_____ . Modal network theory of skin effect in flat conductors. Pr o c . IEEE, 5j4 (9) :1147-51, Sept.1966.			
	Modal network theory of conductors. _____ . skin effect in flat Pr o c .		1966
36 _____ . The accurate calculation of skin effect in conductors of complicated shape. IEEE Trans., PAS-87(3)t735-42, Mar.1968.			

Texto original da referência			
	Título	**Autores**	**Ano de publicação**
	r a n s ., PAS-87(3)t735-42, Mar.1968. 37_____ . Skin effect in multiple and polyphase conductors.	T IEEE	
37_____ . Skin effect in multiple and polyphase conductors. IEEE Trans., P A S- 8 8 (3):231-8, Mar.1969.			
	r a n s ., PAS-87(3)t735-42, Mar.1968. 37_____ . Skin effect in multiple and polyphase conductors.	T IEEE	
			1969
38 _____ . Campos eletromagnéticos modernos. Sao Paulo, Editora Polígono, 1971. s.p.			
39 WIAK, S. & ZAKRZEWSKI, K. Numerical calculation of transients in electrical circuits containing elements with nonlinear eddy-current skin effect. IEE Proc. , 1 3 4 (9A)s 741-6, Nov.1987.			
40 ZERVAS, M.N. & KRIEZIS, E.E. Integral formulation for the calculation of the field and the forces in a system of conducting cylindrical shells: a general approach. IEE P r o c . , 134(58):269-75, Sept.1987.			

ANEXOS

ANEXO A – MAPA DE RECURSO OAI-ORE SERIALIZADO POR ATOM, EMBUTIDO EM RESPOSTA OAI-PMH

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/">
  <responseDate>2017-12-08T04:23:40Z</responseDate>
  <request verb="GetRecord" identifier="oai:dspace.c3sl.ufpr.br:1884/18388"
  ↪  metadataPrefix="ore">http://acervodigital.ufpr.br/oai/request</request>
  <GetRecord>
    <record>
      <header>
        <identifier>oai:dspace.c3sl.ufpr.br:1884/18388</identifier>
        <datestamp>2016-04-07T09:59:15Z</datestamp>
        <setSpec>com_1884_39643</setSpec>
        <setSpec>com_1884_284</setSpec>
        <setSpec>col_1884_39646</setSpec>
      </header>
      <metadata>
        <atom:entry xmlns:atom="http://www.w3.org/2005/Atom"
        ↪  xmlns:dcterms="http://purl.org/dc/terms/"
        ↪  xmlns:oreatom="http://www.openarchives.org/ore/atom/">
          <atom:id>http://hdl.handle.net/1884/18388/ore.xml</atom:id>
          <atom:link rel="alternate" href="http://hdl.handle.net/1884/18388"/>
          <atom:link rel="http://www.openarchives.org/ore/terms/describes"
          ↪  href="http://hdl.handle.net/1884/18388/ore.xml"/>
          <atom:link type="application/atom+xml" rel="self"
          ↪  href="http://hdl.handle.net/1884/18388/ore.xml#atom"/>
          <atom:published>2009-11-10T18:36:47Z</atom:published>
          <atom:updated>2009-11-10T18:36:47Z</atom:updated>
          <atom:source>
            <atom:generator>RepositÃ³rio Digital Institucional da Universidade
            ↪  Federal do ParanÃ¡</atom:generator>
          </atom:source>
          <atom:title>Avaliação da atividade reprodutiva da ictiofauna capturada
          ↪  na pesca artesanal de arrasto camaroeiro pela comunidade de
          ↪  Itapema do Norte, Itapoa, litoral norte de Santa
          ↪  Catarina</atom:title>
          <atom:author>
            <atom:name>Pina, Juliana Ventura de</atom:name>
```

```
</atom:author>
<atom:category label="Aggregation"
↪  term="http://www.openarchives.org/ore/terms/Aggregation"
↪  scheme="http://www.openarchives.org/ore/terms/"/>
<atom:category
↪  scheme="http://www.openarchives.org/ore/atom/modified"
↪  term="2009-11-10T18:36:47Z"/>
<atom:category label="DSpace Item" term="DSpaceItem"
↪  scheme="http://www.dspace.org/objectModel/"/>
<atom:link rel="http://www.openarchives.org/ore/terms/aggregates"
↪  href="http://acervodigital.ufpr.br/bitstream/1884/18388/1/
↪  DISSERTACAO_JULIANA%20VENTURA%20DE%20PINA.pdf"
↪  title="DISSERTACAO_JULIANA VENTURA DE PINA.pdf"
↪  type="application/pdf" length="2861830"/>
<oreatom:triples>
  <rdf:Description
   ↪  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   ↪  rdf:about="http://hdl.handle.net/1884/18388/ore.xml#atom">
    <rdf:type rdf:resource="http://www.dspace.org/objectModel/
     ↪  DSpaceItem"/>
    <dcterms:modified>2009-11-10T18:36:47Z</dcterms:modified>
  </rdf:Description>
  <rdf:Description
   ↪  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   ↪  rdf:about="http://acervodigital.ufpr.br/bitstream/1884/18388/
   ↪  1/DISSERTACAO_JULIANA%20VENTURA%20DE
   ↪  %20PINA.pdf">
    <rdf:type rdf:resource="http://www.dspace.org/objectModel/
     ↪  DSpaceBitstream"/>
    <dcterms:description>ORIGINAL</dcterms:description>
  </rdf:Description>
  <rdf:Description
   ↪  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   ↪  rdf:about="http://acervodigital.ufpr.br/bitstream/1884/18388/
   ↪  2/DISSERTACAO_JULIANA%20VENTURA%20DE
   ↪  %20PINA.pdf.txt">
    <rdf:type rdf:resource="http://www.dspace.org/objectModel/
     ↪  DSpaceBitstream"/>
    <dcterms:description>TEXT</dcterms:description>
```

```
                </rdf:Description>
                <rdf:Description
                ↪  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
                ↪  rdf:about="http://acervodigital.ufpr.br/bitstream/1884/18388/
                ↪  3/DISSERTACAO_JULIANA%20VENTURA%20DE
                ↪  %20PINA.pdf.jpg">
                  <rdf:type rdf:resource="http://www.dspace.org/objectModel/
                  ↪  DSpaceBitstream"/>
                  <dcterms:description>THUMBNAIL</dcterms:description>
                </rdf:Description>
              </oreatom:triples>
            </atom:entry>
          </metadata>
        </record>
      </GetRecord>
</OAI-PMH>
```

ANEXO B – EXCERTO DA SAÍDA XML DO PARSCIT PARA O PRESENTE TRABALHO

```
<algorithms version="110505">
  <algorithm name="ParsCit" version="110505">
    <citationList>
      <citation valid="true">
        <authors>
          <author>C L GILES COUNCILL</author>
          <author>KAN</author>
        </authors>
        <date>2008</date>
        <marker>COUNCILL, KAN, 2008</marker>
        <rawString>COUNCILL; C. L. GILES; KAN, 2008.</rawString>
      </citation>
      <citation valid="true">
        <authors>
          <author>Stéphane BALDI</author>
          <author>L HARGENS</author>
        </authors>
        <title>Reassessing the N-rays reference network: The role of self citations
        ↪ and negative citations.</title>
        <date>1995</date>
        <journal>Scientometrics, Kluwer, Dordrecht,</journal>
        <volume>34</volume>
        <pages>239--253</pages>
        <marker>BALDI, HARGENS, 1995</marker>
        <rawString>BALDI, Stéphane; HARGENS, L. Reassessing the N-rays
        ↪ reference network: The role of self citations and negative citations.
        ↪ Scientometrics, Kluwer, Dordrecht, v. 34, n. 2, p. 239–253,
        ↪ 1995.</rawString>
      </citation>
      <citation valid="false">
        <pages>10--1007</pages>
        <marker/>
        <rawString>DOI: 10.1007/BF02020422.</rawString>
      </citation>
      <citation valid="true">
        <authors>
```

```
      <author>Stéphane BALDI</author>
      <author>Lowell L HARGENS</author>
    </authors>
    <title>Re-examining Price's Conjectures on the Structure of Reference
    ↪ Networks: Results from the Special Relativity, Spatial Diffusion
    ↪ Modeling and Role Analysis Literatures.</title>
    <date>1997</date>
    <journal>Social Studies of Science, SAGE, Thousand Oaks,</journal>
    <volume>27</volume>
    <pages>669--687</pages>
    <marker>BALDI, HARGENS, 1997</marker>
    <rawString>BALDI, Stéphane; HARGENS, Lowell L. Re-examining
    ↪ Price's Conjectures on the Structure of Reference Networks: Results
    ↪ from the Special Relativity, Spatial Diffusion Modeling and Role
    ↪ Analysis Literatures. Social Studies of Science, SAGE, Thousand
    ↪ Oaks, v. 27, n. 4, p. 669–687, 1 ago. 1997. DOI:
    ↪ 10.1177/030631297027004004. Disponível em:
    ↪ &amp;amp;lt;http://www.jstor.org/stable/ 285561&amp;amp;gt;. Acesso em:
    ↪ 8 dez. 2017.</rawString>
  </citation>
  <citation valid="true">
    <authors>
      <author>José Manuel BARRUECO</author>
      <author>Thomas KRICHEL</author>
    </authors>
    <title>Building an autonomous citation index for grey literature: the
    ↪ Economics working papers case.</title>
    <date>2004</date>
    <journal>In: INTERNATIONAL CONFERENCE ON GREY
    ↪ LITERATURE,</journal>
    <volume>6</volume>
    <pages>10760--5879</pages>
    <location>New York. Proceedings. Amsterdam: TextRelease,</location>
    <marker>BARRUECO, KRICHEL, 2004</marker>
    <rawString>BARRUECO, José Manuel; KRICHEL, Thomas. Building an
    ↪ autonomous citation index for grey literature: the Economics
    ↪ working papers case. In: INTERNATIONAL CONFERENCE ON
    ↪ GREY LITERATURE, 6., 2004, New York. Proceedings.
    ↪ Amsterdam: TextRelease, 2005. HDL: 10760/5879.</rawString>
```

```
</citation>
<citation valid="true">
  <authors>
    <author>Acesso em</author>
  </authors>
  <title>18 ago.</title>
  <date>2016</date>
  <marker>em, 2016</marker>
  <rawString>Acesso em: 18 ago. 2016. BEAGRIE, Neil; LAVOIE, Brian;
  ↪ WOOLLARD, Matthew. Keeping Research Data Safe 2: Final
  ↪ Report.</rawString>
</citation>
<citation valid="true">
  <authors>
    <author>UK Salisbury</author>
  </authors>
  <title>88 p. Disponível em: Acesso em: 7 dez.</title>
  <date>2010</date>
  <marker>Salisbury, 2010</marker>
  <rawString>Salisbury, UK, 2010. 88 p. Disponível em:
  ↪ &amp;amp;lt;http://repository.essex.ac.uk/2147/&amp;amp;gt;. Acesso em: 7
  ↪ dez. 2017. BERLEZE, Sérgio Luiz Meister. Efeitos pelicular e de
  ↪ proximidade em condutores nao-magnéticos.</rawString>
</citation>
<citation valid="true">
  <title>Dissertação (Mestrado em Física) – Universidade Federal
  ↪ do</title>
  <date>1988</date>
  <volume>1884</volume>
  <pages>36657</pages>
  <location>Paraná, Curitiba,</location>
  <contexts>
```

```
    <context position="12929" citStr="(1988)" startWordPosition="4221"
     ↪ endWordPosition="4221">. . . . . 31 31 32 34 5
     ↪ CONCLUSÃO . . . . . .
     ↪ 36 REFERÊNCIAS . . . . . .
     ↪ . . . . . 37 APÊNDICES 43 APÊNDICE A EXEMPLO DE
     ↪ ARQUIVO DE CONFIGURAÇÃO DO SPRING BOOT .
     ↪ APÊNDICE B EXEMPLO DE ARQUIVO DE CONFIGURAÇÃO
     ↪ DE SERVIÇOS APÊNDICE C CONFIGURAÇÃO PARA
     ↪ RECORD LINKAGE . . . . . 46 APÊNDICE D
     ↪ REFERÊNCIAS BIBLIOGRÁFICAS EM SALVI (2009)
     ↪ EXTRAÍDAS E SEGMENTADAS ATRAVÉS DO PARSCIT . . . .
     ↪ . . . . . 48 APÊNDICE E REFERÊNCIAS
     ↪ BIBLIOGRÁFICAS EM BERLEZE (1988) EXTRAÍDAS E
     ↪ SEGMENTADAS ATRAVÉS DO PARSCIT . . . . . 60 .
     ↪ . . 45 ANEXOS ANEXO A 44 66 MAPA DE RECURSO
     ↪ OAI-ORE SERIALIZADO POR ATOM, EMBUTIDO EM
     ↪ RESPOSTA OAI-PMH . . . . . 67 10 1
     ↪ INTRODUÇÃO O número de vezes que um determinado trabalho
     ↪ ou autor é citado por outros é um indicador de seu impacto no
     ↪ campo da pesquisa científica, tanto diretamente como componente
     ↪ de métricas como o índice h (HIRSCH, 2005) e o fator de impacto
     ↪ de periódicos (GARFIELD, 2006). Para obter este número, é
     ↪ preciso identificar quais os trabalhos que fazem referência
     ↪ a</context>
  </contexts>
  <marker>1988</marker>
  <rawString>Dissertação (Mestrado em Física) – Universidade Federal do
   ↪ Paraná, Curitiba, 1988. HDL: 1884/ 36657.</rawString>
</citation>
<citation valid="true">
  <authors>
    <author>Priscilla CAPLAN</author>
  </authors>
  <title>Reference Linking for Journal Articles: Promise, Progress and
   ↪ Perils. portal: Libraries and the Academy, Johns Hopkins</title>
  <date>2001</date>
  <volume>1</volume>
  <pages>351--356</pages>
  <publisher>University Press,</publisher>
```

```
    <location>Baltimore,</location>
    <marker>CAPLAN, 2001</marker>
    <rawString>CAPLAN, Priscilla. Reference Linking for Journal Articles:
    ↪ Promise, Progress and Perils. portal: Libraries and the Academy,
    ↪ Johns Hopkins University Press, Baltimore, v. 1, n. 3, p. 351–356,
    ↪ 2001.</rawString>
  </citation>
  <citation valid="false">
    <pages>10--1353</pages>
    <marker/>
    <rawString>DOI: 10.1353/pla.2001.0036.</rawString>
  </citation>
  <citation valid="true">
    <authors>
      <author>Osvaldo Marcio CAVALIERI</author>
    </authors>
    <title>Um método complementar ao processo de sanitização de registros
    ↪ duplicados em bases de dados CADSUS-multiplataforma.</title>
    <date>2014</date>
    <booktitle>Dissertação (Mestrado em Informática) – Universidade
    ↪ Federal do Paraná,</booktitle>
    <pages>1884--36297</pages>
    <location>Curitiba,</location>
    <contexts>
```

```
            <context position="31325" citStr="CAVALIERI, 2014"
             ↪ startWordPosition="7086" endWordPosition="7087"> (2012)
             ↪ propõe o uso de chaves derivadas de dados bibliográficos para
             ↪ combinação de citações. Fedoryszak, Tkaczyk e Bolikowski (2013)
             ↪ utiliza SVMs e CRFs em conjunto com o framework Apache
             ↪ Hadoop1 para combinação de citações em larga escala. 2.4
             ↪ RECORD LINKAGE O processo de record linkage, também
             ↪ conhecido como resolução de entidades ou deduplicação, almeja
             ↪ identificar registros que correspondem à um mesmo indivíduo,
             ↪ objeto ou evento (FELLEGI; SUNTER, 1969). Assume-se que tais
             ↪ registros são similares, mas não necessariamente idênticos, do
             ↪ contrário seria trivial a identificação de duplicatas (CAVALIERI,
             ↪ 2014). O primeiro passo para a deduplicação é a limpeza e
             ↪ padronização dos dados, de modo a remover ou corrigir dados
             ↪ inconsistentes que podem levar a resultados errôneos nas etapas
             ↪ posteriores, e a eliminar diferenças entre representações de um
             ↪ mesmo valor, tornando mais fácil a comparação de registros
             ↪ (CHRISTEN; CHURCHES; HEGLAND, 2004). Os procedimentos
             ↪ exatos variam de acordo com as necessidades de cada conjunto de
             ↪ dados, mas costumam incluir técnicas como normalização de
             ↪ strings (envolvendo conversão para todas letras maiúsculas ou
             ↪ minúsculas, remoção de diacríticos e caracteres
             ↪ não-alfanuméri</context>
          </contexts>
          <marker>CAVALIERI, 2014</marker>
          <rawString>CAVALIERI, Osvaldo Marcio. Um método complementar ao
           ↪ processo de sanitização de registros duplicados em bases de dados
           ↪ CADSUS-multiplataforma. Dissertação (Mestrado em Informática) –
           ↪ Universidade Federal do Paraná, Curitiba, 2014. HDL:
           ↪ 1884/36297.</rawString>
        </citation>
      </citationList>
    </algorithm>
</algorithms>
```