IVISON DA COSTA RUBIM

MAPEAMENTO DE INCONSISTÊNCIAS NA PLATAFORMA LATTES

Dissertação apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Mestre. Área de Concentração: Engenharia de Software.

Orientador: Profa. Vanessa Braganholo Murta

Niterói

2014

Ficha Catalográfica elaborada pela Biblioteca da Escola de Engenharia e Instituto de Computação da UFF

R896 Rubim, Ivison da Costa
Mapeamento de inconsistências na Plataforma Lattes. –
Niterói, RJ : [s.n.], 2014.
87f.

Dissertação (Mestrado em Computação) - Universidade Federal Fluminense, 2014.
Orientador(a): Vanessa Braganholo Murta

1. Base de dados. 2. Plataforma Lattes. 3. Currículo. 4. Ambiguidade de dados. 5. Desambiguação de dados. 6. Inconsistência de dados. I. Título.

CDD 005.1

IVISON DA COSTA RUBIM

MAPEAMENTO DE INCONSISTÊNCIAS NA PLATAFORMA LATTES

Dissertação apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Mestre. Área de Concentração: Engenharia de Software.

Aprovada em Setembro de 2014.

BANCA EXAMINADORA

_____

Prof. D.Sc. VANESSA BRAGANHOLO MURTA – Orientador

UFF

_____

Prof. D.Sc. MARIA LUIZA MACHADO CAMPOS

UFRJ

_____

Prof. D.Sc. JOSÉ VITERBO FILHO

UFF

Niterói

2014

À Maria Amélia, Ine, Ian (in memoria), Ilson e netos.

AGRADECIMENTOS

A todas as pessoas que eu encontrei ao longo da vida. Represento o resultado desses encontros. Agradeço especialmente aos meus avós. O vovô Genaro ensinou-me que uma batalha é ganha por aquele que acreditar na vitória por mais tempo.

Deixemos alguém numa situação melhor do que ela estava quando nós a encontramos. Estou tentando Vô!

A vovó Isaura pelo carinho que ultrapassa o tempo. Obrigado pelas vitaminas e o profundo carinho. Compartilho tudo vovó!

Ao vovô Elízio por me ensinar que toda a construção exige esforço e uma habilidade só explicável pela experiência. Sempre cheguei na hora combinada vovô querido. A vovó "Chichica" me ensinou a ser paciente e a ouvir a voz divina.

Ao meu querido e saudoso pai Wilson e minha mãe Alayde pelo esforço, afeto e dedicação na minha criação. O seu garoto "Pai" está indo! Obrigado pelas colchas de retalhos mamãe!

Aos meus tios e tias que foram pessoas fundamentais na minha formação. Sou um privilegiado por tê-los conhecido.

A Maria Amélia por ser amor eterno, pela dedicação e pelos filhos que tivemos. Só nós sabemos o que vivemos e como vivemos!

Aos meus amados filhos Ine, Ian (saudades) e Ilson. Três seres lindos e que me norteiam até hoje.

A minha primeira netinha Maria Clara, carinhosamente chamada de "Picute", por me despertar para a visão lúdica, própria da infância. Prometo meu ser a você!

A Força Aérea Brasileira e ao Serviço Federal de Processamento de Dados por sedimentarem em mim a necessidade de evolução. Sou e serei sempre um servidor da sociedade.

A todos os professores do IC, especialmente minha orientadora Vanessa Braganholo Murta, pelos ensinamentos e compreensão com minhas limitações e dedicação. Muito obrigado Professora!

Ao Professor Leonardo Gresta Paulino Murta (Leo) você foi fundamental. Levarei suas aulas sempre comigo!

A professora Simone de Lima Martins pela especial e inesquecível compreensão com minhas limitações no início do mestrado.

A todos os colegas de mestrado pelas dicas, auxílios e, sobretudo, pela companhia. Agradeço especialmente aos colegas.

David Barreto pela parceira em LABGC. Disciplina muito forte, tal como você.

Cristiano Cesário por aceitar a parceria em GC e ISMA. Foi gratificante pensarmos juntos. Aprendi muito com você!

Gleiph e a Karen pelos conselhos e sugestões sobre como me conduzir no mestrado.

Ao STI/UFF por disponibilizar os Currículos Lattes do corpo docente da UFF, viabilizando a realização desse trabalho.

Aos funcionários da secretaria do IC pelos auxílios e orientações, especialmente a Teresa!

Por fim agradeço as incertezas por estimularem minha curiosidade e me manterem atento e disposto a buscar mais esclarecimentos.

Tudo deveria se tornar o mais simples possível, mas não mais simples do que isso.

(Albert Einsten)

RESUMO

A Plataforma Lattes vem adquirindo expressividade crescente como principal fonte de informações referentes à comunidade de pesquisadores brasileiros, estudantes, gestores e demais atores do sistema nacional de ciência, tecnologia e inovação. Entretanto, a integridade desse relevante instrumento de aferição da produção bibliográfica nacional pode ser afetada pelo efeito da ambiguidade ou inconsistências em citações de coautoria.

Um primeiro passo para a solução desse problema reside na identificação de tais inconsistências. Dessa forma este trabalho traça um mapa de inconsistências identificadas em citações de coautoria nos segmentos "Artigos Publicados" e "Trabalhos em Eventos", ambos componentes da Plataforma Lattes. Para tanto, foi adotada uma heurística especificada a partir de um levantamento de hipóteses que teve por base a análise de um volume expressivo de currículos Lattes, combinada com um tratamento de similaridade.

A abordagem para esse trabalho está fortemente amparada em procedimentos que vêm sendo praticados visando tratar dois conceitos que atingem notadamente as bibliotecas digitais; quais sejam: a duplicidade e a ambiguidade. Esses fenômenos distorcem resultados de consultas na medida em que admitem como iguais objetos diferentes ou consideram diferentes objetos iguais.

Palavras-chave: Deduplicação, Desambiguação, Inconsistência, Plataforma Lattes.

ABSTRACT

The Lattes Platform is acquiring increasing expressiveness as the main source of information regarding the community of Brazilian researchers, students, managers and other actors in the national system of science, technology and innovation. However, the integrity of this important tool for gauging the national bibliographic production may be affected by the effect of ambiguities or inconsistencies in citations coauthoring.

A first step in order to solve this problem lies in identifying such inconsistencies. Thereby this work traces a map of inconsistencies identified in coauthoring in journal and conference papers in the Lattes Platform. For that, we adopted a specific heuristic developed from a survey of hypothesis where a significant volume of Lattes curriculum was analyzed, combined with a treatment of similarity.

The approach to this work is strongly supported by procedures that have been performed aiming to address two concepts that mainly affect the digital libraries, which are: duplication and ambiguity. These phenomena distort query results, because they admit as equals different objects or different objects as equal.

Keywords: Deduplication, Disambiguation, Inconsistency, Lattes Platform.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ALIAS: *Active Learning Led Interactive Alias Suppression*

CL: Currículo Lattes

CNPq: Conselho Nacional de Pesquisa

IC: Instituto de Computação

LCS: *Longest Common Subsequence*

MII: Mapa de Inconsistências Individualizadas

MIS: Mapa de Inconsistências Sumarizadas

UFF: Universidade Federal Fluminense

SUMÁRIO

CAPÍTULO 1 – INTRODUÇÃO

1.1 MOTIVAÇÃO

A expressividade do Currículo Lattes (CL) como principal fonte de informações referente à comunidade de pesquisadores brasileiros, estudantes, gestores, profissionais e demais atores do sistema nacional de Ciência, Tecnologia e Inovação vem crescendo significativamente. Esse fato pode ser observado na Tabela 1 que apresenta o volume de currículos cadastrados discriminando-os por área de atuação. Essa crescente expressividade torna esse instrumento um dos principais recursos para a aferição e acompanhamento da produção de conhecimento científico no Brasil.

A Plataforma Lattes vem sendo intensamente utilizada e, atualmente, constitui um importante recurso não só para apoiar ações de planejamento, gestão e operacionalização do fomento do CNPq, como também de outras agências de fomento federais e estaduais, de fundações estaduais de apoio à ciência e tecnologia, de instituições de ensino superior e de institutos de pesquisa. Na atualidade é possível afirmar que a Plataforma Lattes se tornou estratégica não só para as atividades de planejamento e gestão, mas também para a formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação.

Tabela 1 - Totais de Currículo Lattes

Totais de Currículos cadastrados: 3.029.468			
Doutores	**Mestres**	**Estudantes**	**Outros**
184.143	325.271	1.248.258	1.271.796

Em um breve histórico, conforme citado pelo Conselho Nacional de Pesquisa (2012) a preocupação em manter uma base de currículos dos pesquisadores brasileiros começa a adquirir corpo em meados dos anos 80, com a adoção de um formulário padrão para registro dos currículos dos pesquisadores brasileiros. Nessa época, os órgãos governamentais já se ressentiam de um cadastro de currículos que pudesse subsidiá-los em ações dirigidas à definição de políticas voltadas para ciência e tecnologia.

Essa iniciativa evoluiu e no final dos anos 90 surge o sistema de informações denominado Plataforma Lattes, como resultado de um trabalho conjunto tendo como participantes o grupo Stela, vinculado à Universidade Federal de Santa Catarina, o grupo C.E.S.A.R[1] vinculado à Universidade Federal de Pernambuco, a empresa Multisoft e

[1] http://www.cesar.org.br/site/cesar/organizacao/

profissionais das Superintendências de Informática e Planejamento do CNPq. Esse sistema tem como principal componente o Currículo Lattes (CL).

Mais recentemente, em julho de 2005, o CNPq criou a Comissão para Avaliação do Lattes, com o objetivo de avaliar, reformular e aprimorar a Plataforma Lattes, tendo como premissas básicas para o seu processo evolutivo, torná-la mais racional, prática e confiável. Essa iniciativa tem um papel relevante no reconhecimento internacional desse instrumento como um exemplo significativo de boas práticas no sentido de disponibilizar uma ampla e confiável base de dados para a extração de métricas (LANE, 2010). Contudo, conforme Ferreira, Gonçalves e Laender (2012) a ambiguidade no contexto de citações bibliográficas está se tornando um problema de amplitude universal, que afeta a qualidade dos serviços, o conteúdo de bibliotecas digitais e sistemas similares, tais como a própria Plataforma Lattes. O desafio em lidar com esse fenômeno abrange diversos métodos de deduplicação e desambiguação. Esses métodos, de forma geral, tentam encontrar alguma similaridade entre os registros de citações de autoria e podem explorar tanto técnicas supervisionadas (BORGES; BECKER; *et al.*, 2011; SARAWAGI; BHAMIDIPATY, 2002) quanto não supervisionadas (BORGES; CARVALHO; *et al.*, 2011; YANG *et al.*, 2008).

A notória expressividade da Plataforma Lattes bem como o seu grande potencial para se tornar a principal fonte de informações a respeito da produção científica nacional, desperta o interesse para a verificação de possíveis inconsistências referenciais especificamente relacionadas com citações de coautoria nos segmentos “Artigos Publicados” e “Trabalhos em Eventos”. Esses dois segmentos armazenam atributos qualificadores de cada uma das publicações de um determinado CL. Logo, provavelmente, esses seriam os segmentos acessados tanto para o fornecimento de informações a respeito da produção científica nacional quanto para a geração de indicadores bibliométricos sobre o mesmo tema. Portanto, por analogia, é aceitável supor que essas possíveis inconsistências referenciais possuem características e efeitos semelhantes aos causados pelos fenômenos da ambiguidade e duplicidade nas bibliotecas digitais. Sendo assim, existe possibilidade que um determinado artigo publicado ou trabalho em evento informado num determinado currículo contenha inconformidades em relação aos currículos dos seus respectivos coautores. Essas inconsistências, se verificadas, poderão comprometer a confiabilidade desse importante instrumento de gestão, além de contribuir para aumentar o esforço despendido na preparação dos dados necessários à geração de indicadores bibliométricos fundamentais à análise da produção científica nacional (MUGNAINI; JANNUZZI; QUONIAM, 2004).

Dongwon L. *et al*. (2007) consideram a manutenção da consistência de citações em produção científica como uma questão não trivial, envolvendo os seguintes fatores: erro na entrada de dados, formato da citação, falta ou não aplicação de normas, falta ou não cumprimento de padrões, imperfeições em software de coleta de citações, nomes comuns de autores, abreviações comuns para locais de publicação e citação de dados em larga escala. Por analogia, é aceitável supor a incidência de alguns desses fatores como causa de inconsistências no CL.

O interesse por esse trabalho está centrado na busca por uma sistemática de aferição do nível de inconsistência da Plataforma Lattes e, em decorrência, auxiliar na compreensão das causas dessas possíveis inconsistências, subsidiando, assim, providências para aumentar a confiabilidade desse importante recurso.

1.2 OBJETIVO

O objetivo dessa dissertação é propor uma abordagem destinada à emissão de um mapa de inconsistências da Plataforma Lattes, como forma de despertar a comunidade de pesquisadores para a necessidade de se aperfeiçoar técnicas de medição e avaliação desse instrumento quanto a sua consistência e integridade. Cabe ressaltar que uma sistemática de acompanhamento do nível de consistência da Plataforma Lattes torna-se progressivamente importante, tendo em vista não só a sua crescente expressividade, como também o seu objetivo primordial que é subsidiar o processo decisório das organizações usuárias do Sistema Nacional de Ciência, Tecnologia e Inovação por meio da integração dos currículos de pesquisadores, grupos de pesquisa e de instituições.

A abordagem praticada nesse trabalho tem por premissa buscar responder a seguinte questão:

As técnicas e métodos adotados no processo de desambiguação e deduplicação em bibliotecas digitais, guardadas as especificidades, são efetivos na verificação de inconsistências na Plataforma Lattes?

1.3 METODOLOGIA

Visando atingir o objetivo proposto, a metodologia adotada, assim como todos os experimentos desenvolvidos, foi norteada pelos seguintes fundamentos:

- Especificação de uma heurística com base em um levantamento de hipótese.
- Aplicação da heurística especificada combinando-a com um tratamento de similaridade.

- Avaliação da abordagem proposta em um conjunto de dados reais, originados da Plataforma Lattes.

Inicialmente foi realizado um levantamento bibliográfico com o intuito de identificar as principais sistemáticas e características técnicas mais frequentemente utilizadas no processo de deduplicação e desambiguação.

A seguir, com o objetivo de praticar uma abordagem indutiva à questão em estudo, foi idealizado e executado um levantamento de hipótese amparado pela análise visual do segmento "Artigos Publicados" dos currículos Lattes de professores pertencentes ao Instituto de Computação (IC) da UFF. Dessa forma, algumas hipóteses de inconsistência puderam ser verificadas e subsidiaram a especificação de requisitos para o desenvolvimento de procedimentos automatizados visando o intencionado mapeamento.

Em um momento subsequente foi projetada e povoada uma base de dados contendo currículos Lattes de professores do IC/UFF e de alguns coautores. A seguir, foram desenvolvidos dois experimentos implementando o procedimento de verificação de inconsistências, respectivamente, para cada segmento tratado; dando origem a um mapeamento detalhado de inconsistências por currículo Lattes, o qual é complementado por um mapa sumarizando as inconsistências encontradas.

Em seguida foram desenvolvidos mais dois experimentos destinados a quantificar as inconsistências por unidade e instituição detentoras de um conjunto de currículos Lattes. Esses novos experimentos emitem um mapeamento sumarizado de inconsistências para cada unidade e um sumário total da instituição considerada, para cada um dos segmentos considerados.

Como forma de avaliar a efetividade dos experimentos, as medidas de precisão (*precision*) e cobertura (*recall*) foram adotadas por meio da realização de um teste de sensibilidade abrangendo os currículos dos professores do IC/UFF. Este teste indicou os limites ideais de similaridade, os quais foram aplicados nas execuções finais dos experimentos.

Finalmente a base de dados foi ampliada visando armazenar os currículos Lattes de todos os professores da UFF, totalizando 3805 CL's, quando então houve a execução final dos experimentos originando os resultados finais.

1.4 CONTRIBUIÇÕES

Essa dissertação almeja contribuir diretamente para o ambiente acadêmico propondo uma abordagem destinada à detecção de inconsistências na Plataforma Lattes, por meio da emissão de um mapa de inconsistências. Essa abordagem está amparada em dois pilares, a saber:

- Desenvolvimento e aplicação de uma heurística, fortemente influenciada pelos métodos de tratamento de ambiguidades e duplicidades em bibliotecas digitais, destinada à verificação de inconsistências referenciais entre CL's, especificamente nos segmentos "Artigos Publicados" e "Trabalhos em Eventos".
- Tratamento de similaridade visando identificar não só currículos de coautores, como também citações correlatas em currículos de coautores.

Entretanto, a busca por uma sistemática de aferição da integridade das informações existentes na Plataforma Lattes, mais especificamente nos currículos Lattes possui uma dimensão maior do que o conhecimento expressado nesse trabalho. Portanto, a principal contribuição dessa dissertação é a perspectiva desse trabalho evidenciar a fragilidade em que se encontra esse importante recurso de medição da produção científica nacional e, consequentemente, fazer emergir soluções visando o seu aperfeiçoamento.

1.5 ORGANIZAÇÃO

Após esse capítulo introdutório, esta dissertação está organizada em seis capítulos. O Capítulo 2 aborda a fundamentação teórica e trabalhos relacionados ao tema, discorrendo sobre as principais sistemáticas, metodologias e conceitos relacionados ao processo de desambiguação e deduplicação, além da estratégia de tratamento de similaridade amparada na edição de distância. A contextualização apresentada nesse capítulo é fundamental para caracterizar a importância desse estudo e, sobretudo, para possibilitar a conexão entre as práticas aplicadas em bibliotecas digitais e a abordagem proposta nesse trabalho.

O Capítulo 3 detalha a estratégia adotada para viabilizar o mapeamento de inconsistências, tendo como ponto de partida um levantamento de hipótese e uma fase analítica complementar, a qual viabilizou a especificação da heurística. Esse capítulo é finalizado com a apresentação dos mapas individual e sumarizado de inconsistências.

O Capítulo 4 detalha a infraestrutura computacional desenvolvida para amparar os experimentos realizados.

O Capítulo 5 analisa detalhadamente os resultados experimentais obtidos. Cada segmento abrangido pelo trabalho é analisado e, por fim, os resultados globais da instituição são discutidos. Complementando esses resultados, é apresentada uma projeção em função dos valores de precisão e cobertura obtidos em um teste de sensibilidade e, como forma de avaliar a efetividade desses resultados, o capítulo estabelece a conexão entre os resultados experimentais obtidos e a questão enunciada, a qual dá causa a essa dissertação. São apresentadas também algumas limitações da abordagem proposta.

O Capítulo 6 conclui a dissertação tecendo considerações sobre o estudo realizado e sugerindo trabalhos futuros nesse mesmo tema.

CAPÍTULO 2 - IDENTIFICAÇÃO DE RÉPLICAS E AMBIGUIDADES

2.1 INTRODUÇÃO

Conforme Singhal (2001), a necessidade de armazenar e recuperar informação tornou-se cada vez mais importante ao longo dos séculos e, com o surgimento dos computadores, os usuários perceberam que eles poderiam ser usados para armazenar e recuperar automaticamente grandes quantidades de informação. Retrospectivamente, Bush (1945) deu origem à ideia de acesso automático a uma grande quantidade de dados armazenados. Na década de 1950, essa ideia se materializou em descrições mais concretas sobre como informações textuais poderiam ser pesquisadas automaticamente. Vários trabalhos surgiram em meados de 1950 os quais abordavam a ideia básica da busca de texto em computador. Um dos métodos mais influentes foi descrito por HP Luhn em 1957, no qual ele propôs usar palavras como unidades de indexação de documentos e uma medida de sobreposição de palavras como critério de recuperação (LUHN, 1957).

A evolução dos métodos e técnicas de armazenamento de informações textuais em meio digital culminou com o surgimento das bibliotecas digitais. Essas bibliotecas são sistemas complexos, envolvendo um conjunto de componentes, tais como: coleções de informações, um sistema de computação oferecendo um conjunto de funcionalidades de acesso a essas coleções, usuários e um ambiente para o qual o sistema foi projetado (FUHR *et al.*, 2007). Um dos fatores que contribuem para a complexidade de sistemas dessa natureza é a possibilidade da existência de ambiguidades e duplicidades na representação dos dados, tendo em vista o fato dessas coleções de informações procederem de diversas origens, as quais podem adotar diversos padrões de armazenamento. Portanto, ambiguidades e duplicidades causam efeitos indesejáveis, não obstante serem inerentes a sistemas dessa natureza. Além do mais, de acordo com Newcombe *et al.* (1985), constituem um problema antigo e de solução não trivial.

Diante da complexidade crescente na obtenção de informações por parte das bibliotecas digitais, alguns esforços foram realizados no sentido de simplificar esse processo. Segundo Oliveira (2005) a *Open Archives Initiative* (OAI)[2] propõe padrões visando a interoperabilidade na disseminação de informações em bibliotecas digitais. Essa abordagem é

[2] http://www.openarchives.org

baseada em uma coleta periódica de dados originados de diferentes fontes por meio de um protocolo denominado *Open Archives Initiative Protocol for Metadata Harvesting*[3].

De forma mais ampla, a partir desses esforços e da evolução dos métodos de aquisição de informações, emerge e se intensifica o problema causado pela ambiguidade e duplicidade. Esse fato evidencia, conforme mencionado anteriormente, a ausência de um padrão único de representação dos dados, não obstante os esforços realizados por projetos como o *Scientific Eletronic Library Online*[4] (SciELO) e a Plataforma Lattes. Essa ausência de padronização torna possível que dados a respeito de um mesmo objeto tenham representações diferentes, dependendo da fonte da informação considerada. Por outro lado, é também possível que dados referentes a objetos diferentes tenham uma mesma representação. Essa é, essencialmente, a situação que caracteriza os fenômenos da duplicidade e ambiguidade. O efeito desse fenômeno reduz a qualidade dos serviços prestados pelas bibliotecas digitais e sistemas correlatos, na medida em que fragmentam o resultado de consultas, disponibilizando um resultado incompleto ou, em alguns casos, gerando redundâncias nas informações fornecidas.

Os avanços da sociedade brasileira, principalmente os científicos e tecnológicos, geram uma demanda crescente pela produção de indicadores quantitativos em ciência, tecnologia e inovação no país por parte dos governos federal, estadual e também pela própria comunidade científica nacional. Esses indicadores são, portanto, relevantes para subsidiarem a definição de diretrizes, alocação de investimentos e recursos, formulação de programas e também para a avaliação do desenvolvimento do país sob o ponto de vista científico e tecnológico (MUGNAINI; JANNUZZI; QUONIAM, 2004). Consequentemente, tornam-se relevante desambiguar e deduplicar as fontes de informações utilizadas na geração desses indicadores, como forma de mantê-los íntegros e consistentes. Segundo os mesmos autores, esses indicadores normalmente são gerados a partir de uma base multidisciplinar específica de cada país, tal como a Plataforma Lattes. É conveniente destacar que parte significativa do tempo despendido na transformação dos dados bibliográficos em indicadores bibliométricos é dedicada ao reconhecimento da forma com que os dados estão estruturados e no processo de transformação com o objetivo de possibilitar a apuração desses citados indicadores.

Nos dias atuais é crescente a complexidade na obtenção de informações e a evolução social possui uma correlação direta com essa complexidade. Nesse sentido, e para viabilizar o processo de tomada de decisões, Dos Santos (2004) afirma que a informação é um recurso

[3] http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[4] http://www.scielo.br/

gerencial imprescindível às instituições, percorrendo-as em todos os sentidos, ultrapassando inclusive suas fronteiras. Este mesmo autor propõe um agrupamento de informações com base em definições de informações mais presentes em debates sobre políticas de informações, quais sejam: informação como recurso, informação como mercadoria, informação como uma percepção de padrões e informação como uma força constitutiva da sociedade. O grupo que desperta a atenção nessa dissertação é o último e, conforme a Figura 1, somente informações consideradas como uma força constitutiva na sociedade incorporam todos os interesses representativos de fenômenos que ocorrem em todos os níveis da estrutura social. Nesse contexto, é importante ressaltar, inserem-se os indicadores sociais, inclusive os bibliométricos.

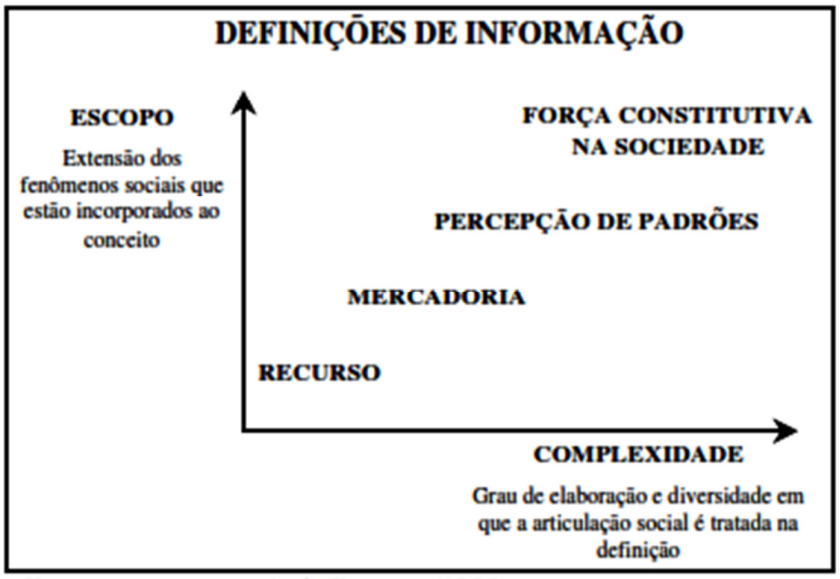

Figura 1 - Hierarquia das definições de informação – Fonte: (DOS SANTOS, 2004)

2.2 O PROCESSO DE DEDUPLICAÇÃO

O processo de deduplicação é intensamente executado no âmbito das bibliotecas digitais e sistemas correlatos, sendo considerado um passo fundamental na atividade de integração de dados provenientes de fontes diversas e envolve, essencialmente, a identificação e eliminação de duplicidades de registros relativos a uma mesma entidade ou objeto do mundo real, tendo em vista o fato dessas informações estarem submetidas a padrões de representação e armazenamento diferentes (CARVALHO *et al.*, 2006; BORGES; CARVALHO; *et al.*, 2011; FERREIRA; GONÇALVES; LAENDER, 2012).

Ao longo dos últimos anos, várias estratégias de deduplicação foram adotadas e testadas, resultando em diversos artigos acadêmicos. Observa-se uma quase unanimidade em

relação ao fato do processo automatizado de deduplicação recorrer às sistemáticas de aprendizado de máquina, heurísticas e também a funções de similaridade para dar o tratamento adequado ao processo, de forma a se obter o resultado mais efetivo possível.

Sarawagi e Bhamidipaty (2002) afirmam que o principal desafio do processo de deduplicação é encontrar uma função capaz de distinguir quando dois registros se referem à mesma entidade, a despeito de possíveis erros e inconsistências nos dados. Dessa forma, a sistemática apresentada pelos autores é implementada por um sistema de deduplicação baseado em aprendizado de máquina denominado de ALIAS[5], o qual, utilizando métodos de aprendizado ativo, possibilita a construção de uma função de deduplicação baseada em um método interativo de desafios de descobertas a partir de pares de registros de treinamento. A ideia central da abordagem é a criação simultânea de várias funções de similaridade redundantes e a exploração de divergências entre elas visando descobrir novos tipos de inconsistências entre duplicidades em um banco de dados de interesse.

Num momento seguinte, Carvalho *et al.* (2006) propõem uma abordagem que apresenta um novo processo de aprendizado de máquina combinando evidências fornecidas por dados de bibliotecas digitais para, por meio de programação genética, inferir uma função de similaridade no nível de registro, capaz de afirmar se dois registros são réplicas ou não. Para tanto essa função deriva de uma combinação de funções de similaridade aplicadas no nível de campo, ponderadas por pesos. Dessa forma, o registro com o maior nível global de similaridade é considerado o mais similar em relação ao registro sob análise. Um fato que diferencia essa abordagem de outras que utilizam o aprendizado de máquina é o fato dos métodos tradicionais requererem a realização de uma fase de treinamento para cada tarefa de deduplicação, ao passo que o método proposto por Carvalho *et al.* tem por objetivo o aprendizado de uma função de similaridade que pode ser desenvolvida a partir de vários conjuntos de dados distintos, de um mesmo domínio. Isso evita a necessidade da realização de um treinamento para cada tarefa de deduplicação.

Borges e Carvalho *et al.* (2011) apresentam uma abordagem heurística não supervisionada, baseada na deduplicação de metadados bibliográficos, a qual dedica uma especial atenção aos campos que se referem a nomes de autores visando identificar corretamente redundâncias nesses registros de metadados. Para tanto, o processo parte de um mapeamento entre os campos de metadados representados em diferentes padrões, sendo o foco principal dessa sistemática a construção de funções de similaridade especialmente

[5] Active Learning Led Interactive Alias Suppression

desenvolvidas para o domínio de bibliotecas digitais. Os conteúdos dos metadados são comparados por meio de uma dessas determinadas funções escolhidas de acordo com o domínio de cada metadado. Essa solução especifica três funções, a saber:

- *IniSim*

  Identifica variações na representação do nome de um coautor considerando erros ortográficos, inversões, abreviações e omissões de nomes.

- *NameMatch*

  Compara conjuntos de nomes próprios, sendo especificamente aplicada para comparar nomes de autores associados a dois objetos digitais distintos.

- *MetadataMatch*

  Tenta encontrar dois registros de metadados visando determinar se eles correspondem a uma réplica ou não.

Um aspecto interessante nessa abordagem é o fato dela não requerer qualquer tipo de treinamento, diferentemente de várias estratégias baseadas em técnicas de aprendizado de máquina encontradas na literatura.

Em uma obra subsequente, Borges e Becker *et al.* (2011) informam que a atividade de deduplicação, no domínio de artigos científicos em bibliotecas digitais, é geralmente baseada na semântica de alguns metadados específicos e, dentre os mais utilizados, estão os que representam os autores e o título do objeto digital. Essa obra apresenta uma análise comparativa entre alguns algoritmos de classificação (Naíve Bayes, Ripper e C4,5) e discute uma abordagem que combina funções de similaridade e algoritmos de classificação para identificar duplicidades em registros de metadados bibliográficos, os quais foram introduzidos por Borges e Carvalho *et al.* (2011), supracitados.

Uma conceituação singular na literatura consultada sobre o processo de deduplicação é apresentada por Christen (2011). O autor considera deduplicação o processo de descoberta de registros referenciando uma mesma entidade, porém em um mesmo banco de dados. Tal processo, quando aplicado a vários bancos de dados, é denominado de "*record linkage*". Esse trabalho apresenta um amplo levantamento sobre técnicas de indexação de registros entre bancos de dados, considerando a estratégia de blocagem. Essa estratégia gera blocos de registros de tal forma que somente registros de um mesmo bloco são comparados no processo de deduplicação. Nesse contexto, torna-se fundamental uma boa escolha dos atributos a serem considerados como chave de blocagem. Duas são as constatações, a saber: a qualidade dos atributos influenciam a qualidade do resultado final e a distribuição de frequência dos valores

dos atributos também afetam o tamanho dos blocos gerados. Portanto, alguns pontos são muito importantes para a efetividade dessa sistemática, quais sejam: a escolha dos atributos a serem utilizados para a geração dos blocos, a variedade e valoração dos parâmetros a serem informados pelos usuários e a possível sensibilidade entre alguns desses parâmetros.

2.3 O PROCESSO DE DESAMBIGUAÇÃO

O processo de desambiguação de citações está intimamente relacionado às bibliotecas digitais e é um dos problemas centrais na atividade de integração e limpeza dos dados. Conforme citado anteriormente, existe atualmente uma proliferação de dados bibliográficos oriundos das mais diversas fontes e a integração entre eles faz-se necessária como forma de prestação de um serviço efetivo de disponibilização de informações. Para tanto é necessário identificar ambiguidades em citações bibliográficas.

As estratégias para desambiguar citações bibliográficas são correlatas às adotadas no processo de deduplicação; ou seja, são fortemente inspiradas na combinação de aprendizado de máquina, heurísticas e tratamento por similaridade.

Na abordagem apresentada por Yang *et al.* (2008) é proposta uma sistemática centrada em dois tipos de correlações entre citações. A primeira, denominada Correlação de Tópico (*Topic Correlation*), mede a similaridade entre tópicos de duas citações por meio de métricas implementadas por funções de similaridade, tais como a função distância co-seno (*Cosine Similarity Metric* - CSM), utilizada para estimar a similaridade entre dois vetores, onde cada vetor representa o atributo título do artigo; e a função sigmod modificada (*Modified Sigmod Function* - MSF) que se baseia na coocorrência de características em dois conjuntos de atributos correspondentes. A segunda correlação, denominada Correlação Web, baseia-se na premissa de que as citações de pesquisadores são geralmente listadas em suas publicações ou nas publicações de seus coautores. Logo, se duas citações ocorrem na mesma página Web, existe grande probabilidade delas pertencerem ao mesmo indivíduo. Portanto, a Correlação Web significa a frequência de coocorrência de duas citações em páginas Web.

Pereira *et al.* (2009) apresentam um método que trata tanto o problema da polissemia (vários autores com o mesmo nome) quanto a ocorrência de sinônimos (diferentes nomes para um mesmo autor), por meio de uma clusterização hierárquica baseada em informações obtidas na Web, tais como currículo vitae e publicações referentes às citações ambíguas. A estratégia básica é, uma vez obtidos os documentos, separá-los agrupando citações que ocorrem em um mesmo documento, independentemente da grafia dos autores no documento. Por outro lado, separam-se em grupos distintos autores cujas citações não aparecem juntas em nenhum dos

documentos, mesmo tendo a mesma grafia. Em um momento subsequente e considerando o fato de uma citação específica poder constar em vários documentos, é aplicado o método de clusterização hierárquica por meio da atribuição de pesos visando ponderar a importância de cada documento no processo de desambiguação.

Mais recentemente Ferreira, Gonçalves e Laender (2012) propuseram uma taxonomia hierárquica, conforme a Figura 2, agrupando os métodos mais representativos de desambiguação de nome de autor encontrados na literatura.

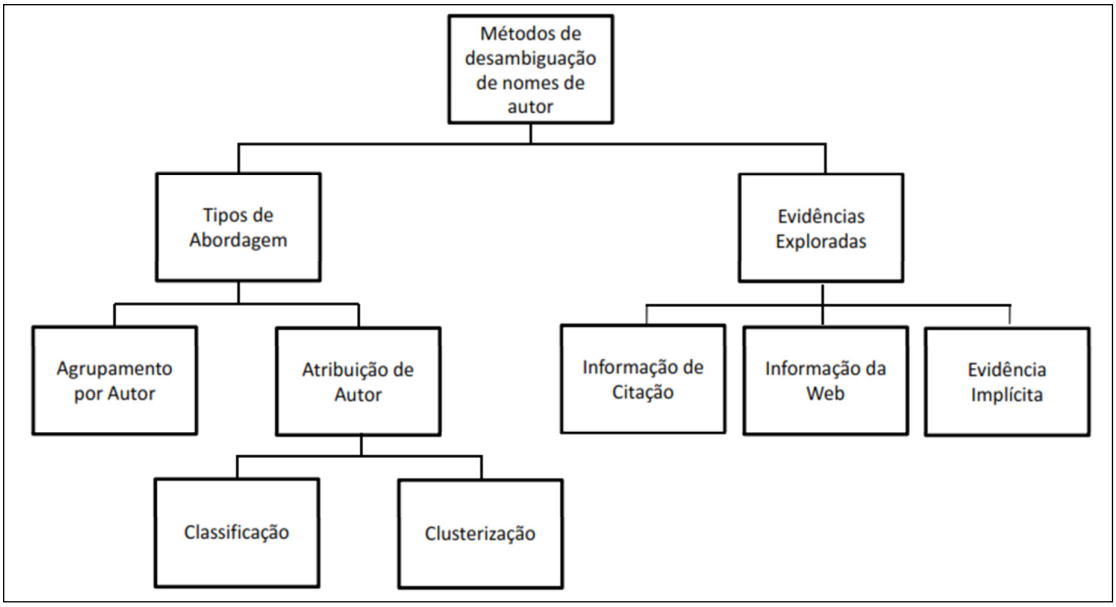

Figura 2 - Taxonomia proposta - Fonte: (FERREIRA; GONÇALVES; LAENDER, 2012)

O agrupamento por autor reúne processos que exploram sistemáticas de desambiguação baseadas no agrupamento de referências relativas a um mesmo autor por meio da adoção de técnicas de similaridade entre alguns atributos de referência. A atribuição de autor reúne métodos que atribuem diretamente as referências a seus respectivos autores. De forma alternativa, os métodos podem ser agrupados de acordo com as evidências exploradas no processo de desambiguação, quais sejam: informação de citação, que considera a extração de atributos diretamente obtidos das citações; informação da web, que reúne métodos que obtêm dados recuperados da web, utilizados como informações adicionais sobre o perfil de publicação do autor e, por último, evidência implícita que reúne métodos de inferência a partir de elementos disponíveis. Essas inferências buscam estimar uma distribuição que é utilizada como evidência para o cálculo da similaridade entre citações.

Não obstante os esforços e dinâmicas adotadas visando à implementação de processos automatizados para a desambiguação de citações, cabe referenciar Elliot (2010) que apresenta iniciativas consideradas manuais com esse objetivo. De fato, o problema da desambiguação adquiriu conotação multidisciplinar abrangendo áreas do conhecimento tais como Biblioteconomia, Ciência da Informação, além da Ciência da Computação. Os esforços para desambiguar citações vêm demandando não só esforços locais, específicos para um determinado contexto, como também esforços envolvendo sistemas nacionais e até internacionais. Em reconhecimento aos problemas decorrentes da desambiguação, a Biblioteca do Congresso Americano[6] (*Library of Congress*) instituiu uma sistemática voltada para manter um arquivo contendo registros de nomes de autores (*Library of Congress Authority File – LCAF*), manualmente gerado como parte do Programa de Cooperação de Nomes de Autores[7] (*Name Authority Cooperative Program - NACO*). Existem aproximadamente 400 (quatrocentas) instituições participantes no mundo, sendo o esforço de atualização compartilhado pelos diversos participantes. Entretanto, apesar da utilidade no fornecimento de informações sobre autoria de publicações de livros monográficos, essa estratégia não foi capaz de abranger nomes de autores de artigos individuais. Outra iniciativa, ainda referenciada pelo mesmo autor, visando à desambiguação manual, é encontrada no Projeto Nome da Biblioteca Britânica (*British Library's Names Project*) que surgiu em resposta ao crescente número de repositórios institucionais no Reino Unido e a consequente necessidade de controle de autoria de artigos.

2.4 A SIMILARIDADE EM CADEIAS DE CARACTERES

A noção de similaridade é utilizada nas mais diversas áreas do conhecimento acadêmico. De fato, Dutra (2008) enuncia a possibilidade de se considerar que tanto a ciência, enquanto um tipo específico de conhecimento, quanto o conhecimento humano em geral são atividades de construção de modelo, onde esse conceito é entendido como uma réplica da realidade. Nesse sentido a noção de similaridade possui um papel operativo e até mesmo indispensável. É inquestionável a existência de similaridade entre o modelo representado e a realidade. Além do mais, identifica-se uma similaridade mais sofisticada, a qual o autor denomina de "similaridade de estrutura". Essa similaridade está relacionada com a noção de congruência. É intuitivo supor que um modelo, por ser uma representação em escala de um

[6] http://www.loc.gov/

[7] http://id.loc.gov/download/

objeto real, não seja congruente com este. Entretanto, se ampliarmos progressivamente o modelo e reduzirmos da mesma forma o objeto representado, chegaremos a um momento de congruência. Logo, um modelo deve manter as mesmas proporções entre as partes que o compõem. Consequentemente, o conceito de similaridade possui uma existência intrínseca à construção de modelos, como é o caso nessa dissertação, não só na construção do modelo de dados detalhado na Seção 4.2, como também na aplicação de cálculos de similaridade considerados necessários para o mapeamento proposto.

A aplicação do conceito de similaridade no ambiente científico é vasta e necessária como forma de afirmar a igualdade de objetos a partir de seus elementos constitutivos considerados semelhantes. Nesse contexto, recursos algorítmicos para o cálculo de similaridade são objetos de pesquisa visando satisfazer necessidades das mais diversas áreas da pesquisa acadêmica.

Navarro (2001) apresenta um conjunto de técnicas destinadas a comparar *strings* com o objetivo de encontrar *strings* iguais, onde um ou os dois *strings* tenham sido corrompidos por alguma razão. Essas técnicas são úteis em diversas áreas do conhecimento, por exemplo, na biologia computacional, no processamento de sinais e na recuperação de textos. Essa obra aborda algoritmos baseados no conceito de distância de edição e na programação dinâmica. Considerando dois *strings*, a distância de edição significa o menor esforço para tornar esses dois *strings* iguais por meio de operações de exclusão, inserção e substituição de caracteres em um ou nos dois *strings*. Algumas variações desses algoritmos foram projetadas para permitirem somente inserção e exclusão visando atender a situações particulares tais como a necessidade de se identificar a maior sequência comum entre dois *strings* (*longest commom subsequence* - LCS). Outra variação da distância de edição permitindo somente substituição e que tem recebido bastante atenção é a *Hamming distance* que informa o número de posições com conteúdos diferentes em dois *strings*. Em outras palavras, significa o total de substituições necessárias para tornar os dois *strings* iguais.

O procedimento de identificação da maior sequência comum entre dois *strings* também é discutido por Ullman, Aho e Hirschberg (1976), com o objetivo de atender a aplicações genéticas tais como estudos da evolução molecular. Complementarmente Chen, Wan e Liu (2006) apresentam um breve histórico sobre esse algoritmo e abordam o problema de encontrar a maior sequência comum sob o ponto de vista de desempenho.

Além dessas abordagens, é relevante considerar a aplicação da noção de similaridade em ambiente de banco de dados ou coleções de valores. Nesse sentido é oportuno referenciar Dorneles *et al.* (2004) em relação a proposta apresentada para gerar métricas de similaridade

para manipular conjuntos de valores ocorrendo em documentos XML. Esse trabalho reconhece como recorrente o fato dos usuários nem sempre estarem habilitados a formalizar com precisão os argumentos de uma consulta a um banco de dados pelo fato deles conhecerem vagamente os dados disponíveis para consulta. Essa abordagem introduz duas métricas de similaridade, a saber.

Métrica para tratar valores atômicos (MAV): utilizada para tratar um elemento XML atômico. Seria aplicável, por exemplo, a um elemento “nome”, que seja filho de outro elemento XML denominado “pessoa”.

Métrica para tratar valores complexos (MCV): utilizada para tratar um elemento XML complexo. Seria aplicável, por exemplo, ao elemento “conferência” sendo que esse elemento possui os elementos “nome”, “ano” e “endereço” como elementos filhos. Essa métrica faz uma distinção entre elementos complexos, considerando-os como uma tupla ou como uma coleção. Como exemplo de uma coleção pode-se citar um elemento “publicação” que possua o elemento complexo denominado “autor”, o qual possui como filho várias ocorrências do elemento “nome”. Dessa forma são definidas duas classes de elementos complexos: tuplas e coleção.

2.5 CONSIDERAÇÕES FINAIS

Este capítulo apresentou as principais abordagens aplicadas aos processos de deduplicação e desambiguação. Foram também apresentadas algumas abordagens de tratamento de similaridade relevantes para o presente estudo. Essas sistemáticas e abordagens exercem influência marcante na solução proposta para o mapeamento de que trata esse trabalho. Portanto, em função das evidências verificadas, esse mapeamento intencionado está fortemente consubstanciado pela adoção de uma heurística, especificamente desenvolvida, combinada com um tratamento de similaridade, conforme detalhamento apresentado nos capítulos subsequentes.

CAPÍTULO 3 - O MAPEAMENTO DE INCONSISTÊNCIAS

3.1 INTRODUÇÃO

O capítulo anterior apresentou um levantamento bibliográfico sobre as principais sistemáticas com o objetivo de anular ou minimizar os efeitos nocivos causados pela duplicidade de dados em bibliotecas digitais e ambiguidade em citações de autoria. Essas sistemáticas influenciaram determinantemente a solução desenvolvida nessa dissertação.

Este capítulo detalha os passos percorridos para o desenvolvimento do mapeamento de inconsistências em currículos Lattes. O mapeamento identifica inconformidades referenciais entre CL's de coautores, no que diz respeito aos segmentos "Artigos Publicados" e "Trabalhos em Eventos".

A busca pela abordagem mais efetiva e a constatação de que as diversas áreas do conhecimento científico se tangenciam e, não raro, se interseccionam evidenciando um ambiente integrado, contribuiu para que o método indutivo fosse adotado para o problema em questão. Nesse sentido, é pertinente referenciar Michalski (1983) em sua afirmação sobre o aprendizado indutivo ser um processo de aquisição de conhecimento por meio do qual são realizadas inferências a partir de fatos fornecidos pelo ambiente e envolver operações de generalização, transformação, correção e refinamento de representações desse conhecimento desejado.

Tendo em vista o objetivo proposto, o restante deste capítulo está organizado da seguinte forma. A Seção 3.2 apresenta um levantamento de hipótese de inconsistência. A Seção 3.3 apresenta uma análise investigativa complementar. A Seção 3.4 detalha a heurística desenvolvida, enquanto a Seção 3.5 aborda o tratamento de similaridade adotado. As Seções 3.6 e 3.7 discorrem sobre os mapas de inconsistências individualizadas e sumarizadas, respectivamente e a Seção 3.8 apresenta considerações finais, encerrando do capítulo.

3.2 LEVANTAMENTO DE HIPÓTESE

Inicialmente foi realizada uma comparação visual do segmento "Artigos Publicados" abrangendo 58 CL's de professores do IC/UFF, totalizando 679 artigos verificados. Dessa forma, algumas hipóteses de inconsistências puderam ser verificadas e, assim, subsidiaram a especificação de um procedimento automatizado tendo em vista o intencionado mapeamento abrangendo os segmentos de interesse.

Essa comparação buscou identificar, com base empírica, as seguintes hipóteses de

inconsistências, preliminarmente imaginadas:

Erro de referência do artigo publicado: O artigo publicado constante em um determinado currículo Lattes sendo analisado, não é encontrado no currículo Lattes de um dos coautores.

Exemplo: Em um determinado currículo Lattes em análise consta o artigo publicado "*Using graph cuts in GPUs for color based human skin segmentation"* e no currículo Lattes de um coautor consta "*Colour based human skin segmentation using graph-cuts in GPUs*"

Erro de referência na citação de coautoria: O nome de um determinado coautor, constante em um determinado artigo publicado, está diferente do nome para citações bibliográficas informado no currículo do coautor.

Exemplo: Em um determinado artigo consta o coautor "Couve, J." e no currículo do coautor consta "COUVE, J. C.".

Todos os artigos publicados constantes em cada um dos currículos considerados tiveram suas existências verificadas visualmente nos currículos dos respectivos coautores. Nesse momento a relação de coautoria foi também comparada.

A tabulação desse procedimento preliminar pode ser observada na Tabela 2, onde fica destacada a maior incidência da inconsistência do tipo "Erro de Referência na Citação de Coautoria". Esse fato sugere que a aplicação de uma função de similaridade, combinada com uma heurística, poderá apresentar um resultado bastante efetivo na identificação das inconformidades buscadas pelo mapeamento. Entretanto, não obstante essa convicção formada, essas hipóteses de inconsistência foram consideradas insuficientes para um mapeamento mais expressivo.

Tabela 2- Resultado do Mapeamento Visual das Hipóteses de Inconsistência

Erro de Referência do Artigo Publicado	**Erro de Referência na Citação de Coautoria**
46	103

3.3 ANÁLISE INVESTIGATIVA COMPLEMENTAR

A imersão no processo investigativo e analítico complementando o levantamento de hipótese demandou a especificação e o desenvolvimento de uma infraestrutura computacional, detalhada no Capítulo 4, como forma de apoiar e possibilitar a obtenção de cruzamento de informações praticamente impossíveis de serem obtidas por outros meios, devido ao volume

de dados envolvido e a variedade de atributos a serem verificados. É oportuno ressaltar que todo esse processo investigativo complementar foi norteado por Chater (1997), onde é enunciado que a busca pela simplicidade deve ser o objetivo fundamental do processo cognitivo e os indivíduos tendem a buscar, naturalmente, as explicações mais simples para os fenômenos observados, e assim foi feito.

A infraestrutura desenvolvida possibilitou a análise de uma base de dados contendo os CL's de todos os professores do IC/UFF com seus respectivos artigos publicados e trabalhos em eventos. Foram inseridos também dados de CL's referentes a alguns coautores com o intuito de ampliar a base de comparação e facilitar a avaliação do comportamento dos futuros experimentos. Dessa forma, uma primeira versão da base de dados abrangeu 107 CL's, com as características de volume apresentadas pela Tabela 3.

Tabela 3 - Características de volume

Curriculum Vitae	Nome Citação	Artigo Publicado	Autor Artigo	Trabalho Evento	Autor Trab Evento
107	334	1.976	7.158	7.096	24.673

Após algumas consultas foi possível conhecer as características de domínio de alguns atributos persistidos e então definir quais deles seriam relevantes não só para a verificação de inconsistências, como também para subsidiar a especificação de uma heurística capaz de obter a maior efetividade possível na identificação das inconsistências buscadas e, assim, compor a especificação de requisitos para o primeiro experimento. A Tabela 4 apresenta os percentuais de tuplas com os respectivos atributos nulos ou considerados inválidos, referentes ao segmento "Artigos Publicados", enquanto a Tabela 5 apresenta os mesmos percentuais relativos ao segmento "Trabalhos em Eventos".

Tabela 4 - Totais de registros com respectivo atributo nulo (Artigos Publicados)

Nr Volume	Ano Publicação	Pg Inicial	Pg Final	Pgs Inicial e Final	Nr DOI	Nr ISSN
73 (3,69%)	0	63 (3,19%)	173 (8,75%)	63 (3,19%)	1.122 (56,78%)	189 (9,56%)

Tabela 5 - Totais de registros com respectivo atributo nulo (Trabalhos em Eventos)

Nr Volume	Ano Realização	Pg Inicial	Pg Final	Pgs Inicial e Final	Nr DOI	Nr ISBN
4.762 (67,11%)	0	2.638 (37,17%)	2.725 (38,40%)	2.634 (37,12%)	6.527 (91,98%)	5.577 (78,59%)

É relevante destacar a significativa quantidade de artigos publicados e trabalhos em eventos com os respectivos números de "DOI" nulos, totalizando 56,78% e 91,98%, respectivamente. Este fato desaconselha a utilização desse atributo como chave de acesso em quaisquer dos dois segmentos do CL. Merece também especial atenção o total de "Trabalhos em Eventos" com o atributo "ISBN" nulo. O atributo "Ano Publicação" por ser de preenchimento obrigatório na Plataforma Lattes constitui um bom atributo para ser utilizado como argumento de acesso. Sendo assim, uma vez localizado o título de um determinado artigo ou trabalho em evento, os seguintes atributos foram escolhidos para comporem os parâmetros de verificação: número do volume, ano da publicação, página inicial, página final, número do DOI, número do ISSN (para artigos publicados), número do ISBN (para trabalhos em eventos), nome do periódico (para artigos publicados), nome do evento (para trabalho em evento) e a ordem de autoria.

3.4 A HEURÍSTICA

Michaewicz e Fogel (2004) constatam que a sociedade se depara com problemas cada vez mais complexos. De fato, o mundo real está em constante mutação e a importância de uma solução eficaz para problemas nunca foi tão grande quanto na atualidade, pois a tecnologia disponível atualmente permite-nos afetar o ambiente de tal forma que os impactos de uma decisão podem acarretar consequências irreversíveis no futuro. Essa mesma tecnologia continua a se expandir e esse fato torna a solução de problemas cada vez mais complexa, na medida em que aumenta a quantidade de fatores a serem considerados. Logo, uma boa solução para um problema deve começar com uma adequada compreensão do propósito a ser atingido. Entretanto, decompor esse propósito subjetivo em objetivos mensuráveis de tal forma que seja possível verificar possíveis diferenças entre a abordagem e a realidade é uma atividade complexa, que requer prática e não há nenhum substituto para essa experiência. É aceitável, portanto, entender as soluções heurísticas como uma forma de solucionar problemas de forma satisfatória.

A palavra "heurística" admite muitas interpretações. Na sua origem grega significa simplesmente "encontrar ou descobrir". Porém, em um sentido mais contemporâneo, a heurística deve ser utilizada, sendo até mesmo indispensável ao processo cognitivo, para a solução de problemas que não podem ser tratados pela lógica e nem pela teoria da probabilidade. De forma consistente com essa afirmação, as soluções heurísticas podem ser vistas como uma abordagem para a solução de problemas e é necessariamente incompleta em relação ao conhecimento disponível sobre eles (GIGERENZER; ENGEL, 2006).

Além do mais a heurística, a lógica e a probabilidade são três ideias centrais na história intelectual da mente. A heurística, particularmente, possui a característica de ser frugal na medida em que ignora parte das informações. Uma heurística não tenta aperfeiçoar, no sentido de encontrar a melhor solução; mas sim satisfazer. Isto é, encontrar uma solução boa e suficiente para um determinado problema. Portanto, as heurísticas não tentam encontrar uma solução ótima, mas sim encontrar uma solução que satisfaça um determinado nível de satisfação desejado (GIGERENZER, 2008).

Diante dos fatos, os métodos heurísticos são muito úteis como um importante recurso para a descoberta de soluções para problemas, na medida em que permitem a obtenção de uma solução próxima a um determinado grau de satisfação.

Com base na literatura consultada e após a análise investigativa detalhada na Seção 3.3, foi possível definir a heurística apresentada abaixo. Vale esclarecer que a mesma heurística é aplicada aos dois segmentos do CL tratados nessa dissertação.

Início

Para cada Currículo Lattes a ser verificado;
Obter os limites aceitáveis de similaridade para os atributos passíveis de tratamento por similaridade [Título do artigo/trabalho, nome do periódico/evento e nome coautor];
Para cada Artigo Publicado/Trabalho em Evento
Obter coautores;
Para cada Coautor
Localizar o Currículo Lattes do coautor pelo atributo Nome;
Se Currículo Lattes do coautor não for localizado pelo atributo Nome
Então:
Localizar o Currículo Lattes do coautor pelo Identificador Lattes;
Se Currículo Lattes do coautor não for localizado pelo Identificador Lattes
Então:
Se o nome do Coautor contiver ponto ou vírgula
Então:
Localizar o Currículo Lattes do coautor por meio do Nome de Citação;
Senão:
Recuperar todos os Currículos Lattes cujos nomes dos titulares contenham as palavras que compõem o nome do coautor procurado;

Submeter os nomes dos titulares de cada currículo Lattes recuperado ao tratamento de similaridade

Apropriar o Currículo Lattes cujo nome do titular seja o mais similar, desde que igual ou superior ao limite desejado;

Fim se

Fim se

Fim se

Se Currículo Lattes do Coautor foi localizado

Então:

Localizar o Artigo Publicado/Trabalho em Evento no currículo do coautor, pelo atributo Título do Artigo/Trabalho em Evento;

Se o Artigo Publicado/ Trabalho em Evento foi localizado pelo Título do Artigo/Trabalho em Evento;

Então:

Comparar o atributo de verificação [nr volume, ano publicação, página inicial, página final, nr DOI, nr ISSN/ISBN (conforme o segmento do CL), nome periódico/evento (conforme o segmento do CL) e a ordem de autoria] informando as inconsistências;

Senão

Localizar o Artigo Publicado/Trabalho em Evento no currículo do coautor, em função dos atributos [Identificador do Currículo, ano publicação/evento, número do volume, páginas inicial e final];

Se encontrou Artigos Publicados/Trabalhos em Eventos

Então:

Apropriar o Artigo publicado/Trabalho em Evento cujo título seja o mais similar dentre os recuperados, desde que a similaridade seja igual ou superior ao limite informado;

Comparar a ordem de autoria, nr ISSN/ ISBN (conforme o segmento do CL), nome periódico/evento (conforme o segmento do CL) e nr DOI, informando possíveis inconsistências;

Senão:

Localizar o Artigo Publicado/Trabalho em Evento no currículo do coautor, em função dos atributos mais estáveis [Identificador do CL do coautor e ano publicação/evento];

Se encontrou Artigos Publicados/Trabalhos em Eventos

Então:

Apropriar o Artigo Publicado/Trabalho em Evento cujo título seja o mais similar dentre os recuperados, desde que a similaridade seja igual ou superior ao limite informado;

Comparar o atributo de verificação [nr volume, página inicial, página final, nr DOI, nr ISSN/ISBN (conforme o segmento do CL), nome

```
                                periódico/evento (conforme o segmento do CL) e a ordem de autoria]
                                informando as inconsistências;
                            Senão:
                                Informar "Inconsistência: Artigo Publicado/Trabalho em Evento não
                                localizado"
                            Fim se
                        Fim se
                    Fim se
                Senão:
                    Informar "Currículo Lattes não localizado. Verificação não realizada"
                Fim se
Fim
```

Após obter a identificação do CL e os limites toleráveis de similaridade para os atributos título do artigo/trabalho em evento, nome do periódico/evento e nome do coautor, o procedimento recupera o CL a ser verificado e os "Artigos Publicados" ou "Trabalhos em Eventos" referenciados, doravante denominados de item. A seguir, o procedimento tenta recuperar, para cada item, a correspondente referência no currículo de cada um dos coautores. Caso encontre, os atributos de verificação são comparados e as possíveis inconsistências identificadas são apresentadas.

Os currículos dos coautores são recuperados da base de dados, inicialmente pelo atributo "nome". Se não forem localizados, o procedimento tenta recuperá-los pelo atributo identificador Lattes (muitas vezes nulo). Na hipótese de fracasso nessa segunda tentativa e se o nome do coautor contiver um ponto e/ou um ponto e vírgula (o que caracteriza um nome de citação), o procedimento tentará localizar o CL do coautor pelo atributo "nome de citação". Se ainda assim não houver sucesso, uma última tentativa é realizada e o procedimento recupera todos os currículos cujos nomes dos titulares contenham pelo menos uma das palavras que componha o nome do coautor citado. Vale destacar que esse procedimento é implementado por meio da aplicação da cláusula "*like*" (recurso da linguagem SQL). Os registros recuperados, nesse caso, são então submetidos ao tratamento de similaridade e o nome do coautor mais similar com similaridade igual ou superior ao limite informado é submetido à verificação. Caso contrário, é emitida uma mensagem informando a impossibilidade de proceder à verificação, motivada pela não localização do currículo do coautor.

Eventualmente um item pode não ser localizado no currículo de um coautor por apresentar acentuada diferença de grafia, conforme detectado na Seção 3.2. Nesse caso o

procedimento faz uma tentativa de localizá-lo considerando os atributos de verificação obtidos no item referenciado pelo currículo em análise, quais sejam: identificador do currículo do coautor, ano de publicação do artigo/trabalho em evento, número do volume e páginas inicial e final. Se houver recuperação de itens, então o procedimento verifica a similaridade entre o título original do artigo/trabalho em evento e o título de cada uma das instâncias recuperadas e considera como similar o título com o maior percentual de similaridade, desde que a similaridade seja igual ou superior à tolerância informada.

Ainda assim, o item pode não ser localizado. Nesse caso, uma última tentativa de localização é feita quando o procedimento tenta identificar o item considerando os atributos mais bem comportados (mais estáveis), quais sejam: o identificador do CL do coautor e o ano de publicação do artigo/trabalho em evento. Para cada item recuperado por esses argumentos de acesso é calculada a similaridade entre o título original e o respectivo título recuperado. O item com maior similaridade, mais uma vez, desde que com similaridade superior ou igual ao limite informado, terá seus atributos de verificação comparados, quando então as inconsistências identificadas serão informadas. Nessa fase é esperado um volume maior de recuperação de itens, por esse motivo essa opção é a última a ser aplicada.

O procedimento apresenta o resultado de cada verificação realizada no segmento do CL considerado e finaliza a execução com a emissão de um mapa sumarizado de inconsistências identificadas, o qual apresenta totalizações referentes às comparações realizadas em cada um dos atributos de verificação. Esse mapa é detalhado na Seção 3.7.

3.5 O TRATAMENTO DE SIMILARIDADE

A necessidade de aplicação do conceito de similaridade neste presente trabalho insere-se num contexto menos complexo do que as referências citadas na Seção 2.4 e está relacionada a três situações específicas, discriminadas abaixo:

- Necessidade de se localizar por similaridade o CL de um coautor a partir do nome do coautor obtido no CL em análise. De acordo com a heurística detalhada na Seção 3.4, o último recurso de localização do CL de um coautor ocorre por meio da recuperação dos CL's de todos os titulares cujos respectivos nomes possuam pelo menos uma das palavras que componha o nome do titular do CL em análise. Esse processo é viabilizado pela aplicação da cláusula "*like*" (recurso da linguagem SQL). Os nomes dos titulares desses CL's são então submetidos ao tratamento de similaridade, sendo apropriado o

mais similar, desde que a similaridade seja igual ou maior do que o limiar de tolerância informado.

- Necessidade de se localizar por similaridade o título do artigo publicado/trabalho em evento. Existem duas hipóteses de aplicação de similaridade. Na primeira, o tratamento de similaridade é aplicado a todos os títulos de publicações recuperados com os seguintes atributos iguais: identificador do CL do coautor, ano publicação/evento, número do volume, páginas inicial e final. A segunda hipótese submete ao tratamento de similaridade todos os títulos de publicações recuperados a partir dos atributos considerados mais estáveis, que são: identificador do CL do Coautor e ano publicação/evento. Em ambas as hipóteses o titulo mais similar será considerado, desde que a sua similaridade seja maior ou igual ao limiar de tolerância informado.
- Necessidade de se localizar por similaridade o nome do periódico/evento, como forma de compensar divergências. Na hipótese do nome do periódico/evento ser diferente desse mesmo atributo na publicação de origem, a heurística submete esses atributos ao tratamento de similaridade e, tal como nos tratamentos anteriores, considera os atributos iguais se eles possuírem similaridade maior ou igual ao limiar de tolerância informado.

O algoritmo adotado para a verificação de similaridade foi o "Longest Common Subsequence" (LCS). Este algoritmo é considerado como uma solução clássica e muito utilizada, conforme menção anterior, em várias áreas do conhecimento científico. Seu objetivo é, dada duas sequências de entrada, identificar a maior sequência comum de caracteres entre essas sequências fornecidas. Não é necessário que a maior sequência comum contenha elementos consecutivos, mas tão somente que seja composta por elementos que ocorram na mesma ordem de apresentação. Sua complexidade assintótica é O(n+m) onde "n" e "m" representam os tamanhos das sequências de entrada (CORMEN *et al.*, 2009).

Em um tratamento preliminar, as sequências de entrada são ajustadas para caracteres maiúsculos, sendo também retirados espaços extras de forma que sobre somente um espaço entre as palavras que compõem cada uma dessas sequências.

É relevante esclarecer que o percentual de similaridade entre as duas sequências é obtido dividindo-se o dobro do tamanho da maior sequência comum pela soma dos tamanhos das duas sequências de caracteres de entrada.

3.6 O MAPEAMENTO DE INCONSISTÊNCIAS INDIVIDUALIZADAS

O mapeamento de inconsistências individualizadas (MII) foi o primeiro experimento a ser desenvolvido e analisa um determinado CL, identificando e detalhando possíveis inconsistências em cada um dos itens verificados em cada um dos segmentos abrangidos. Seu desenvolvimento foi desdobrado em dois programas para o tratamento individualizado de cada um dos segmentos tratados e possui como argumento de entrada a identificação de um determinado CL e os percentuais limites de similaridade aceitáveis para cada um dos atributos passíveis de serem tratados dessa forma, conforme Seção 3.5. Para "Artigos Publicados" os seguintes atributos possuem tratamento de similaridade: nome do periódico, título do artigo e o nome de coautores. Para o segmento "Trabalhos em Eventos" os atributos são: título do trabalho, nome do evento e nome de coautores.

Obtidos os dados de entrada, o procedimento apresenta, além dos dados do titular do CL, o total de itens a serem analisados, conforme a Figura 3 e, a seguir, o resultado da verificação de cada um dos itens citados. Em conformidade com a heurística enunciada na Seção 3.4, é detalhado o resultado de cada comparação realizada em cada um dos coautores referenciados.

```
>>>> Conexão feita com Sucesso <<<<<
================================================================================================================
Identificação do currículo a ser verificado:1
Percentual Limite de Similaridade Aceitável para Título do Artigo... [0.00 a 100.00]? 60
Percentual Limite de Similaridade Aceitável para Título do Periódico [0.00 a 100.00]? 60
Percentual Limite de Similaridade Aceitável para o Nome de Coautores [0.00 a 100.00]? 65
=================================== MAPA DE INCONSISTÊNCIAS INDIVIDUALIZADAS ===================================
IDENTIFICADOR LATTES.........: 0060120644445370 <---->  NOME: Vanessa Braganholo Murta
TOTAL DE ARTIGOS LOCALIZADOS: 18
============================== INÍCIO DA VERIFICAÇÃO DE ARTIGOS PUBLICADOS ==================================
Titulo Artigo..: Rumo ao título de Doutor/Mestre
DOI Artigo.....: Não Informado
Periódico......: Revista de Informática Teórica e Aplicada (Impresso)
ISSN Periódico.: 01034308  => Ano Public: 2004 => Nr. Vol: 10 [Pag. Inicial: 99 <->  Pag. Final: 112]
Limites de Similaridade Considerados ==> Titulo Artigo: 0.6 ==> Titulo Periódico: 0.6 ==> Nome Coautor: 0.65
==============================>> VERIFICAÇÃO DE COAUTORIA - Total de Autores: 4 <<=============================

Coautor: Mirella Moura Moro --> Nome Citação:[MORO, M. M.] --> OrdemAutoria: 1 --> IdentCnpq: 6408321790990372
Currículo Coautor NÃO localizado pelo NOME.
Currículo Coautor NÃO localizado pelo Id. LATTES. Possível Limitação da Base de dados
[#RESULTADO VERIFICAÇÃO#]: Verificação não realizada.

Coautor: Vanessa Braganholo Murta --> Nome Citação:[BRAGANHOLO, V.] --> OrdemAutoria: 2 --> IdentCnpq:
>>>>>>>> TITULAR DO CURRÍCULO SOB VERIFICAÇÃO

Coautor: André Costi Nácul --> Nome Citação:[NÁCUL, A. C.] --> OrdemAutoria: 3 --> IdentCnpq: 8612167299434512
Currículo Coautor NÃO localizado pelo NOME.
Currículo Coautor NÃO localizado pelo Id. LATTES. Possível Limitação da Base de dados
[#RESULTADO VERIFICAÇÃO#]: Verificação não realizada.

Coautor: Miguel Fornari --> Nome Citação:[FORNARI, M.] --> OrdemAutoria: 4 --> IdentCnpq: 9396708253007179
Currículo Coautor NÃO localizado pelo NOME.
Currículo Coautor NÃO localizado pelo Id. LATTES. Possível Limitação da Base de dados
[#RESULTADO VERIFICAÇÃO#]: Verificação não realizada.

================================================================================================================
```

Figura 3 – Exemplo MII (Início do procedimento)

A Figura 4 apresenta, como exemplo, o segmento de um MII no qual é possível observar os dados de apresentação do artigo em análise e, em seguida, o resultado da verificação referencial em cada um dos coautores. Cabe destacar que o CL do coautor "Celso Carneiro Ribeiro" não foi localizado pelo nome, mas sim pelo identificador Lattes. Nesse mesmo exemplo é possível observar que a comparação da ordem de autoria ocorre por similaridade pelo fato de existir divergência entre o nome do titular do CL do coautor e a respectiva citação feita pelo titular do CL em análise.

```
================================================================================================
Titulo Artigo..: Developing SPMD Applications with Load Balancing
DOI Artigo.....: 10.1016/S0167-8191(03)00060-7
Periódico......: Parallel Computing
ISSN Periódico.: 01678191  => Ano Public: 2003 => Nr. Vol: 29 [Pag. Inicial: 743 <->  Pag. Final: 766]
Limites de Similaridade Considerados ==> Titulo Artigo: 0.6 ==> Título Periódico: 0.6 ==> Nome Coautor: 0.65
==============================>> VERIFICAÇÃO DE COAUTORIA - Total de Autores: 3 <<==============================

Coautor: Celso Carneiro Ribeiro --> Nome Citação:[RIBEIRO, C. C.] --> OrdemAutoria: 2 --> IdentCnpq: 3614186131432854
Currículo Coautor NÃO localizado pelo NOME.
Currículo Coautor LOCALIZADO pelo Id. Lattes
--> Artigo Localizado pelo Título
>>> Título do Periódico é igual
>>> Ano da Publicação é igual
>>> Número do Volume é igual
>>> Páginas [inicial e final] são iguais.
>>### ORDEM DE AUTORIA DIFERENTE - Avaliação por Similaridade
>>### Nome na Origem.: Celso Carneiro Ribeiro
>>### Nome no Coautor: Celso da Cruz Carneiro Ribeiro [Similaridade: 0,85]
----->> Computada igualdade. Similaridade igual ou acima do limite considerado.
>>> ISSN do Periódico é igual
### DOI Não Informado no currículo sob verificação e/ou no do Coautor
[#RESULTADO VERIFICAÇÃO#]: Inconsistências --> Coautor não Localizado pelo nome

Coautor: Noemi Rodriguez --> Nome Citação:[RODRIGUEZ, N.] --> OrdemAutoria: 3 --> IdentCnpq: 4933326132948063
Currículo Coautor NÃO localizado pelo NOME.
Currículo Coautor NÃO localizado pelo Id. LATTES. Possível Limitação da Base de dados
[#RESULTADO VERIFICAÇÃO#]: Verificação não realizada.
```

Figura 4 – Exemplo MII (Tratamento de um item)

Outro exemplo que ratifica a heurística é apresentado na Figura 5, onde se observa que o titular do CL em análise ao invés de informar o nome do coautor "Isabel Cristina Mello Rosseti", informou o nome de citação "Rosseti, Isabel". A heurística aplicada percebe esse fato e, como o identificador Lattes é nulo, recupera o CL do coautor a partir desse nome de citação informado.

Ao término da verificação é apresentado um mapa contendo a sumarização das verificações realizadas. Este mapa está detalhado na próxima seção.

```
=============================================================================================================
Titulo Artigo..: A hybrid data mining GRASP with path-relinking
DOI Artigo.....: 10.1016/j.cor.2012.02.022
Periódico......: Computers & Operations Research
ISSN Periódico.: 03050548  => Ano Public: 2013 => Nr. Vol: 40 [Pag. Inicial: 3159 <->  Pag. Final: 3173]
Limites de Similaridade Considerados ==> Titulo Artigo: 0.6 ==> Título Periódico: 0.6 ==> Nome Coautor: 0.65
==============================>> VERIFICAÇÃO DE COAUTORIA - Total de Autores: 4 <<==============================

Coautor: Hugo Barbalho --> Nome Citação:[BARBALHO, Hugo] --> OrdemAutoria: 1 --> IdentCnpq:
Currículo Coautor NÃO localizado pelo NOME.
Id Lattes é nulo. Tentativa de localização do Currículo do Coautor pelo nome de citação e cláusula Like.
------->>> Tentativa sem Sucesso
[#RESULTADO VERIFICAÇÃO#]: Verificação não realizada.

Coautor: Rosseti, Isabel --> Nome Citação:[Rosseti, Isabel] --> OrdemAutoria: 2 --> IdentCnpq:
Currículo Coautor NÃO localizado pelo NOME.
Id Lattes é nulo. Tentativa de localização do Currículo do Coautor pelo nome de citação e cláusula Like.
--->>>Currículo Lattes Localizado pelo Nome de Citação.
------>> Identif. Curriculo: 4  <<-->> Nome do Titular: Isabel Cristina Mello Rosseti
--> Artigo Localizado pelo Título
>>> Título do Periódico é igual
>>> Ano da Publicação é igual
>>> Número do Volume é igual
>>> Páginas [inicial e final] são iguais.
-->>### Nome do Coautor aparenta ser um nome de citação
-->>### ORDEM DE AUTORIA DIFERENTE
>>> ISSN do Periódico é igual
### DOI Não Informado no currículo sob verificação e/ou no do Coautor
[#RESULTADO VERIFICAÇÃO#]: Inconsistências --> Coautor não Localizado pelo nome
```

Figura 5 – Exemplo MII (Apropriação de coautor)

3.7 O MAPEAMENTO DE INCONSISTÊNCIAS SUMARIZADAS

O Mapa de inconsistências sumarizadas (MIS) deriva do experimento MII e tem por objetivo quantificar as verificações realizadas fornecendo subsídios quantitativos para um processo analítico. Ele pode ser emitido abrangendo tanto um determinado CL quanto uma determinada unidade ou instituição. O MIS, apresentado na Figura 6, está estruturado em três blocos de informações, a saber:

- Sumarização referente a itens: esse bloco apresenta totais relativos aos itens verificados. Os totais informados discriminam os itens conforme o recurso da heurística utilizado na localização do item. É relevante destacar que, eventualmente, a soma entre os totais de localizações por nome e por similaridade, pode resultar num valor maior do que o total de artigos, representado pela letra "B". Isto se deve ao fato de um mesmo artigo poder ser localizado várias vezes, inclusive de forma diferentes para cada um dos seus coautores.
- Sumarização referente a coautores: esse bloco apresenta totais relativos aos coautores. São informados totais em função de cada método da heurística utilizado para localizar os coautores. Quais sejam: localização pelo nome do coautor, localização pelo identificador Lattes, localização pelo nome de citação e localização pela cláusula "*like*" combinada com tratamento de similaridade. É oportuno ressaltar que um mesmo coautor pode ser contabilizado várias vezes pelo

fato de ser referenciado em vários itens. O total de verificações realizadas abrange todas as comparações feitas buscando identificar as inconsistências consideradas.

- Sumarização referente a inconsistências: esse bloco discrimina os totais apurados referentes a cada tipo de inconsistência considerada.

Conforme citado, a sumarização referente às inconsistências é gerada em função de um CL, unidade ou instituição. A taxa de inconsistência, expressada pela letra "Z" no MIS, permite avaliar o volume de inconsistências em relação ao total de verificações realizadas. Por fim, alguns dados adicionais são apresentados, quantificando as localizações de coautores para efeito de verificação da ordem de autoria.

```
================================================================================================================

>>>>>>>>>>>> Mapa de Inconsistências Sumarizadas por Unidade - ARTIGOS PUBLICADOS <<<<<<<<<<<< Em: 18/8/2014
Instituicao: 55 - UNIVERSIDADE FEDERAL FLUMINENSE  --> Unidade: 9 - INSTITUTO DE COMPUTAÇÃO
                  Limites de Similaridade Considerados
    Título Artigo: 0.6 <-> Título Periódico: 0.6 <-> Nome Coautor: 0.65
[A]>>Total de Currículos na Unidade (Analisados).....................: 29
[B]>>>>Total de Artigos Publicados na Unidade.......................: 381
[C]    Total Localizações de Artigos pelo Nome .....................: 162
[D]    Total Localizações de Artigos por Similaridade ..............: 36
[E]    Total de Localizações e VERIFICAÇÕES [C+D]...................: 198
[F]    TAXA DE VERIFICAÇÕES EM ARTIGOS PUBLICADOS [([E]*100)/[B]].....: 51,97
[G]>>Total de Coautores na Unidade (ACEITA possíveis duplicidades)...: 1128
[H]    Total de Coautores Localizados pelo Nome.....................: 113
[I]    Total de Coautores Localizados pelo Id Lattes [Nome Diferente].....: 81
[J]    Total de Coautores Localizados pelo Nome de Citação.............: 12
[K]    Total de Coautores Localizados pela cláusula LIKE com Similaridade.: 108
[L]    Total de Coautores Localizados e Verificados [H+I+J+K]..........: 314
[M]    TAXA DE VERIFICAÇÕES EM COAUTORES {[L]*100)}/[G].................: 27,84
[N]    Total de Verificações Realizadas ..............................: 4910
------------------------ INCONSISTÊNCIAS ENCONTRADAS -------------------------
[O]     Artigos não encontrados em coautores:..............: 116
[P]     Título do Periódico não localizado:................: 4
[Q]     Ordem de Coautoria diferente:......................: 52
[R]     Ano da Publicação diferente: ......................: 9
[S]     Volume diferente:..................................: 22
[T]     Páginas [Inicial ou final]diferentes:..............: 47
[U]     DOI diferente:.....................................: 3
[V]     ISSN diferente:....................................: 17
[X]     TOTAL DE INCONSISTÊNCIAS...........................: 270
[Z]     TAXA DE INCONSISTÊNCIA DA UNIDADE ([X]*100/N).....: 5,5
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
DADOS ADICIONAIS: Ordem de Autoria Apropriada pelo Nome .............: 85
                  Ordem de Autoria Apropriada por Similaridade......: 27
                  Ordem de Autoria Apropriada pelo Nome de Citação..: 28
================================================================================================================
```

Figura 6 – Exemplo MIS (Mapa sumarizado)

3.8 CONSIDERAÇÕES FINAIS

Este capítulo detalhou a abordagem desenvolvida visando o mapeamento de inconsistências tendo como objeto a Plataforma Lattes. Conforme apresentado, o mapeamento está fortemente amparado na combinação de um método heurístico com um tratamento de similaridade.

Execuções sucessivas desses experimentos evidenciaram a necessidade de se determinar os limites de similaridade ideais, para então aplicá-los na base de currículo de professores da UFF, composta por 3805 currículos, de forma a obter a maior efetividade possível por meio da redução tanto dos falsos positivos quanto dos falsos negativos. Para tanto foi desenvolvido um teste de sensibilidade, o qual está detalhado no Capítulo 5.

CAPÍTULO 4 - A INFRAESTRUTURA COMPUTACIONAL

4.1 INTRODUÇÃO

Este capítulo tem por objetivo apresentar o ambiente computacional desenvolvido para viabilizar a execução dos experimentos detalhados no capítulo anterior. Para tanto, a especificação dessa infraestrutura de suporte levou em consideração não só a possibilidade de subsidiar futuras aplicações relacionadas ao tema dessa dissertação, como também a flexibilidade como requisito não funcional com vistas a facilitar futuras modificações estruturais que se fizerem necessárias com o objetivo de auxiliar a realização de outros possíveis trabalhos acadêmicos.

De fato, o esforço despendido no desenvolvimento de infraestruturas para apoiar a execução de experimentos científicos não deve ser desperdiçado. Portanto, é desejável a prática de reutilização de infraestruturas como forma de aumentar a efetividade da atividade de pesquisa. Essa última consideração está em sintonia com Sayão e Sales (2012) no que diz respeito à importância que vem sendo atribuída ao grande volume de dados derivados de pesquisa. Esses autores introduzem o conceito de curadoria de dados digitais de pesquisa com a finalidade de preservar esses dados, tendo em vista o fato das pesquisas científicas atualmente se depararem com um paradigma no qual a ciência unifica de forma crescente experimentos, teorias e simulações por meio do uso intenso de dados obtidos por instrumentos cada vez mais sofisticados ou gerados por simulação, armazenados em bases de dados e disponíveis para serem integrados e analisados pelo ambiente acadêmico.

Visando atender a finalidade enunciada foi implementado um banco de dados para armazenar os dados a serem analisados nesta dissertação. Esse banco de dados ficará disponível para pesquisas futuras que venham a ser conduzidas por outros alunos. O capítulo foi estruturado da seguinte forma. A Seção 4.2 descreve o modelo de dados especificado para representar os conceitos relativos aos currículos Lattes abrangidos pelo trabalho. A Seção 4.3 detalha o procedimento desenvolvido para povoar o banco de dados e, por fim, a Seção 4.4 tece considerações finais sobre o capítulo.

4.2 O MODELO DE DADOS

O modelo proposto foi derivado da *Document Type Definition* (DTD) do currículo Lattes, disponibilizada pelo CNPQ (2012b). Inicialmente, e com o objetivo específico de subsidiar a análise investigativa citada na Seção 3.3 do capítulo anterior, o modelo de dados

representou somente as entidades: *CurriculumVitae*, *NomeCitacao*, *ArtigoPublicado*, *TrabalhoEvento*, *AutorArtigo* e *AutorTrabEvento*. Em um momento subsequente, objetivando a futura execução do mapeamento de inconsistências com maior amplitude e flexibilidade, esse modelo foi ampliado por meio da incorporação dos conceitos de "*Instituicao*" e "*Unidade*", ambos definidos a seguir.

Apesar da similaridade entre as estruturas dos segmentos tratados por esse trabalho, quais sejam "Artigos Publicados" e "Trabalhos em Eventos" e também com o objetivo de levar flexibilidade à representação dos conceitos de interesse, conforme supracitado, optou-se por manter representações distintas para esses dois segmentos, os quais podem ser considerados gêneros de um conceito mais abrangente "Produção Acadêmica", por exemplo. A Figura 7 apresenta a versão final do modelo de dados.

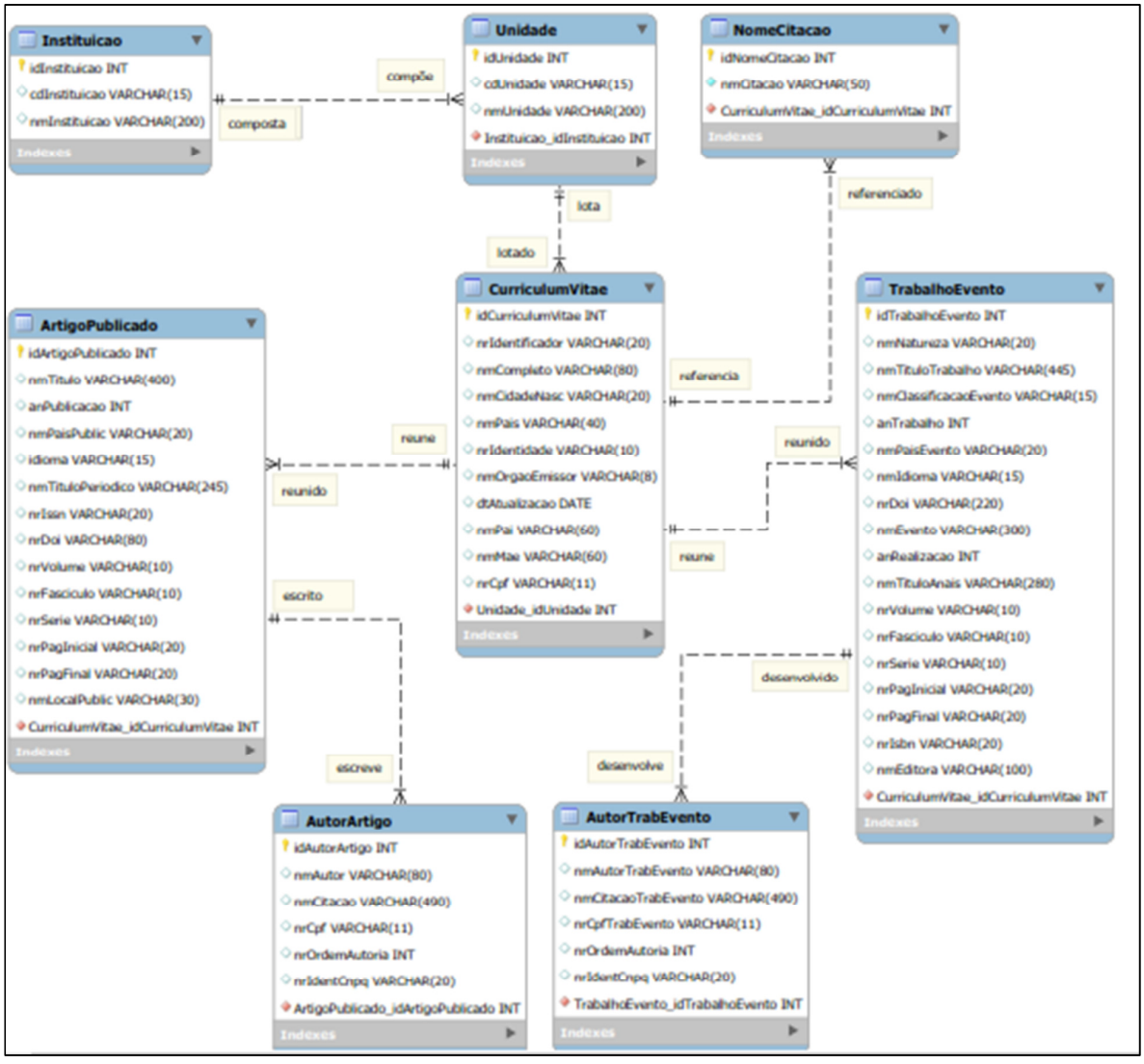

Figura 7 – Modelo de Dados

Com o objetivo de contextualização é apresentada a seguir a descrição de cada uma das entidades que compõem o modelo de dados. Essas descrições são complementadas pelo ANEXO A, que dicionariza os atributos de cada entidade e, complementarmente, informa a correspondência de cada um desses atributos em relação à DTD do arquivo XML de entrada. Dessa forma, é possível identificar a origem dos dados de cada um dos atributos modelados.

- **Instituição** - Representa todas as organizações de ensino, pesquisa ou fomento que sejam do interesse da Plataforma Lattes, tais como: Universidades Federais, Universidades Estaduais, Institutos de Pesquisa, etc.
- **Unidade** - Representa todas as repartições que compõem uma determinada "*Instituição*". Normalmente este conceito está associado à estrutura organizacional da "Instituição". Por exemplo: Departamento de Geologia, Instituto de Computação, etc.
- **CurriculumVitae** - Representa os currículos de todos os profissionais que mantenham registro na Plataforma Lattes, tais como: pesquisadores, professores, alunos de doutorado, alunos de mestrado e de graduação.
- **ArtigoPublicado** - Representa todos os artigos publicados pelo titular de um determinado currículo, em coautoria ou não.
- **TrabalhoEvento** - Representa todos os trabalhos publicados em eventos, no qual o titular do currículo seja autor ou coautor.
- **AutorArtigo** - Representa os coautores de um determinado artigo publicado.
- **AutorTrabEvento** - Representa os coautores de um determinado trabalho em evento.

O modelo de dados representa o fato de que uma determinada "Instituição" é composta por "Unidades". Estas, por sua vez, abrigam um conjunto de "CurriculumVitae". Um determinado "CurriculumVitae" reúne (possivelmente vários) "ArtigoPublicado" e também "TrabalhoEvento". Tanto "ArtigoPublicado" quanto "TrabalhoEvento" são descritos por "AutorArtigo" e "AutorTrabEvento", respectivamente. É oportuno informar que as características dos atributos que qualificam cada uma das entidades representadas foram definidas a partir de sucessivos testes visando determinar não só o formato, mas também o tamanho e a obrigatoriedade. Nesse processo foi constatada a existência de vários atributos com conteúdo incompatível com o seu significado. Estas inconformidades estão indicadas no Capítulo 5.

Para a implementação do modelo, foi utilizado o software *MySQL Workbench*[8]. Esse software é uma bancada integrada para desenvolvimento utilizando o SGBD MySQL. Trata-se de uma ferramenta que disponibiliza funcionalidades destinadas à administração e projeto de banco de dados, além de recursos para o desenvolvimento de modelos de dados, dentre outros.

É importante ressaltar que essa versão final do modelo de dados, tal como descrito, além de possuir os atributos necessários à realização do mapeamento proposto, pode também suportar uma ampla gama de experimentos voltados para analisar não só dados curriculares como também gerar indicadores em função das unidades detentoras de currículos e, numa amplitude maior, das instituições. Portanto, conforme mencionado anteriormente, o modelo desenvolvido tem potencial para ser reutilizado em futuros experimentos acadêmicos.

4.3 O PROCEDIMENTO DE POVOAMENTO DO BANCO DE DADOS

Os CL's utilizados no experimento detalhado na Seção 3.3 foram obtidos individualmente, de forma manual na própria Plataforma Lattes, no formato XML. Esses currículos foram submetidos a um fluxo de processamento conforme o diagrama da Figura 8. Num momento subsequente, quando as características dos dados já eram bem conhecidas, os CL's do corpo docente da UFF foram obtidos, no mesmo formato XML, junto a Superintendência de Tecnologia de Informação da UFF (STI/UFF) e submetidos ao mesmo tratamento, de forma a viabilizar os experimentos finais. É oportuno ressaltar que nesse momento houve a reestruturação do banco de dados com o objetivo de incorporar os conceitos de "Instituicao" e "Unidade", conforme citado na Seção 4.2.

O programa que gera scripts de inserção foi desenvolvido na linguagem Java e acessa os CL's no formato XML por meio da API SAX[9]. Seu desenvolvimento foi sucessivamente ajustado em função da identificação de novas características dos dados a serem inseridos. Vale destacar que, como forma de facilitar a identificação e o tratamento de possíveis interrupções no momento da carga no banco de dados, foi adotada a estratégia de gerar um script de inserção no banco para cada 100 (cem) currículos. Portanto, para povoar toda a base de dados, foram gerados e executados 38 scripts de inserção. Essa estratégia se revelou eficiente na medida em que permitiu a identificação de várias inconformidades entre os dados processados e o projeto físico do banco de dados, principalmente no que diz respeito ao

[8] http://www.mysql.com/products/workbench/

[9] http://www.saxproject.org/

tamanho dos atributos modelados.

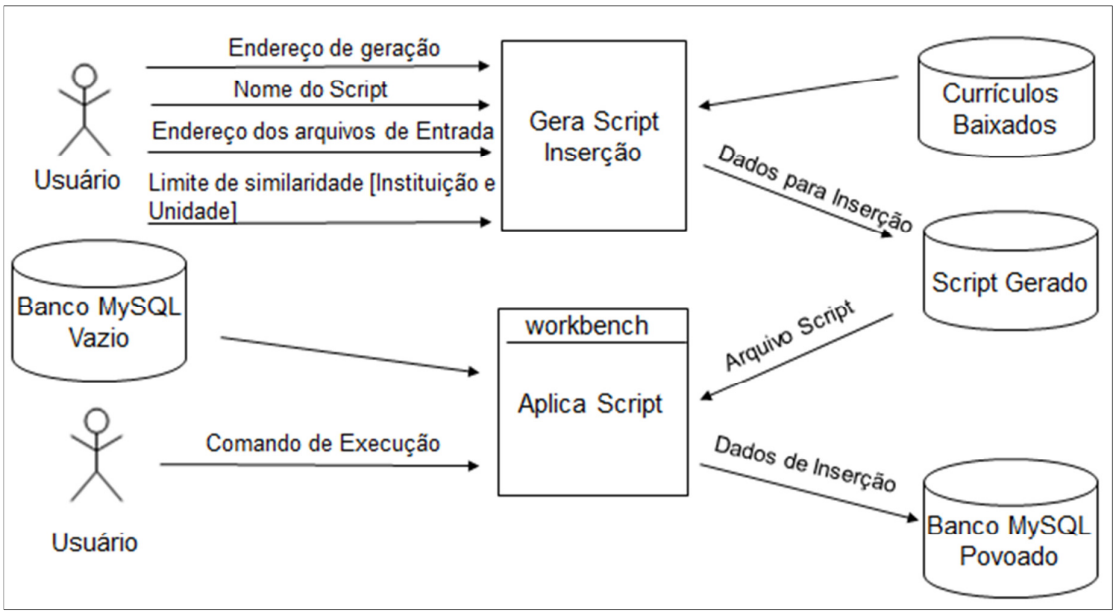

Figura 8 – Diagrama de Carga

A versão final do procedimento de geração do script de inserção possui o seguinte comportamento.

Início

Obter o endereço de geração do script;
Obter o nome do arquivo que conterá o script;
Obter o endereço de localização dos currículos de entrada no formato XML;
Obter o limite de similaridade aplicável para a Unidade e para a Instituição
Normalizar a codificação do arquivo de entrada para [UTF-8];
Abrir o primeiro arquivo de entrada;
Enquanto existir arquivo de entrada a ser processado;
Se Identificador Lattes não existir
Então:
Identificador Lattes ← nome do arquivo de entrada (é sempre o identificador);
Fim se
Transformar o atributo Data da Última Atualização para o formato "aaaammdd";
Desmembrar o atributo Nome Citação em função da existência do separador [;]. Gerar tantas instâncias quantos forem os nomes de citação;
Colocar cada um dos atributos não nulos entre aspas;
Se Instituição não existir nos dados de entrada
Então:
Instituição ← "*Instituição não informada no currículo relacionado*"
Gravar inserção na tabela [Instituicao];

```
Fim se
Se Instituição existir na base de dados
Então:
        Apropriar essa instância encontrada na base de dados
Senão:
        Se existir ocorrência de Instituição com similaridade >= tolerância informada para similaridade
        Então:
                Apropriar a Instituição encontrada por similaridade
        Senão:
                Apropriar a nova instância de Instituição
                Gravar inserção na tabela [Instituicao];
        Fim se
Fim se
Se Unidade não existir no elemento "CODIGO-UNIDADE" do arquivo de entrada E Unidade não
    existir no elemento "CODIGO-ORGAO" do arquivo de entrada
Então:
        Unidade ← "Unidade não informada no currículo relacionado"
        Gravar inserção na tabela [Unidade];
Fim se
Se Unidade existir na base de dados, relacionada à Instituição considerada
Então:
        Apropriar essa instância encontrada na base de dados
Senão:
        Se existir ocorrência de Unidade com similaridade >= tolerância informada para similaridade
        Então:
                Apropriar a Unidade encontrada por similaridade
        Senão:
                Apropriar a nova instância de Unidade
                Gravar inserção na tabela [Unidade];
        Fim se
Fim se
Gravar inserção na tabela [CurriculumVitae];
Gravar inserção na tabela [NomeCitacao];
Obter o primeiro Artigo Publicado;
Enquanto existir Artigo Publicado
    Normalizar espaço entre as palavras que formam os atributos [Nome do Periódico, Título do
            Artigo];
    Retirar possíveis espaços do atributo Número do Fascículo;
    Retirar possíveis espaços do atributo Número do DOI;
    Retirar possíveis aspas e/ou vírgulas dos atributos [Nome do Periódico, Título do Artigo];
```

```
            Colocar cada um dos atributos não nulos entre aspas;
            Gravar inserção na tabela [ArtigoPublicado];
            Obter Autor do Artigo Publicado;
            Enquanto existir Autor do Artigo Publicado
                  Colocar cada um dos atributos não nulos entre aspas;
                  Gravar inserção na tabela [AutorArtigo];
                  Obter o próximo Autor do Artigo Publicado;
            Fim Enquanto
            Obter o próximo Artigo Publicado;
      Fim Enquanto

      Obter o primeiro Trabalho em Evento;
      Enquanto existir Trabalho em Evento
            Normalizar espaço entre as palavras que formam os atributos [Título do Trabalho, Nome do
                  Evento, Título dos Anais];
            Retirar possíveis aspas e/ou vírgulas dos atributos [Título do Trabalho, Nome do Evento, Título
                  dos Anais];
            Retirar possíveis espaços do atributo Número do Fascículo;
            Retirar possíveis espaços do atributo Número do DOI;
            Se o total de caracteres do atributo Ano de Realização do Evento for diferente de 4
            Então:
                  Considerar o atributo Ano de Realização como nulo;
            Fim se
            Colocar cada um dos atributos não nulos entre aspas;
            Gravar inserção na tabela [TrabalhoEvento];
            Obter Autor do Trabalho em Evento;
            Enquanto existir Autor do Trabalho em Evento
                  Colocar cada um dos atributos não nulos entre aspas;
                  Gravar inserção na tabela [AutorTrabEvento];
                  Obter o próximo Autor do Trabalho em Evento;
            Fim Enquanto
            Obter o próximo Trabalho em Evento
      Fim Enquanto
      Obter o próximo arquivo de entrada a ser processado;
Fim Enquanto
```

Fim

Esse procedimento de geração do script de inserção inicia com a obtenção dos seguintes argumentos de entrada: endereço de geração do *script*, nome do arquivo que conterá o script, endereço dos arquivos XML contendo os CL's a serem processados e o limite de

similaridade aceitável tanto para a "Instituição" quanto para a "Unidade". A seguir os dados são normalizados para a codificação UTF-8[10] e o procedimento então percorre todos os arquivos XML, acessando-os no endereço informado. O procedimento assegura obrigatoriamente a identificação do CL por meio da apropriação do nome do arquivo como o identificador Lattes, pois foi observado que esse atributo não é informado em muitos currículos. É importante informar que os nomes de citação compõem um único atributo no arquivo de entrada. Esse fato exigiu um tratamento adicional para desmembrar todos os nomes de citação informados pelo titular do CL, de forma a manter a normalização do modelo de dados.

A seguir o procedimento acessa o elemento "ENDERECO-PROFISSIONAL" para apropriar tanto a instituição quanto a unidade. Caso a instituição não tenha sido informada, o procedimento considera a instituição inexistente gerando uma instância com o texto "Instituição não informada no currículo relacionado". Na hipótese da instituição ter sido informada, o procedimento verifica se essa instância existe cadastrada na base de dados. Se existir então essa instância é considerada para estabelecer a relação com a tabela de "Unidade". Caso contrário, é calculada a similaridade entre a instituição informada e as instituições existentes na base de dados. A ocorrência mais similar é apropriada, desde que a similaridade seja maior ou igual ao limite de similaridade informado no início do procedimento. Se o limite de similaridade não for satisfeito, então o procedimento considera a instância da instituição informada no CL como uma nova tupla a ser inserida na base de dados e efetiva a inserção. Logo após, o procedimento verifica se a unidade foi informada analisando primeiramente o atributo "CODIGO-UNIDADE", do mesmo elemento "ENDERECO-PROFISSIONAL". Caso este atributo esteja nulo, o atributo "CODIGO-ORGAO" será verificado. Esse comportamento foi adotado devido ao fato de a unidade ser informada tanto no primeiro atributo referenciado quanto no segundo. Se ainda assim não existir conteúdo, o procedimento assume a inexistência da unidade atribuindo o texto "Unidade não encontrada no currículo relacionado". Se a unidade tiver sido informada, então o procedimento verificará se essa instância existe cadastrada na base de dados, associada à instituição anteriormente tratada, considerando-a previamente cadastrada, caso exista. De outra forma, o procedimento verificará a unidade mais similar existente na base de dados e apropriará a mais similar, desde que a similaridade seja igual ou superior ao limite informado. Não existindo similaridade aceitável, o procedimento fará a inserção de uma nova unidade a

[10] http://www.unicode.org/

partir dos dados apropriados. É conveniente esclarecer que o comportamento tanto para a inserção da instituição quanto para a inserção da unidade tem por objetivo evitar a inserção de ocorrências em duplicidade nessas respectivas tabelas.

Todos os “Artigos Publicados” e “Trabalhos em Eventos” são então acessados e os atributos de interesse recebem um tratamento especificamente definido para garantir a uniformidade da representação e evitar possíveis rejeições devido a restrições impostas pelo banco de dados. É assegurado somente um espaço entre as palavras que compõem os atributos: título do artigo, título do trabalho, nome do periódico, nome do evento e título dos anais. Possíveis aspas, vírgulas e barras existentes também são retiradas desses atributos. Adicionalmente, são retirados quaisquer espaços existentes nos atributos número do fascículo e número do DOI.

4.4 CONSIDERAÇÕES FINAIS

Este capítulo detalhou a infraestrutura computacional desenvolvida para suportar o mapeamento de que trata essa dissertação. A especificação e o desenvolvimento dessa infraestrutura foram realizados levando em consideração a possibilidade de reutilização desse recurso como forma de facilitar a realização de futuros experimentos de natureza análoga à desse trabalho. Atualmente existe uma proliferação crescente de dados resultantes de experimentos acadêmicos e também de dados gerados para subsidiar pesquisas científicas. Esse fato desperta o interesse para uma reflexão, considerada oportuna, no sentido de se aperfeiçoar o reuso desses dados com o objetivo de reduzir o tempo despendido na preparação de futuros ambientes computacionais e, consequentemente, aumentar o tempo dedicado ao experimento científico propriamente dito.

CAPÍTULO 5 - RESULTADOS EXPERIMENTAIS

5.1 INTRODUÇÃO

Os capítulos anteriores detalharam o processo de desenvolvimento destinado à realização do mapeamento de que trata essa dissertação. Entretanto, a análise dos resultados experimentais obtidos é de fundamental importância para verificar a efetividade desses resultados como forma de responder a questão enunciada na Seção 1.2, qual seja: *As técnicas e métodos adotados no processo de desambiguação e deduplicação em bibliotecas digitais, guardadas as especificidades, são efetivos na verificação de inconsistências na Plataforma Lattes?*

A análise dos resultados obtidos em sucessivas execuções com variações tanto dos limites de similaridade quanto de volume da base de dados indicam um resultado bastante satisfatório. Contudo, esses sucessivos experimentos evidenciaram a necessidade de um amparo mais consistente visando identificar um limiar ideal de similaridade. Essa necessidade também está alinhada com o sentido de acentuar a abordagem sistemática e os níveis de controle requeridos por um trabalho acadêmico, tal como preconizado por Scalaton e Garcia (2014) quando afirmam que um experimento controlado é aquele em que acontece a investigação de uma hipótese supondo-se uma relação causal entre características de interesse do fenômeno em estudo. Essa relação causal deve ser mapeada em variáveis de forma a que se possam aferir os resultados obtidos. Além do mais, essas variáveis de controle devem ser criteriosamente manipuladas visando sempre à geração de resultados que possibilitem conclusões corretas. Seguindo esse preceito, foi desenvolvido um procedimento complementar visando avaliar a sensibilidade dos limiares de similaridade para os segmentos do CL abrangidos por esse trabalho. Em um momento subsequente o MIS foi gerado tendo como limites de similaridade os valores indicados pelo teste de sensibilidade.

Os resultados experimentais obtidos foram apurados em cada um dos segmentos abrangidos por essa dissertação e, num momento seguinte, foi feita a totalização desses resultados, originando a apuração global da Universidade Federal Fluminense.

O ambiente de hardware, sobre o qual os experimentos foram desenvolvidos, foi composto por um computador com um processador INTEL(R) CORE(TM) I7 4500-U, memória RAM de 8 GB, SO Windows 8.

O restante desse capítulo está estruturado da seguinte forma. A Seção 5.2 apresenta detalhes da fase de análise preliminar dos dados de entrada. A Seção 5.3 discute o teste de

sensibilidade realizado com a intenção de aumentar a efetividade dos experimentos finais. A Seção 5.4 discute detalhadamente os resultados finais obtidos, discriminando-os para cada um dos segmentos abrangidos e totalizando a instituição analisada. A Seção 5.5 apresenta possíveis ameaças à validade dos resultados obtidos, identificadas no decorrer do desenvolvimento dos experimentos e, por fim, a Seção 5.6 tece considerações finais sobre o capítulo.

5.2 ANÁLISE PRELIMINAR DOS DADOS

Conforme informado na Seção 4.3, a versão final da base de dados é composta pelos CL's de todos os docentes da UFF. Esses dados foram obtidos junto à Superintendência de Tecnologia de Informação (STI) da UFF. A Tabela 6 apresenta o volume de dados para cada uma das tabelas na versão final da base de dados.

Tabela 6 – Característica de volume das tabelas

Instituição	unidade	curriculumlattes	nome citacao	artigo publicado	trabalho evento	autor artigo	autortrab evento
294	797	3.805	7.158	50.574	95.425	184.034	328.492

Inicialmente esses novos currículos recebidos foram submetidos a uma análise preliminar com a intensão de se identificar possíveis conteúdos com características não previstas pela análise investigativa tratada na Seção 3.3. Foram verificados totais expressivos de conteúdo nulo em alguns atributos, conforme apresentado na Tabela 7 e Tabela 8. É importante frisar a grande quantidade de registros com o atributo "nrDoi" nulo nos dois segmentos abrangidos, quais sejam 78,27% em "Artigos Publicados" e 99,30% em "Trabalhos em Eventos". Foram também localizados 67 registros no segmento "Artigo Publicado" com o conteúdo do atributo "Ano Publicação" igual a 1900. Enquanto que no segmento "Trabalhos em Eventos" foram localizados 3.714 registros com o atributo "Ano Realização" também igual a 1900. Logo, é aceitável considerar essas ocorrências também como inconsistências devido à impossibilidade dos respectivos autores terem redigido essas obras nesse ano.

Tabela 7 – Totais de registros com respectivo atributo nulo (Artigos Publicados)

Nr Volume	Ano Publicação	Pg Inicial	Pg Final	Pgs Inicial e Final	Nr DOI	Nr ISSN
2.154 (4,26%)	0	1.845 (3,65%)	6.494 (12.84%)	1.842 (3,64%)	39.582 (78,27%)	5.415 (10,71%)

Tabela 8 – Totais de registros com respectivo atributo nulo (Trabalhos em Eventos)

Nr Volume	Ano Realização	Pg Inicial	Pg Final	Pgs Inicial e Final	Nr DOI	Nr ISBN
65.022 (68,14%)	0	56.217 (58,91%)	62.493 (65,49%)	56.066 (58,75%)	94.754 (99,30%)	81.849 (85,77%)

Foram também identificados alguns conteúdos considerados em desconformidade com o sentido semântico do respectivo atributo. A Tabela 9 e a Tabela 10 apresentam alguns exemplos dessas inconformidades nas colunas intituladas "Nr Página Inicial" do segmento "Artigo Publicado" e "Trabalhos em Eventos", respectivamente.

Tabela 9 - Exemplo de inconformidade em conteúdo (Artigo Publicado)

Título do Artigo	Nr Página Inicial	Nr Página Final
Effect of Plasma Toroidal Flows on Poloidal Ion Rotation and HC Condutivity in Edge Plasmas of Tokamaks	22222222222	
Estimating the Extent of Underreporting of Mortality among HIV-Infected Individuals in Rio de Janeiro, Brazil	101007190827070	
Laparoscopic Partial Nephrectomy Under Warm Ischemia Reduces the Glomerular Density in a Pig Model	120104064747007	710
Cranial Pole Nephrectomy in the Pig Model: Anatomic Analysis of Arterial Injuries in Tridimensional Endocasts	120221104407005	
Magnetic Mn and Co Complexes with a Large Polycyclic Aromatic Substituted Nitronylnitroxide	120222092334008	3145
Os arquitetos e o patrimônio	http://www.comc	
Um autor necessário, um romance de urgência: acerca de O AMOR É FODIDO, de Miguel Esteves Cardoso	http://www.crit	
Augusto Malta e Marc Ferrez: Olhares sobre a construção de uma metrópole	http://www.deze	
Restrictions on the Use Cadmium Coating in Industries	http://www.ejou	
SOBRE A ATIVIDADE DO PENSAMENTO E A BANALIDADE DO MAL EM HANNAH ARENDT	http://www.filo	

Tabela 10 – Exemplo de inconformidade em conteúdo (Trabalho em Evento)

Título do Trabalho em Evento	Nr Pagina Inicial	Nr Página Final
Alternative photoinitiators: stress reduction while maintaining DC and crosslink density.	Abstract 0133	
Operações Fundamentais: Ideias, Significados e Algorítmos	3608_2062_ID	
Finite Element Simulation of a 2-D Non-Linear Heat Transfer Problem using a Minimum Principle	1Eduardo D. Cor	
A Trajetória da Seguridade e da Assistência Social vista a Partir do Desenvolvimento do Benefício de Prestação Continuada (BPC)	nºtrabalho1063	
O papel das instituições de ensino superior junto à unidades de conservação	T13_0606_3259	
Remoção de Rodamina B de Efluentes Industriais Empregando Espumas de Poliuretano.	p. AB-059-20	
Mechanical and Thermal Properties of Graphene Nanomeshes	opl.201.186	
Avaliação dos fatores de risco para o desenvolvimento da cárie da primeira infância	Abstr. PB058	
Effect of the LASER preparation on the hybrid layer thickness	Abstract 1526	
Un Motor sin Rodamientos Controlado con DSP	paper 08-12.p1p	paper 08-12.p9

É muito importante ressaltar a quantidade significativa de currículos que não possuem a "Universidade Federal Fluminense" como instituição a que pertencem, mas sim diversas outras, conforme informado pela Tabela 11. Lembrando que a base de currículos fornecida pelo STI/UFF abrange todo o corpo docente da UFF e, portanto, todos esses currículos

deveriam mencionar a UFF como instituição. Adicionalmente, foi também notada a existência de várias unidades vinculadas a UFF com uma quantidade duvidosa de CL associado. Por exemplo, várias unidades possuem somente um CL vinculado. Alguns CL's também estão vinculados a unidades inexistentes.

Tabela 11 – Totais de Currículos Lattes

Instituições	
UFF	**Outras**
2.733 (71,83%)	1.072 (28,17%)

5.3 O TESTE DE SENSIBILIDADE

Conforme citado anteriormente, o teste de sensibilidade foi considerado necessário para possibilitar a identificação do limiar ideal de similaridade para o título do artigo/ evento e para o nome do coautor visando aumentar a efetividade do mapeamento de inconsistência. Esses dois atributos são considerados fundamentais para o mapeamento, pois o nome do coautor é o argumento central de busca para a localização do coautor por similaridade e o título do artigo/evento também é o principal argumento de consulta para a sua localização por similaridade no CL de um determinado coautor.

A busca por um limiar de similaridade que garanta a maior efetividade na execução dos experimentos implica na obtenção de uma métrica de precisão e de uma métrica de cobertura. Entretanto a necessidade de se combinar a maior precisão com a maior cobertura, visando à obtenção da maior efetividade, é traduzida pela média harmônica dessas duas medidas, a qual é denominada de *F-Measure*. Portanto, os limiares de similaridade mais efetivos para os atributos considerados serão indicados pelo maior *F-Measure* (HAN; KAMBER; PEI, 2012).

Pelo fato da geração do gabarito do experimento exigir uma atividade de conferência visual, para a realização do teste de sensibilidade optou-se por considerar um subconjunto dos dados contendo 43 CL's de professores credenciados na pós-graduação do IC/UFF. A Figura 9 apresenta o procedimento executado para a geração dos gabaritos para os dois segmentos tratados.

O procedimento de geração do mapa para a obtenção do gabarito foi derivado do MII e tem o objetivo de gerar um mapa contendo os registros referentes às tentativas de localização bem sucedidas (positivos) e outro mapa contendo os registros referentes às tentativas de localização mal sucedidas (negativos). Para esse procedimento foram adotados

limiares baixos de similaridade, a saber: 50% para o título do artigo/evento e 55% para o nome do coautor.

Esses mapas gerados para conferência detalham, para cada CL, os resultados encontrados em relação à localização de cada CL de coautor e também o respectivo título do artigo ou trabalho em evento, conforme o segmento em processamento. Dessa forma é possível, por meio de uma cuidadosa conferência visual, identificar a assertividade do processamento e, ainda por meio manual, corrigir os falsos positivos e falsos negativos encontrados. Cabe esclarecer que falsos positivos são aquelas localizações de publicações consideradas como falsas e os falsos negativos são aquelas tentativas de localização sem sucesso e que a análise visual identifica como falsas, ou seja, a publicação existe.

Para a geração do gabarito os falsos positivos transformam-se em verdadeiros negativos e os falsos negativos transformam-se em verdadeiros positivos. O resultado final dessa conferência gera os dados para povoar quatro tabelas de gabarito para os segmentos tratados. O gabarito é composto de duas tabelas para cada segmento do CL em estudo. Uma para as ocorrências positivas encontradas e outra para as negativas.

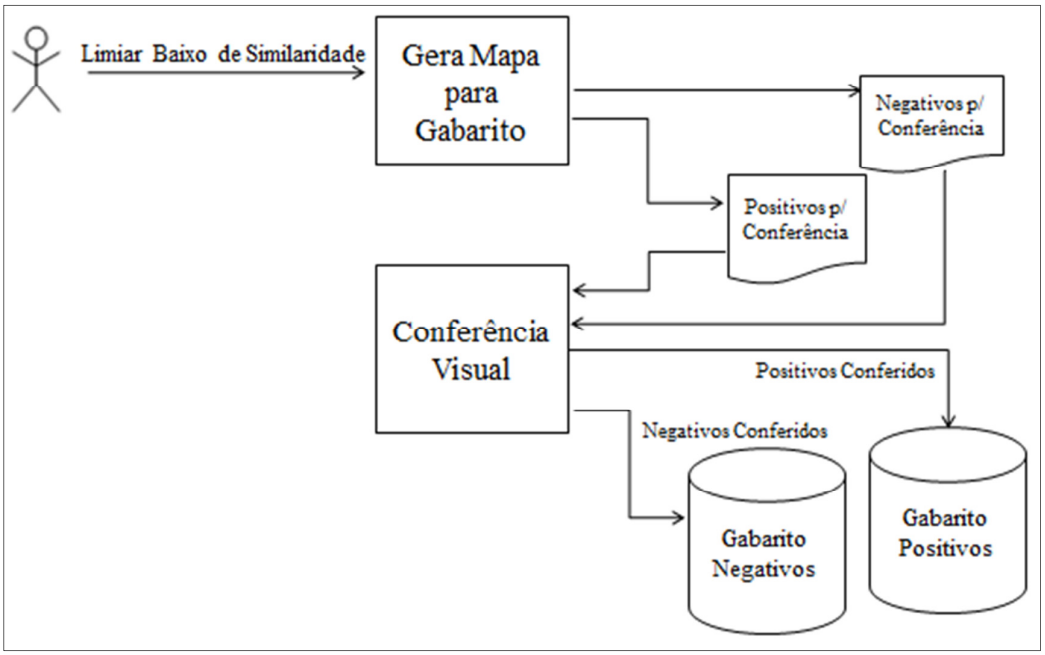

Figura 9 – Procedimento de geração do gabarito

O procedimento de geração da matriz de sensibilidade é executado conforme informado na Figura 10. Essa matriz reúne o resultado de várias execuções de uma versão do MII adaptada para comparar e quantificar cada uma das verificações realizadas em relação ao

gabarito, gerando os dados necessários para o cálculo das medidas de precisão e cobertura. A adaptação realizada no MII implementa o seguinte comportamento.

```
Início
    Obter dados de entrada
    Obter CL para verificação
    Enquanto existir CL para verificação
        Obter publicação
        Enquanto existir publicação
            Obter coautor
            Enquanto existir coautor, desde que diferente do titular do CL sendo verificado
                Se houve a localização do CL do coautor
                Então:
                    Se houve a localização da publicação
                    Então:
                        Se o Gabarito de positivos contiver a associação entre o CL sendo
                        verificado, o coautor e a publicação
                        Então:
                            Incrementar o contador de verdadeiro positivo
                        Senão:
                            Incrementar o contador de falso positivo
                        Fim se
                    Senão:
                        Se o Gabarito negativo contiver a associação entre o CL sendo
                        verificado, o coautor e a publicação
                        Então:
                            Incrementar o contador de verdadeiro negativo
                        Senão:
                            Incrementar o contador de falso negativo
                        Fim se
                    Fim se
                Senão:
                    Se o Gabarito negativo contiver a associação entre o CL sendo verificado, o
                    coautor e a publicação
                    Então:
                        Incrementar o contador de verdadeiro negativo
                    Senão:
                        Incrementar o contador de falso negativo
                    Fim se
                Fim se
```

Fim enquanto

Fim enquanto

Fim enquanto

Fim do procedimento

O procedimento inicia solicitando o limite inferior de similaridade para o título do artigo/evento e para o nome do coautor; a seguir é informado o valor de incremento a ser aplicado ao limite de similaridade, o total de iterações e o nome do arquivo de saída (contendo os dados da matriz).

A realização do teste iniciou com um limite inferior de similaridade no valor de 40%. Esse valor inicial foi incrementado do valor de 0.05% em iterações sucessivas e o último limiar de similaridade avaliado foi 90% tanto para o título do artigo/evento quanto para o nome do coautor.

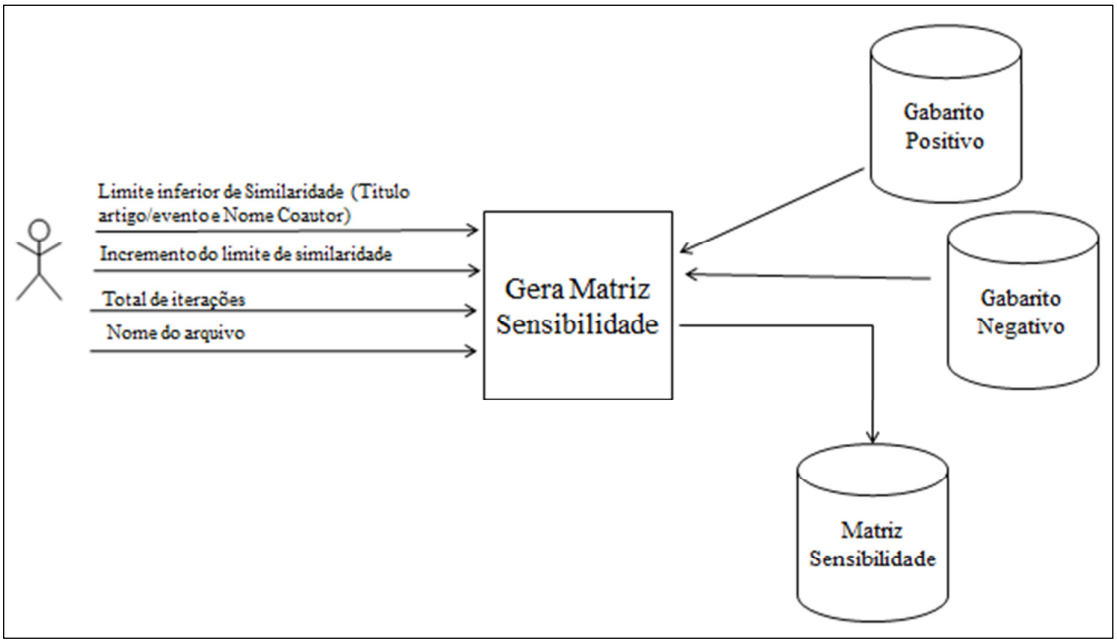

Figura 10 – Procedimento de geração da matriz de sensibilidade

Após a obtenção dos dados de entrada, o procedimento inicia o processo de mapeamento conforme o MII adaptado. Essa adaptação foi realizada para acessar as tabelas contendo os gabaritos de positivos e negativos, de tal forma que, para cada localização de um currículo de coautor em uma determinada publicação, o procedimento verifica no gabarito se o titular do CL, considerando a publicação em análise, existe associado ao coautor. Se existir então um contador de localização de verdadeiros positivos é incrementado. Caso contrário, um contador de falsos positivos é incrementado. Caso um determinado currículo de coautor,

em uma determinada publicação, não seja localizado, o procedimento busca identificar esses dados no gabarito de negativos. Se existir, então o contador de verdadeiros negativos é incrementado. Caso contrário, o contador de falsos negativos é incrementado.

A Tabela 12 apresenta os valores mais expressivos da matriz de sensibilidade gerada para o segmento "Artigos Publicados", a qual indica como os mais efetivos os limiares de similaridade de 55% para o título do artigo e 50% para o nome de coautor. Esses limiares produziram um F-Measure de 0,9548. Os valores mais expressivos da matriz de sensibilidade do segmento "Trabalhos em Eventos" são apresentados na Tabela 13, onde é possível verificar como os mais efetivos os limiares de 65% para o título do trabalho em evento e 55% para o nome do coautor. Esses percentuais produziram um F-Measure de 0,9130. Esses são os limiares de similaridade aplicados na execução do experimento final abrangendo toda a base de docentes da UFF. As matrizes de sensibilidade com todos os valores apurados são apresentadas no ANEXO B.

Tabela 12 – Matriz de Sensibilidade (Artigos Publicados)

Análise de sensibilidade para os limites informados de similaridade - ARTIGOS PUBLICADOS								
Similaridade Artigo	**Similaridade Coautor**	**Verdadeiro Positivo**	**Falso Positivo**	**Verdadeiro Negativo**	**Falso Negativo**	**Precisão**	**Recall**	**F1-Measure**
0.55	0.40	349	6	1.889	27	0,983099	0,928191	0,954856
0.55	0.45	349	6	1.889	27	0,983099	0,928191	0,954856
0.55	0.50	349	6	1.889	27	0,983099	0,928191	0,954856
0.55	0.55	348	6	1.890	27	0,983051	0,928000	0,954733
0.50	0.40	349	7	1.888	27	0,980337	0,928191	0,953552
0.50	0.45	349	7	1.888	27	0,980337	0,928191	0,953552
0.50	0.50	349	7	1.888	27	0,980337	0,928191	0,953552

Tabela 13 – Matriz de Sensibilidade (Trabalhos em Eventos)

Análise de sensibilidade para os limites informados de sensibilidade - TRABALHOS EM EVENTOS								
Similaridade Trabalho	**Similaridade Coautor**	**Verdadeiro Positivo**	**Falso Positivo**	**Verdadeiro Negativo**	**Falso Negativo**	**Precisão**	**Recall**	**F1-Measure**
0.65	0.55	1.192	96	6.784	131	0,925466	0,900983	0,913060
0.55	0.55	1.205	119	6.764	115	0,910121	0,912879	0,911498
0.60	0.55	1.194	108	6.773	128	0,917051	0,903177	0,910061
0.65	0.60	1.181	96	6.784	142	0,924824	0,892668	0,908462
0.50	0.50	1.219	148	6.738	98	0,891734	0,925588	0,908346
0.55	0.60	1.194	118	6.765	126	0,910061	0,904545	0,907295
0.50	0.45	1.220	154	6.731	98	0,887918	0,925645	0,906389

5.4 ANÁLISE DOS RESULTADOS OBTIDOS

Os resultados obtidos na análise preliminar, conforme a Seção 5.2, indicaram uma expressiva quantidade de unidades contendo somente um CL associado e um total de 227 unidades com menos de 10 CL's vinculados. Esse fato põe em dúvida a composição dessas unidades e, por essas quantidades de CL não serem razoáveis, foi adotado um critério de corte de forma que o experimento final executado abrangesse somente unidades com 10 (dez) ou mais CL's. Seguindo essa premissa, foi emitido um MIS para cada unidade contendo 10 ou mais CL's, totalizando 59 MIS emitidos e 2147 CL's analisados. Nesse processo foram aplicados os limiares de similaridade indicados pelo teste de sensibilidade realizado para cada segmento do CL abrangido. Os dados mapeados, referentes a cada unidade, foram consolidados em função de um conceito existente no organograma da instituição UFF[11] denominado "Unidade de Ensino". Esse conceito, conforme o ANEXO C, abriga várias unidades que foram mapeadas pelo experimento final. Dessa forma, não só é esperada maior expressividade na apresentação desses resultados, como também é apresentada uma associação entre os dados curriculares e a estrutura organizacional da instituição mapeada (UFF).

É oportuno ressaltar que algumas unidades mapeadas não existem nesse citado organograma. Contudo sempre foi possível classificá-las em uma das "Unidades de Ensino" por meio da adoção de um critério subjetivo de afinidade funcional. Como exemplo cita-se o "Departamento de Ciência da Computação". Essa unidade, apesar de não existir no organograma da UFF, possui afinidade funcional com a unidade "Instituto de Computação". Portanto foi classificada na Unidade de Ensino "Campus da Praia Vermelha". O ANEXO D apresenta a composição de cada uma das unidades de ensino, em relação às unidade tratadas pelo MIS.

5.4.1 RESULTADOS OBTIDOS EM ARTIGOS PUBLICADOS

Com o objetivo de seguir a estrutura do MIS (veja a Seção 3.7), os respectivos MIS emitidos com os limiares de similaridade indicados pelo teste de sensibilidade, inicialmente foram consolidados em função dos resultados obtidos em relação à forma com que os artigos publicados foram localizados. A Tabela 14 apresenta as consolidações classificadas em ordem decrescente do percentual de localização por similaridade. Esse percentual de localização é

[11] http://www.uff.br/sites/default/files/images/organogramarelatoriogestao.png

calculado, para cada método de localização, em função do total de publicações em cada unidade de ensino considerada.

É possível observar que o método de localização por similaridade foi mais efetivo na unidade de ensino "Unidades Isoladas de Niterói". É válido também observar que o percentual de localização de artigos publicados pelo título se mantém razoavelmente próximo ao percentual de localização por similaridade em praticamente todas as unidades de ensino. Esse fato, considerando os percentuais apresentados, sugere baixa qualidade no preenchimento do título do artigo publicado, não obstante os razoáveis percentuais obtidos nas localizações por similaridade. É importante destacar que em condições ideais os artigos publicados deveriam ser localizados sempre pelo título, em um CL de coautor.

Tabela 14 – Localização de Artigos Publicados por Unidade de Ensino UFF

Localização de Artigos Publicados por Unidade de Ensino da UFF				
Nome da Unidade de Ensino	**Total de Artigos Publicados**	**Localização pelo Título**	**Localização por Similaridade**	**Total**
UNIDADES ISOLADAS DE NITERÓI	7.821	1.862 (23,81%)	2.075 (26,53%)	3.937 (50,34%)
ORGÃOS SUPLEMENTARES	2.639	875 (33,16%)	695 (26,34%)	1.570 (59,49%)
CAMPUS DO VALONGUINHO	6.096	1.971 (32,33%)	1.506 (24,70%)	3.477 (57,04%)
CAMPUS DA PRAIA VERMELHA	6.621	2.270 (34,28%)	1.513 (22,85%)	3.783 (57,14%)
UNIDADES DO INTERIOR	2.257	444 (19,67%)	227 (10,06%)	671 (29,73%)
CAMPUS DO GRAGOATÁ	7.263	878 (12,09%)	599 (8,25%)	1.477 (20,34%)

Os totais apresentados pela Tabela 15 permitem avaliar a efetividade de cada um dos métodos adotados para localizar um determinado coautor no segmento em consideração. Esses dados são apresentados em ordem decrescente do percentual de localização de coautores pelo método que aplica a cláusula "*like"* (recurso da linguagem SQL) combinada com um tratamento de similaridade (veja o MIS na Seção 3.7) e são apresentados, da esquerda para a direita, seguindo a ordem de aplicação pela heurística, tal como explicitado na Seção 3.4.

É importante ressaltar que as localizações de coautores pelo nome e pelo identificador Lattes representam os métodos mais precisos de localização. Entretanto é notório o baixo

percentual de localização de coautores pelo nome. Esse fato demonstra a baixa qualidade no preenchimento desse atributo pelos titulares dos currículos, apesar das apropriações por similaridade. Outro fato que chama a atenção é o baixo percentual de localização pelo identificador Lattes, lembrando que esse atributo sempre é preenchido com o nome do arquivo de entrada para a geração do script de carga no banco de dados. A combinação desses fatos sugere que os artigos publicados possuem uma baixa taxa de coautores pertencentes à instituição analisada, ou seja, a UFF.

A localização pelo nome de citação, apesar do baixo percentual, demonstra uma relativa uniformidade de apropriação entre as unidades de ensino e comprova a assertividade desse método na heurística desenvolvida.

O expressivo percentual de localização pela combinação da cláusula "*like"* com um tratamento de similaridade pode se justificar pela apropriação de falsos positivos, não obstante a aplicação de limiares de similaridade recomendados pelo teste de sensibilidade.

Tabela 15 – Localizações de Coautores por Unidade de Ensino UFF

Localização de Coautores por Unidade de Ensino da UFF - ARTIGOS PUBLICADOS						
Unidade de Ensino	**Total de Coautores**	**Localização pelo Nome**	**Localização pelo Id Lattes**	**Localização pelo Nome Citação**	**Localização por Like/ Similaridade**	**Total**
ORGÃOS SUPLEMENTARES	9.881	826 (8,36%)	586 (5,93%)	138 (1,40%)	3.493 (35,35%)	5.043 (51,04%)
UNIDADES ISOLADAS DE NITERÓI	21.227	3.263 (15,37%)	892 (4,20%)	212 (1,00%)	7.314 (34,46%)	11.681 (55,03%)
CAMPUS DO GRAGOATÁ	10.251	792 (7,73%)	453 (4,42%)	178 (1,74%)	3.336 (32,54%)	4.759 (46,42%)
UNIDADES DO INTERIOR	4.826	444 (9,20%)	183 (3,79%)	59 (1,22%)	1.270 (26,32%)	1.956 (40,53%)
CAMPUS DO VALONGUINHO	17.991	1.896 (10,54%)	1.406 (7,82%)	515 (2,86%)	4.289 (23,84%)	8.106 (45,06%)
CAMPUS DA PRAIA VERMELHA	21.459	1.558 (7,26%)	1.891 (8,81%)	280 (1,30%)	3.289 (15,33%)	7.018 (32,70%)

A Tabela 16 informa os totais de inconsistências verificadas em cada uma das unidades de ensino, classificados em ordem decrescente pelo total de inconsistências de cada unidade. Logo abaixo de cada total de inconsistência é apresentado o respectivo percentual de inconsistência em relação ao total de coautores existentes na respectiva unidade de ensino. Com esses dados é possível concluir pela maior incidência da inconsistência do tipo "Artigo não encontrado em Coautores". Esse fato pode se justificar por uma apropriação indevida de coautor, caracterizando falsos positivos na identificação do CL de coautores. Vale lembrar que as verificações de inconsistências ocorrem a partir da localização de artigos publicados no CL de coautores. Portanto a apropriação indevida de um CL de coautor certamente

influenciará no aumento desse tipo de inconsistência. Outro fato a destacar refere-se ao baixo percentual de inconsistência do tipo "Nr DOI diferente". Esse resultado está fortemente influenciado pela baixa taxa de preenchimento desse atributo na base de dados, conforme apresentado na Tabela 7.

A inconsistência do tipo "Nr ISSN diferente" apresenta um baixo percentual de incidência, considerando que o atributo "Nr ISSN" possui apenas 10,71% de conteúdo nulo, conforme a mesma Tabela 7. Porém, em uma análise sob outra perspectiva, é aceitável considerar esse percentual de conteúdo nulo bastante expressivo tendo em vista o fato de esse atributo ser muito importante para a identificação de publicações relacionadas a esse segmento.

Tabela 16 – Inconsistências por Unidade de Ensino – Artigos Publicados

Totais de Inconsistências por Unidade de Ensino UFF - Artigos Publicados												
Nome da Unidade	**Total de CL**	**Total de coautores**	**Artigo não encontrado em coautores**	**Título periódico não localizado**	**Ordem coautoria diferente**	**Ano publicação diferente**	**Nr volume diferente**	**Páginas (inicial ou final) diferentes**	**Nr DOI diferente**	**Nr ISSN diferente**	**Total**	**Total Verificações**
UNIDADES ISOLADAS DE NITERÓI	435	21.227	7.744 (36,48%)	141 (0,66%)	995 (4,69%)	60 (0,28%)	637 (3,00%)	1.388 (6,54%)	13 (0,06%)	483 (2,28%)	11.461 (53,99%)	128.929
ORGÃOS SUPLEMENTARES	152	9.881	3.473 (35,15%)	73 (0,74%)	461 (4,67%)	33 (0,33%)	266 (2,69%)	552 (5,59%)	6 (0,06%)	252 (2,55%)	5.116 (51,78%)	59.761
UNIDADES DO INTERIOR	251	4.141	1.285 (31,03%)	24 (0,58%)	161 (3,89%)	15 (0,36%)	131 (3,16%)	216 (5,22%)	2 (0,05%)	120 (2,90%)	1.954 (47,19%)	24.962
CAMPUS DO GRAGOATÁ	532	10.251	3.282 (32,02%)	47 (0,46%)	512 (4,99%)	33 (0,32%)	223 (2,18%)	565 (5,51%)	12 (0,12%)	158 (1,54%)	4.832 (47,14%)	58.940
CAMPUS DO VALONGUINHO	399	19.384	4.327 (22,32%)	154 (0,79%)	1.233 (6,36%)	93 (0,48%)	564 (2,91%)	1.100 (5,67%)	35 (0,18%)	455 (2,35%)	7.961 (41,07%)	98.928
CAMPUS DA PRAIA VERMELHA	367	21.396	3.197 (14,94%)	183 (0,86%)	1.485 (6,94%)	84 (0,39%)	498 (2,33%)	1.154 (5,39%)	55 (0,26%)	496 (2,32%)	7.152 (33,43%)	99.991

Considerando que os resultados obtidos no teste de sensibilidade possuem proporcionalidade em relação à totalidade dos dados utilizados como entrada para o experimento, as medidas de precisão e cobertura foram aplicadas aos totais de inconsistências com o objetivo de obter os valores da precisão e da cobertura dos resultados obtidos. Observando os dados apresentados pela Tabela 17, os quais estão classificados em ordem decrescente do valor da projeção da cobertura, é aceitável concluir que o experimento realizado nesse segmento do CL apresenta uma efetividade bastante significativa. Os totais de inconsistências obtidos expressam uma cobertura de 92,81% e os valores apresentados na coluna "Projeção cobertura" projetam os totais de inconsistência em um cenário de 100% de cobertura para a projeção de precisão obtida.

Os valores apresentados na coluna "Projeção precisão" expressam o total de inconsistências, considerando a precisão obtida pelo teste de sensibilidade, qual seja 98,30%.

Significa dizer que, do total de inconsistências encontradas, o teste de sensibilidade estima que 98,30% podem ser consideradas verdadeiras e esse valor resultante possui 92,81% de cobertura. Sendo assim, é aceitável a projeção de que esse segmento do CL em estudo apresenta os totais de inconsistências indicados pela coluna "Projeção cobertura". Complementando esse raciocínio, ainda é possível concluir que esse alto percentual de cobertura associado a um alto percentual de precisão, conforme apresentado, são indicativos de que esse segmento do CL apresenta alta taxa de inconformidade.

Tabela 17 – Medidas de precisão e cobertura – Artigos Publicados

Projeção das medidas de precisão e cobertura por Unidade de Ensino UFF - Artigos Publicados			
Unidade de Ensino	**Total Inconsistência**	**Projeção Precisão (Precision = 98,30%)**	**Projeção Cobertura (Recall = 92,81%)**
UNIDADES ISOLADAS DE NITERÓI	11.461	11.266,16	12.138,95
CAMPUS DO VALONGUINHO	7.961	7.825,66	8.431,92
CAMPUS DA PRAIA VERMELHA	7.152	7.030,42	7.575,06
ORGÃOS SUPLEMENTARES	5.116	5.029,03	5.418,63
CAMPUS DO GRAGOATÁ	4.832	4.749,86	5.117,83
UNIDADES DO INTERIOR	1.954	1.920,78	2.069,59

5.4.2 RESULTADOS OBTIDOS EM TRABALHOS EM EVENTOS

Tal como detalhado na seção anterior, os resultados obtidos nesse segmento seguem a estrutura do MIS detalhada na Seção 3.7. A Tabela 18 apresenta as totalizações de localizações de trabalhos em eventos classificadas em ordem decrescente em função do método de localizações por similaridade.

Tabela 18 – Localização de Trabalhos em Eventos por Unidade de Ensino UFF

Localização de Trabalhos em Eventos por Unidade de Ensino da UFF				
Nome da Unidade de Ensino	**Total de Trabalhos em Evento**	**Localização pelo Título**	**Localização por Similaridade**	**Total**
UNIDADES ISOLADAS DE NITERÓI	14.763	2.364 (16,01%)	3.285 (22,25%)	5.649 (38,26%)
ORGÃOS SUPLEMENTARES	3.633	604 (16,63%)	680 (18,72%)	1.284 (35,34%)
CAMPUS DO VALONGUINHO	12.329	2.036 (16,51%)	2.102 (17,05%)	4.138 (33,56%)
UNIDADES DO INTERIOR	5.871	1.066 (18,16%)	955 (16,27%)	2.021 (34,42%)
CAMPUS DA PRAIA VERMELHA	13.432	2.690 (20,03%)	1.733 (12,90%)	4.423 (32,93%)
CAMPUS DO GRAGOATÁ	12.114	1.086 (8,96%)	972 (8,02%)	2.058 (16,99%)

Tal como apresentado na seção anterior, o percentual de localização é calculado, para cada método de localização, em função do total de publicações em cada unidade de ensino considerada. Cabe destacar que a unidade de ensino "Unidades Isoladas de Niterói" apresenta

a maior incidência de localização por similaridade e, de forma geral, os percentuais de localização das publicações pelo nome e por similaridade se mantém razoavelmente próximos e não maior que 22,25%.

Os totais apresentados pela Tabela 19 contribuem para a avaliação da efetividade de cada um dos métodos adotados para localizar um determinado coautor. Esses totais, da mesma forma que o segmento tratado anteriormente, estão dispostos em ordem decrescente do percentual de localização de coautores pelo método que aplica a cláusula "*like"* (recurso da linguagem SQL) combinada com um tratamento de similaridade (veja o MIS na Seção 3.7) e são apresentados, da esquerda para a direita, seguindo a ordem de aplicação pela heurística, tal como explicitado na Seção 3.4.

Os dados apresentados informam um baixo percentual de localização de coautores pelo nome. Esse fato pode ser atribuído à baixa qualidade no preenchimento desse atributo pelos titulares dos currículos.

Tabela 19 – Localização de Coautores por Unidade de Ensino UFF – Trabalhos em Eventos

Localização de Coautores por Unidade de Ensino da UFF - TRABALHOS EM EVENTOS						
Unidade de Ensino	**Total de Coautores**	**Localização pelo Nome**	**Localização pelo Id Lattes**	**Localização pelo Nome Citação**	**Localização por Like/ Similaridade**	**Total**
CAMPUS DO GRAGOATÁ	19.282	2.148 (11,14%)	286 (1,48%)	72 (0,37%)	9.376 (48,63%)	11.882 (61,62%)
UNIDADES ISOLADAS DE NITERÓI	42.290	6.555 (15,50%)	1.155 (2,73%)	81 (0,19%)	18.681 (44,17%)	26.472 (62,60%)
UNIDADES DO INTERIOR	13.664	2.154 (15,76%)	242 (1,77%)	60 (0,44%)	5.822 (42,61%)	8.278 (60,58%)
ORGÃOS SUPLEMENTARES	11.974	1.211 (10,11%)	422 (3,52%)	63 (0,53%)	4.588 (38,32%)	6.284 (52,48%)
CAMPUS DO VALONGUINHO	34.194	4.238 (12,39%)	1.531 (4,48%)	147 (0,43%)	12.169 (35,59%)	18.085 (52,89%)
CAMPUS DA PRAIA VERMELHA	33.901	4.029 (11,88%)	1.404 (4,14%)	92 (0,27%)	11.118 (32,80%)	16.643 (49,09%)

Outro fato a destacar é o baixo percentual de localização pelo identificador Lattes. A combinação desses resultados sugere que os trabalhos em eventos publicados possuem uma baixa taxa de coautores pertencentes à instituição analisada, ou seja, a UFF.

A localização de coautores pelo nome de citação apresenta um baixo percentual e uma relativa uniformidade de apropriação entre as unidades de ensino analisadas. Esse fato sugere que o hábito de se preencher o nome de um coautor com o seu respectivo nome de citação não é uma prática comum nesse segmento estudado.

A localização de coautores pelo método da combinação da cláusula "*like"* com um tratamento de similaridade apresenta um percentual bastante significativo, tal como observado no segmento analisado anteriormente. Esse fato pode revelar alta incidência de apropriação de falsos positivos, apesar da aplicação dos limiares de similaridade recomendados pelo teste de sensibilidade.

As inconsistências identificadas estão discriminadas por unidade de ensino na Tabela 20 e classificadas em ordem decrescente do total de inconsistência por unidade de ensino. Logo abaixo de cada total de inconsistência é apresentado o respectivo percentual em relação ao total de coautores existentes na respectiva unidade de ensino.

Esses dados informam que a maior incidência de inconsistência verificada foi a do tipo "Trabalho não encontrado em Coautores". Esse fato pode ser justificado por uma apropriação indevida de coautor, uma vez que a busca por um trabalho em evento é precedida pela identificação do CL do coautor. Portanto, a apropriação indevida de um coautor certamente influenciará no aumento desse tipo de inconsistência.

Tabela 20 – Inconsistências por Unidade de Ensino UFF – Trabalhos em Evento

Totais de Inconsistências por Unidade de Ensino UFF - Trabalhos em Eventos												
Nome da Unidade	**Total de CL**	**Total de coautores**	**Trabalho não encontrado em coautores**	**Título evento não localizado**	**Ordem coautoria diferente**	**Ano publicação diferente**	**Nr volume diferente**	**Páginas (inicial ou final) diferentes**	**Nr DOI diferente**	**Nr ISBN diferente**	**Total**	**Total Verificações**
UNIDADES ISOLADAS DE NITERÓI	435	42.290	20.823 (49,24%)	1.406 (3,32%)	1.452 (3,43%)	165 (0,39%)	1.241 (2,93%)	872 (2,06%)	0 (0,00%)	427 (1,01%)	26.386 (62,39%)	281.371
UNIDADES DO INTERIOR	251	13.664	6.257 (45,79%)	557 (4,08%)	526 (3,85%)	60 (0,44%)	495 (3,62%)	325 (2,38%)	0 (0,00%)	285 (2,09%)	8.505 (62,24%)	90.378
CAMPUS DO GRAGOATÁ	532	19.282	9.747 (50,55)	424 (2,20%)	665 (3,45%)	69 (0,36%)	491 (2,55%)	300 (1,56%)	0 (0,00%)	145 (0,75%)	11.918 (61,81%)	130.927
CAMPUS DO VALONGUINHO	399	34.194	13.947 (40,79%)	1.295 (3,79%)	1.470 (4,30%)	147 (0,43%)	1.179 (3,45%)	694 (2,03%)	1 (0,00%)	312 (0,91%)	19.045 (55,70%)	202.916
ORGÃOS SUPLEMENTARES	152	11.974	5.000 (41,76%)	371 (3,10%)	391 (3,27%)	36 (0,30%)	261 (2,18%)	243 (2,03%)	0 (0,00%)	116 (0,97%)	6.418 (53,60%)	73.641
CAMPUS DA PRAIA VERMELHA	378	33.901	12.220 (36,05%)	1.001 (2,95%)	1.520 (4,48%)	185 (0,55%)	1.198 (3,53%)	620 (1,83%)	3 (0,01%)	334 (0,99%)	17.081 (50,38%)	195.002

Outro fato que merece destaque diz respeito ao baixo percentual de inconsistência do tipo "Nr DOI diferente". Esse resultado está fortemente influenciado pela alta taxa de conteúdo nulo para esse atributo na base de dados, qual seja 99,30%, conforme apresentado na Tabela 8.

A inconsistência do tipo "Nr ISBN diferente" apresenta um baixo percentual de incidência, o qual é admissível ser justificado pela alta taxa de preenchimento nulo, no valor de 85,77%, também conforme a mesma Tabela 8.

Tal como procedido na Seção 5.4.1, foi presumido que os resultados obtidos no teste de sensibilidade possuem proporcionalidade em relação aos dados utilizados como entrada para o experimento. Portanto, as medidas de precisão e cobertura foram aplicadas aos totais

de inconsistências como forma de se dimensionar a precisão e a cobertura dos resultados obtidos em cada unidade de ensino. A Tabela 21 apresenta os resultados das projeções dessas medidas, classificadas em ordem decrescente do valor da projeção da cobertura. Esses dados indicam que o experimento realizado nesse segmento do CL apresenta alta efetividade. Os totais de inconsistências obtidos expressam uma cobertura de 90,09% e, tal como no segmento anteriormente analisado, os valores da coluna "Projeção cobertura" projetam os totais de inconsistência em um cenário de 100% de cobertura para a projeção de precisão obtida. Os valores apresentados na coluna "Projeção precisão" expressam os totais de inconsistências, considerando a precisão obtida pelo teste de sensibilidade, qual seja 92,54%. Isto é, do total de inconsistências encontradas, 92,54% podem ser consideradas verdadeiras e esse valor resultante possui 90,09% de cobertura. Sendo assim, é aceitável inferir que o segmento em estudo apresenta os totais de inconsistências indicados pela coluna "Projeção cobertura". Essa alta taxa de cobertura associada a uma alta precisão são indicativos de que esse segmento do CL também apresenta uma significativa taxa de inconformidade.

Tabela 21 - Medidas de precisão e cobertura – Trabalhos em Eventos

Projeção das medidas de precisão e cobertura por Unidade de Ensino UFF - Trabalhos em Eventos			
Unidade de Ensino	**Total Inconsistência**	**Projeção Precisão Precision = 92,54%**	**Projeção Cobertura Recall = 90,09%**
UNIDADES ISOLADAS DE NITERÓI	26.386	24.417,60	27.130,67
CAMPUS DO VALONGUINHO	19.045	17.624,24	19.582,49
CAMPUS DA PRAIA VERMELHA	17.081	15.806,76	17.563,06
CAMPUS DO GRAGOATÁ	11.918	11.028,92	12.254,35
UNIDADES DO INTERIOR	8.505	7.870,53	8.745,03
ORGÃOS SUPLEMENTARES	6.418	5.939,22	6.599,13

5.4.3 ANÁLISE GLOBAL DOS RESULTADOS OBTIDOS

A análise global dos resultados obtidos reúne totalizações sobre os dois segmentos abrangidos por essa dissertação e também totalizações de inconsistências por unidade de ensino considerando os dois segmentos do CL tratados nessa dissertação. Portanto, esses dados totalizadores possibilitam uma visão analítica global da instituição UFF.

A Tabela 22 totaliza as localizações de publicações para cada um desses segmentos. No segmento artigos publicados a localização pelo título foi mais significativa do que no segmento trabalhos em eventos. Esse fato induz a interpretação de que esse segmento possui mais qualidade no preenchimento do título da respectiva publicação pelos titulares dos CL's analisados. Entretanto o percentual total de localização pelo título, no valor de 19,08%, revela

baixa qualidade no preenchimento do título de publicações, no que pese a maior apropriação de publicações pelo título do que pelo tratamento de similaridade em todos os segmentos considerados e, consequentemente, na instituição como um todo.

Em uma visão geral é aceitável observar que a aplicação da heurística desenvolvida combinada com o tratamento de similaridade praticado foi mais efetiva na localização de artigos publicados do que na localização de trabalhos em eventos. Entretanto, o percentual total de localização, no valor de 36,31%, pode ser considerado baixo e se justificar, conforme citado anteriormente, pela baixa frequência de publicações entre coautores pertencentes à mesma instituição analisada.

Tabela 22 – Localização de publicações por segmento analisado

Localização de Publicações por Segmento Analisado				
Segmento Analisado	**Total de Publicações**	**Localizações pelo Título**	**Localizações por Similaridade**	**Total de Localizações**
Artigos Publicados	32.697	8.250 (25,23%)	6.615 (20,23%)	14.865 (45,46%)
Trabalhos em Eventos	62.142	9.846 (15,84%)	9.727 (15,65%)	19.573 (31,50%)
Total da Instituição	94.839	18.096 (19,08%)	16.342 17,23%)	34.438 (36,31%)

Os totais de localização informados pela Tabela 23 apresentam baixo percentual global de localização de coautores pelo nome. Esse fato pode se justificar por dois motivos. O primeiro é a suposição de que parte significativa dos coautores pertence a outras instituições e um segundo motivo sugere a baixa qualidade no preenchimento do nome dos coautores.

O baixo percentual de localização pelo identificador Lattes pode ratificar a afirmação anterior de que os coautores pertencem a outras instituições. Contudo, vale observar que esse atributo, o identificador Lattes, é pouco preenchido nas tabelas "AutorArtigo" e "AutorTrabEvento" e essas tabelas representam dados de autoria, os quais são básicos para a recuperação de currículos de coautores. Logo, esse fato reduz a possibilidade de recuperação de um currículo de coautor pelo identificador Lattes, apesar desse atributo sempre existir na tabela "CurriculumVitae".

A localização pelo nome de citação, apesar do baixo percentual de incidência, pode ser considerada como um recurso relativamente efetivo no segmento artigos publicado, além de apontar uma distorção semântica na medida em que o nome de citação é erroneamente informado no lugar do nome completo de coautores.

A apropriação de coautores pela combinação da cláusula "*like"* (recurso da linguagem SQL) combinada com um tratamento de similaridade apresenta os maiores percentuais de

localização. Esse fato, conforme comentado na análise individual realizada dos segmentos estudados, pode ser atribuído à apropriação de falsos positivos, não obstante a aplicação dos limiares de similaridade indicados pelo teste de sensibilidade.

Uma percepção global dos resultados indica que houve 52,38% de localizações de coautores e esse fato está coerente com a interpretação de que parte significativa dos coautores pertence a outras instituições.

Tabela 23 – Localização de coautores por segmento analisado

Localização de Coautores por Segmento Analisado						
Segmento Analisado	**Total de Coautores**	**Localizações pelo Nome**	**Localizações pelo Id Lattes**	**Localização pelo Nome Citação**	**Localização por Like/ Similaridade**	**Total**
Artigos Publicados	85.635	8.779 (10,25%)	5.411 (6,32%)	1.382 (1,61%)	22.991 (26,85%)	38.563 (45,03%)
Trabalhos em Eventos	155.305	20.335 (13,09%)	5.040 (3,46%)	515 (0,33%)	61.754 (39,76%)	87.644 (56,43%)
Total da Instituição	240.940	29.114 (12,08%)	10.451 (4,38%)	1.897 (0,79%)	84.745 (35,17%)	126.207 (52,38%)

A Tabela 24 consolida as inconsistências em função dos segmentos do CL tratados nessa dissertação. Os dados estão classificados pelo total de inconsistências verificadas em cada segmento analisado. De forma coerente com a análise individual de cada segmento, a inconsistência mais frequente foi "Publicações não encontrada em coautores". Esse resultado corrobora comentários anteriormente realizados; ou seja, é muito provável que uma quantidade significativa de coautores não pertença à instituição UFF e simultaneamente esteja havendo a apropriação de falsos positivos na localização do currículo de coautores. Não obstante a aplicação de limiares de similaridade recomendados pelo teste de sensibilidade.

O percentual de inconsistência verificado na instituição analisada, qual seja 52,91%, mesmo considerando a possibilidade de apropriação de falsos positivos, conforme menção anterior, pode ser considerado um valor com expressividade suficiente para ameaçar a integridade de possíveis indicadores bibliométricos gerados a partir desses dados.

Tabela 24 – Inconsistências por segmento analisado

Totais de Inconsistências por Segmento do CL Analisado												
Segmento Analisado	**Total de currículos**	**Total de coautores**	**Publicação não encontrada em coautores**	**Origem da publicação não localizada**	**Ordem coautoria diferente**	**Ano publicação diferente**	**Nr volume diferente**	**Páginas (inicial ou final) diferentes**	**Nr DOI diferente**	**Nr ISSN/ ISBN diferente**	**Total Inconsistências**	**Total Verificações**
Trabalhos em Eventos	2.147	155.305	67.994 (43,78%)	5.054 (3,25%)	6.024 (3,88%)	662 (0,43%)	4.865 (3,13%)	3.054 (1,97%)	4 (0,00%)	1.619 (1,04%)	89.353 (57,53%)	974.235
Artigos Publicados	2.136	86.280	23.308 (27,01%)	622 (0,72%)	4.847 (5,62%)	318 (0,37%)	2.319 (2,69%)	4.975 (5,77%)	123 (0,14%)	1.964 (2,28%)	38.476 (44,59%)	471.511
Total da Instituição	2.147	241.585	91.302 (37,79%)	5.676 (2,35%)	10.871 (4,50%)	980 (0,41%)	7.184 (2,97%)	8.029 (3,32%)	127 (0,05%)	3.583 (1,48%)	127.829 (52,91%)	1.445.746

A Tabela 25 dispõe de outra forma os dados globais obtidos. Esses dados sumarizam, por unidade de ensino, as inconsistências verificadas em cada segmento analisado. Os dados estão classificados em ordem decrescente do total de inconsistências da unidade de ensino. Os percentuais das colunas 1 e 2 são calculados em função do total de inconsistências por segmento do CL, e o percentual da coluna 3 é calculado em função do total de inconsistências da instituição. Dessa forma é possível verificar a influência das inconsistências ocorridas em cada unidade de ensino no total de instituição analisada. Portanto, é possível verificar com facilidade que a unidade "Unidades Isoladas de Niterói" apresenta a maior incidência de inconsistências. É importante destacar que essa unidade também apresenta os maiores percentuais de inconsistências nos dois segmentos abrangidos pelo presente estudo. Esse fato desperta a atenção para o fato dos percentuais de inconsistências dos segmentos analisados, em cada unidade de ensino, apresentarem certa uniformidade, com pequena variação, exceto pela unidade "Órgãos Suplementares" que apresenta uma diferença de 6,12% entre as inconsistências de cada segmento.

Tabela 25 – Totalizações de Inconsistências por Unidades de Ensino

Unidade de Ensino	Total de Inconsistência Artigo Publicado (1)	Total de Inconsistências Trabalhos em Eventos (2)	Total Unidade Ensino (3)
UNIDADES ISOLADAS DE NITERÓI	11.461 (29,79%)	26.386 (29,53%)	37.747 (29,53%)
CAMPUS DO VALONGUINHO	7.961 (20,69%)	19.045 (21,31%)	27.006 (21,13%)
CAMPUS DA PRAIA VERMELHA	7.152 (18,59%)	17.081 (19,12%)	24.233 (18,96%)
CAMPUS DO GRAGOATÁ	4.832 (12,56%)	11.918 (13,34%)	16.750 (13,10%)
ORGÃOS SUPLEMENTARES	5.116 (13,30%)	6.418 (7,18%)	11.534 (9,02%)
UNIDADES DO INTERIOR	1.954 (5,08%)	8.505 (9,52%)	10.459 (8,18%)
TOTAL POR SEGMENTO CL	38.476	89.353	127.829

De forma geral o segmento "Trabalhos em Eventos" apresenta a maior incidência de inconsistências, sendo responsável por 57,53% das inconsistências da instituição, conforme informado na Tabela 24.

A projeção global das medidas de precisão e cobertura para a instituição UFF é apresentada na Tabela 26. A partir desses dados é possível inferir que a soma das projeções de precisão para os segmentos considerados representam 93,81% do total de inconsistências verificadas na instituição. Portanto, esse percentual significa a medida de precisão da sistemática aplicada abrangendo toda a instituição. Em complemento, a projeção global de precisão representa 90,92% da projeção global da medida de cobertura. Isso significa que, em termos globais, o experimento realizado apresenta 93,81% de precisão e 90,92% de cobertura.

Essas medidas resultam em um F-Measure de 92,34%, o qual pode ser considerado como bastante efetivo.

Tabela 26 - Medidas de precisão e cobertura por segmento analisado

Projeção das medidas de precisão e cobertura por Segmento Analisado					
Segmento Currículo Lattes	**Total Inconsistência**	**Precisão (%)**	**Projeção Precisão**	**Cobertura (%)**	**Projeção Cobertura**
Artigo Publicado	38.476	98,30%	37.821,9	92,81%	40.751,97
Trabalhos em Eventos	89.353	92,54%	82.687,27	90,09%	91.782,66
Total Instituição	127.829	**93,81%**[12]	119.911,24	**90,92%**[13]	131.883,22

5.5 AMEAÇAS À VALIDADE DOS RESULTADOS OBTIDOS

O processo de desenvolvimento da abordagem apresentada foi continuamente acompanhado por uma criteriosa validação de cada resultado parcial obtido. Entretanto, algumas fases do processo e premissas assumidas podem embutir algumas ameaças a esses citados resultados.

O teste de sensibilidade exigiu a construção de gabaritos de verdadeiros positivos e verdadeiros negativos considerando cada título de publicação existente em cada coautor de cada segmento abrangido. Esse processo, inevitavelmente, exigiu uma cuidadosa conferência visual. Mesmo considerando apenas 43 CL's e tendo sido adotada uma forma de realização lenta e altamente criteriosa, é possível a existência de erros na classificação dessas publicações e, consequentemente, a presença de distorções na apuração dos limiares ideais de similaridade.

Outro fato que vale ser destacado é a presunção de proporcionalidade entre a amostra usada para gerar o gabarito e a base de dados utilizada no experimento final. Essa presunção é o fator que justifica a aplicação das medidas obtidas no teste de sensibilidade na base de dados final e é inferida por se considerar que os dados referenciados possuem mesma natureza semântica e, portanto, guardam alguma proporcionalidade. Contudo, é válido e necessário reconhecer que esses dados de projeção, por carecerem de um amparo mais científico, podem ter a sua validade experimental reduzida e, dessa forma, não expressar na totalidade da efetividade apresentada.

[12] (119911,24*100)/127829

[13] (119911,24*100)/131883,22

5.6 CONSIDERAÇÕES FINAIS

Este capítulo analisou os resultados obtidos a partir da execução dos experimentos finais destinados à emissão do mapeamento de inconsistência, tendo como dados de entrada o conjunto de CL's do corpo docente da Universidade Federal Fluminense.

A execução dos experimentos foi precedida de uma análise preliminar dos dados de entrada com o objetivo de ratificar os requisitos especificados a partir da análise investigativa discutida na Seção 3.3 e também verificar possíveis necessidades de ajustes na heurística detalhada na Seção 3.4, decorrentes de alguma nova característica desses dados de entrada. Essa análise preliminar identificou uma quantidade duvidosa de currículos em várias unidades da instituição considerada e determinou uma seleção nos dados de entrada como forma de reduzir possíveis distorções no mapeamento intencionado. Sendo assim, foram consideradas somente unidades contendo 10 (dez) ou mais currículos.

Em um momento seguinte foi realizado um teste de sensibilidade abrangendo os segmentos "Artigos Publicados" e "Trabalhos em Eventos" com o objetivo de identificar os limiares de similaridade ideais tanto para os títulos de publicações quanto para o nome de coautores nesses dois segmentos e também determinar uma expectativa de efetividade para o experimento final. Esse teste de sensibilidade, por envolver uma atividade manual para a geração de um gabarito para servir de referência, utilizou como amostra os dados de 43 currículos de professores do IC/UFF e o seu resultado foi então aplicado na execução dos experimentos finais.

O experimento final consistiu da emissão de um MIS para cada unidade e segmento considerados e da classificação desses resultados em função do conceito "Unidade de Ensino". Este conceito pertence à estrutura organizacional da UFF e é considerado relevante na medida em que associa os resultados obtidos com a estrutura organizacional dessa instituição. Nesse processo de associação houve a decomposição dos MIS's emitidos em função não só dos segmentos abrangidos, como também dos blocos que compõem esse referido mapa, conforme detalhado na Seção 3.7. Essa decisão foi motivada por facilitar a interpretação dos resultados obtidos na medida em que permite uma análise mais pontual sobre esses resultados.

A seguir, os resultados obtidos em cada segmento tratado foram discutidos separadamente e originaram a consolidação dos resultados para a instituição, a qual foi também analisada.

Com o objetivo de aferir a efetividade dos resultados obtidos tanto em cada segmento quanto na totalidade da instituição, as medidas de precisão e cobertura, obtidas no teste de sensibilidade, foram projetadas e, tal como se supunha, apresentaram resultados bastante satisfatórios e, portanto, respondem satisfatoriamente a questão enunciada nessa dissertação. Entretanto, é fundamental ressaltar que a aplicação dessas medidas de efetividade presume uma proporcionalidade entre os dados de entrada considerados no teste de sensibilidade e os dados utilizados nos experimentos finais. Essa presunção, por carecer de uma comprovação científica, pode apresentar distorções nessas medidas finais de efetividade.

CAPÍTULO 6 – CONCLUSÃO

Esta dissertação detalhou uma abordagem destinada a aferir a integridade da Plataforma Lattes por meio da geração de um mapeamento de inconsistências, especificamente desenvolvidos para os segmentos “Artigos Publicados” e “Trabalhos em Eventos”, ambos integrantes do Currículo Lattes. O desenvolvimento desse estudo foi fortemente inspirado na adoção de técnicas e sistemáticas aplicadas nos processos de desambiguação de citações bibliográficas e deduplicação de publicações no contexto de bibliotecas digitais.

A solução desenvolvida utiliza uma heurística especificamente desenvolvida em função dos estudos preliminares e complementares realizados. Essa heurística recorre a um tratamento de similaridade naquelas situações onde esse recurso possa contribuir para aumentar a efetividade dos resultados. Foi também realizado um teste de sensibilidade com o objetivo de determinar os limiares ideais de similaridade por meio da determinação das medidas de precisão, cobertura e a média harmônica F-Measure entre essas medidas, a qual indicou a previsão de efetividade. Por fim essas medidas foram projetadas nos resultados finais obtidos e indicaram uma efetividade bastante satisfatória.

De acordo com os critérios de medição de inconsistência adotados, a análise realizada demonstra que os dados curriculares desses dois segmentos abrangidos possuem um expressivo volume de inconsistência, o qual pode ser considerado como uma ameaça à integridade da Plataforma Lattes, de tal forma que a sua confiabilidade pode se encontrar em níveis tão baixos que inviabilizem o cumprimento do seu objetivo precípuo que é, em última análise, servir a sociedade como principal fonte de informações referente à comunidade de pesquisadores brasileiros, estudantes, gestores, profissionais e demais atores do sistema nacional de Ciência, Tecnologia e Inovação.

A contribuição dessa dissertação pretende atingir uma dimensão maior do que a apresentação de uma sistemática de aferição da integridade das informações existentes na Plataforma Lattes. Sendo assim, a principal perspectiva de contribuição está centrada em evidenciar a fragilidade em que se encontra o registro de dados nesse importante instrumento de aferição da produção científica nacional, que é a Plataforma Lattes e, como consequência, despertar os interessados para buscar soluções visando o seu aprimoramento.

6.1 TRABALHOS FUTUROS

Existem varias formas de se abordar questões relacionadas com a medição de inconsistências na Plataforma Lattes. Contudo, tal como enfatizado nesse estudo, torna-se necessário propagar, no ambiente acadêmico e nos demais ambientes correlatos, a necessidade de um despertar para o estabelecimento de medidas de confiabilidade para esse instrumento tão relevante. Nesse sentido, este presente trabalho deve ser entendido como uma pequena colaboração e, portanto, pode ser considerada como introdutória.

No decorrer dessa dissertação algumas questões foram se revelando e por algumas limitações principalmente à restrição de tempo, foram deixadas para oportunidades futuras.

É possível imaginar a verificação de integridade sendo ponderada pela frequência com que um determinado titular de um CL publica com um determinado coautor, ou seja, inconsistências verificadas em um coautor que publica com muita frequência com um determinado titular de CL receberia um peso maior do que aquelas verificadas em coautores poucos frequentes.

Outro aspecto a destacar é a possibilidade de se classificar as inconsistências em função de um possível critério que retrate sua gravidade. Por exemplo: um erro na ordem de autoria deve ou não possuir a mesma relevância de um erro nas páginas inicial e final? A inexistência de uma publicação no currículo de um coautor pode ou não ser considerada como mais grave do que uma inconformidade no número do volume?

É ainda possível imaginar critérios de aferição de inconsistências específicos para cada um dos segmentos do CL.

Outra forma de obtenção do mapeamento poderia considerar a possibilidade de uma publicação ser classificada de forma diferente pelos diferentes coautores. Por exemplo, uma determinada publicação pode ser lançada como capítulo de livro por um coautor e como artigo publicado por outro coautor e ainda como trabalho em evento por um terceiro coautor. Portanto, uma abordagem heurística poderia tentar localizar as publicações em diferentes segmentos.

Este estudo, complementado por esse conjunto de sugestões, faz emergir uma reflexão sobre o tema com vistas a buscar alternativas de solução para uma grande questão: como manter a Plataforma Lattes em níveis satisfatórios de integridade, de forma a assegurar a realização do seu objetivo?

REFERÊNCIAS

BORGES, E. N.; BECKER, K.; *et al.* A Classification-Based Approach for Bibliographic Metadata Deduplication. In: IADIS International Conference WWW/INTERNET 2011, 2011, [S.l: s.n.], 2011.

BORGES, E. N.; CARVALHO, M. G.; *et al.* An unsupervised heuristic-based approach for bibliographic metadata deduplication. *Information Processing and Management*, Nr 5. v. 47, p. 706–718, 2011.

BUSH, VANNEVAR. AS WE MAY THINK. *The Atlantic Monthly. The electronic version was prepared by Denys Duchier , April 1994.*., jul. 1945.

CARVALHO, M. G. *et al.* Learning to Deduplicate. *Learning to Deduplicate*, p. 11–15, jun. 2006.

CHATER, Nick. Simplicity and the mind. *The Psychologist*, UK, nov. 1997. , p. 495–498.

CHEN, Y.; WAN, A.; LIU, W. A fast parallel algorithm for finding the longest common sequence of multiple biosequences. *BMC Bioinformatics*, PMID: 17217522PMCID: PMC1780122, v. 7, n. Suppl 4, p. S4, 12 dez. 2006. Acesso em: 20 jun. 2014.

CHRISTEN, P. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 9. v. 24, 7 jun. 2011.

CNPQ. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 15 ago. 2012a.

CNPQ. *CNPq - Demanda e Atendimento: bolsas no país por área de conhecimento*. Disponível em: <http://estatico.cnpq.br/portal/paineis/painel_bolsa_pais_area/index.html>.

CORMEN, T. H. *et al. Introduction to Algorithms*. Third Edition ed. [S.l.]: The MIT Press, 2009.

DONGWON L. *et al.* Are your citation clean? *ACM Communications*, v. 50, n. 12, p. 33 – 38, dez. 2007.

DORNELES, C. F. *et al.* Measuring Similarity Between Collection of Values. In: 6TH Annual ACM International Workshop on WEB Information and Data Management, 2004, New York: [s.n.], 2004. p. 56–63.

DOS SANTOS, M. Contribuição à compreensão da informação no ambiente das organizações: um ensaio teórico. *Vll Semead - Seminário em Administração - FEA-USP*, 11 ago. 2004.

DUTRA, L. H. DE A. A Ciência e o Conhecimento Humano como Construção de de Modelos. *Philósophos - Revista de Filosofia*, v. 11, n. 2, p. 247–286, 6 set. 2008. Acesso em: 19 jun. 2014.

ELLIOT, S. A Survey of Author Name Disambiguation: 2004 to 2010. *Library Philosophy and Practice*, out. 2010.

FERREIRA, A. A.; GONÇALVES, M. A.; LAENDER, A. H. F. A Brief Survey of Automatic Method for Author Name Disambiguation. *SIGMOD Record Web Edition*, 2. v. Vol. 41, p. 15–26, jun. 2012.

FUHR, N. *et al.* Evaluation of digital libraries. *International Journal on Digital Libraries*, v. 8, n. 1, p. 21–38, 1 nov. 2007.

GIGERENZER, G. Why Heuristics Work. *Association for Psychologial Science*, 1. v. 3, p. 20–29, 2008. Acesso em: 16 maio 2014.

GIGERENZER, G.; ENGEL, C. *Heuristics and the Law*. [S.l.]: MIT Press, 2006.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining Concepts and Techniques*. 3rd Edition ed. [S.l.]: Elsevier, 2012.

LANE, J. Let's make science metrics more scientific. *Nature*, v. 464, n. 7288, p. 488–489, 25 mar. 2010.

LUHN, H. P. A Statistical Approach to Mechanized Encoding 1 and Searching of Literary Information. *IBM Journal of Research and Development*, v. vol 1 (4), p. 309–317, out. 1957.

MICHALEWICZ, Z.; FOGEL, D. B. *How to Solve It: Modern Heuristics*. [S.l.]: Springer, 2004.

MICHALSKI, R. S. A theory and methodology of inductive learning. *Artificial Intelligence*, v. 20, p. 111–161, fev. 1983.

MUGNAINI, R.; JANNUZZI, P. DE M.; QUONIAM, L. Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base Pascal. *Ciência da Informação*, v. 33, n. 2, p. 123–131, ago. 2004.

NAVARRO, G. A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, v. 33, n. 1, p. 31–88, mar. 2001.

NEWCOMBE, H. B. *et al.* Automatic Linkage of Vital Records. *Workshop on Exact Matching Methodologies*, v. 130, p. 954–959, dez. 1985.

OLIVEIRA, J. W. A. *Uma Estratégia para Remoção de Ambiguidades na Identificação de Autoria de Objetos Bibliográficos*. 2005. Dissertação de Mestrado – Universidade Federal de Minas Gerais, UFMG, 2005. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/RVMR-6EAGQK/jeanwanderleialvesoliveira.pdf?sequence=1>. Acesso em: 18 mar. 2014.

PEREIRA, D. A. *et al.* Using web information for author name disambiguation. *9th ACM/IEEE-CS joint conference on Digital libraries*, Joint conference on Digital Library. p. 49–58, 2009.

SARAWAGI, S.; BHAMIDIPATY, A. Interactive Deduplication Using Active Learning. ACM SIGKDD International Conference on Knowledge discovery and data mining, 2002, New York, NY, USA. p. 269–278.

SAYÃO, K. F.; SALES, L. F. Curadoria Digital: um novo patamar para preservação de dados digitais de pesquisa. Informação & Sociedade: Estudos. v. 22, p. 179–191, dez. 2012.

SCALATON, L. P.; GARCIA, R. E. Empacotamento de Experimentos Controlados com Abordagem Evolutiva Baseada em Ontologia. *Interciência & Sociedade*, v. Vol. 3, n. 1, p. 20–28, 2014.

SINGHAL, A. Modern Information Retrieval: A Brief Over view. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, p. 1–9, 2001.

ULLMAN, J. D.; AHO, A. V.; HIRSCHBERG, D. S. Bounds on the Complexity of the Longest Common Subsequence Problem. *Journal of the ACM (JACM)*, v. 23, n. 1, p. 1–12, jan. 1976.

YANG, K.-H. *et al.* Author Name Disambiguation for Citations Using Topic and Web Correlation. In: CHRISTENSEN-DALSGAARD, B. *et al.* (Org.). . *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science. [S.l.]: Springer Berlin / Heidelberg, 2008. v. 5173. p. 185–196.

ANEXO A - DESCRIÇÃO DOS ATRIBUTOS MODELADOS E ORIGEM DOS DADOS

Este anexo dicionariza cada um dos atributos que compõem as entidades modeladas. É informada também a origem dos dados que povoam cada um desses atributos, tendo como referência à Data Type Definition (DTD) do Currículo Lattes.

DICIONÁRIO DE ATRIBUTOS COM A ORIGEM DOS DADOS DE CARGA

Em: Jun/2014

Entidade: **Instituicao**

Atributo	Descrição	Elemento XML	Atributo XML	Observações
idInstituicao	Chave primária	Não se aplica	Não se aplica	
cdInstituicao	Código da Instituição	ENDERECO-PROFISSIONAL	CODIGO-INSTITUICAO-EMPRESA	
nmInstituicao	Nome da Instituição		NOME-INSTITUICAO-EMPRESA	

Entidade: **Unidade**

Atributo	Descrição	Elemento XML	Atributo XML	Observações
idUnidade	Chave primária	Não se aplica	Não se aplica	
cdUnidade	Código da Unidade	ENDERECO-PROFISSIONAL	CODIGO-UNIDADE ou CODIGO-ORGAO	
nmUnidade	Nome da Unidade		NOME-UNIDADE ou CODIGO-ORGAO	

Entidade: **CurriculumVitae**

Atributo	Descrição	Elemento XML	Atributo XML	Observações
idCurriculumVitae	Chave Primária	Não se aplica	Não se aplica	
nrIdentificador	Número do identificador Lattes	CURRICULO-VITAE	NUMERO-IDENTIFICADOR	
nmCompleto	Nome completo do titular	DADOS-GERAIS	NOME-COMPLETO	
nmCidadeNasc	Cidade de Nascimento do titular		CIDADE-NASCIMENTO	
nmPais	Nome do País de nascimento do titular		PAIS-DE-NASCIMENTO	
nrIdentidade	número de identidade do titular		NUMERO-IDENTIDADE	Não povoado
nmOrgaoEmissor	Nome do Órgão emissor da identidade		ORGAO-EMISSOR	
dtAtualizacao	Data da última atualização do CV	CURRICULO-VITAE	DATA-ATUALIZACAO	
nmPai	Nome do Pai do titular	DADOS-GERAIS	NOME-DO-PAI	Não povoado
nmMae	Nome da mãe do titular		NOME-DA-MAE	Não povoado
nrCpf	Número do CPF do titular		CPF	Não povoado

Entidade: **NomeCitacao**

Atributo	Descrição	Elemento XML	Atributo XML	Observações
idNomeCitacao	Chave primária			
nmCitacao	Nome de citação do titular	DADOS-GERAIS	NOME-EM-CITACOES-BIBLIOGRAFICAS	Passível de ser desmembrado em várias instâncias

Entidade: **ArtigoPublicado**

Atributo	Descrição	Elemento XML	Atributo XML	Observações
idArtigoPublicado	Chave Primária	Não se aplica	Não se aplica	
nmTitulo	Título do artigo publicado	DADOS-BASICOS-DO-ARTIGO	TITULO-DO-ARTIGO	
anPublicacao	Ano da publicação		ANO-DO-ARTIGO	
nmPaisPublic	Nome do país de publicação		PAIS-DE-PUBLICACAO	
idioma	Idioma de publicação		IDIOMA	
nmTituloPeriodico	Título do periódico de publicação	DETALHAMENTO-DO-ARTIGO	TITULO-DO-PERIODICO-OU-REVISTA	
nrIssn	Número do ISSN		ISSN	
nrDoi	Número do DOI	DADOS-BASICOS-DO-ARTIGO	DOI	
nrVolume	Número do volume de publicação	DETALHAMENTO-DO-ARTIGO	VOLUME	
nrFasciculo	Número do fascículo de publicação		FASCICULO	
nrSerie	Número da série de publicação		SERIE	
nrPagInicial	Número da página inicial de publicação		PAGINA-INICIAL	
nrPagFinal	Número da página final de publicação		PAGINA-FINAL	
nmLocalPublic	Local de publicação		LOCAL-DE-PUBLICACAO	

Entidade: **TrabalhoEvento**				
Atributo	**Descrição**	**Elemento XML**	**Atributo XML**	**Observações**
idTrabalhoEvento	Chave primária	Não se aplica	Não se aplica	
nmNatureza	Natureza do trabalho em evento	DADOS-BASICOS-DO-TRABALHO	NATUREZA	COMPLETO/RESUMO ou RESUMO_EXPANDIDO
nmTituloTrabalho	Título do trabalho em evento		TITULO-DO-TRABALHO	
nmClassificacaoEvento	Classificação do evento	DETALHAMENTO-DO-TRABALHO	CLASSIFICACAO-DO-EVENTO	INTERNACIONAL/ NACIONAL/ REGIONAL/ LOCAL/ NAO_INFORMADO
anTrabalho	Ano do trabalho em evento	DADOS-BASICOS-DO-TRABALHO	ANO-DO-TRABALHO	
nmPaisEvento	Nome do país de realização do evento		PAIS-DO-EVENTO	
nmIdioma	Idioma		IDIOMA	
nrDoi	Número do DOI		DOI	
nmEvento	Nome do evento	DETALHAMENTO-DO-TRABALHO	NOME-DO-EVENTO	
anRealizacao	Ano de realização do evento		ANO-DE-REALIZACAO	
nmTituloAnais	Título dos anais ou proceedings		TITULO-DOS-ANAIS-OU-PROCEEDINGS	
nrVolume	Número do volume referente ao trabalho		VOLUME	
nrFasciculo	Número do fascículo		FASCICULO	
nrSerie	Número de série		SERIE	
nrPagInicial	Número da página inicial		PAGINA-INICIAL	
nrPagFinal	Número da página final		PAGINA-FINAL	
nrIsbn	Número do ISBN		ISBN	
nmEditora	Nome da editora responsável pela publicação do trabalho em evento		NOME-DA-EDITORA	

Entidade: **AutorArtigo**				
Atributo	**Descrição**	**Elemento XML**	**Atributo XML**	**Observações**
idAutorArtigo	Chave primária	Não se aplica	Não se aplica	
nmAutor	Nome completo do autor do artigo	AUTORES	NOME-COMPLETO-DO-AUTOR	
nmCitacao	Nome de citação adotado no artigo		NOME-PARA-CITACAO	
nrCpf	Número do CPF do autor do artigo		CPF	Não povoado
nrOrdemAutoria	Número da ordem de autoria referente ao autor		ORDEM-DE-AUTORIA	
nrIdentCnpq	Número do identificador Lattes		NRO-ID-CNPQ	

Entidade: **AutorTrabEvento**				
Atributo	**Descrição**	**Elemento XML**	**Atributo XML**	**Observações**
idAutorTrabEvento	Chave primária			
nmAutorTrabEvento	Nome completo do autor do trabalho em evento	AUTORES	NOME-COMPLETO-DO-AUTOR	
nmCitacaoTrabEvento	Nome de citação adotado no trabalho		NOME-PARA-CITACAO	
nrCpfTrabEvento	Número do CPF do autor do trabalho		CPF	Não povoado
nrOrdemAutoria	Número da ordem de autoria referente ao autor		ORDEM-DE-AUTORIA	
nrIdentCnpq			NRO-ID-CNPQ	

ANEXO B - MATRIZ DE SENSIBILIDADE

Este anexo apresenta as matrizes de sensibilidade geradas para cada segmento considerado, classificadas pela coluna F1-Measure.

Análise de sensibilidade para os limites informados de similaridade - ARTIGOS PUBLICADOS								
Similaridade Artigo	**Similaridade Coautor**	**Verdadeiro Positivo**	**Falso Positivo**	**Verdadeiro Negativo**	**Falso Negativo**	**Precisão**	**Recall**	**F1-Measure**
0.55	0.40	349	6	1.889	27	0,983099	0,928191	0,954856
0.55	0.45	349	6	1.889	27	0,983099	0,928191	0,954856
0.55	**0.50**	**349**	**6**	**1.889**	**27**	**0,983099**	**0,928191**	**0,954856**
0.55	0.55	348	6	1.890	27	0,983051	0,928000	0,954733
0.50	0.40	349	7	1.888	27	0,980337	0,928191	0,953552
0.50	0.45	349	7	1.888	27	0,980337	0,928191	0,953552
0.50	0.50	349	7	1.888	27	0,980337	0,928191	0,953552
0.50	0.55	348	7	1.889	27	0,980282	0,928000	0,953425
0.45	0.40	349	9	1.886	27	0,974860	0,928191	0,950954
0.45	0.45	349	9	1.886	27	0,974860	0,928191	0,950954
0.45	0.50	349	9	1.886	27	0,974860	0,928191	0,950954
0.45	0.55	348	9	1.887	27	0,974790	0,928000	0,950820
0.60	0.60	342	6	1.890	33	0,982759	0,912000	0,946058
0.70	0.60	342	3	1.890	36	0,991304	0,904762	0,946058
0.75	0.60	342	3	1.890	36	0,991304	0,904762	0,946058
0.65	0.60	342	5	1.890	34	0,985591	0,909574	0,946058
0.65	0.65	337	5	1.890	39	0,985380	0,896277	0,938719
0.60	0.65	337	6	1.890	38	0,982507	0,898667	0,938719
0.70	0.65	337	3	1.890	41	0,991176	0,891534	0,938719
0.75	0.65	337	3	1.890	41	0,991176	0,891534	0,938719
0.65	0.70	336	5	1.890	40	0,985337	0,893617	0,937238
0.70	0.70	336	3	1.890	42	0,991150	0,888889	0,937238
0.75	0.70	336	3	1.890	42	0,991150	0,888889	0,937238
0.60	0.70	336	6	1.890	39	0,982456	0,896000	0,937238
0.40	0.55	348	20	1.876	27	0,945652	0,928000	0,936743
0.70	0.75	333	3	1.890	45	0,991071	0,880952	0,932773
0.75	0.75	333	3	1.890	45	0,991071	0,880952	0,932773
0.60	0.75	333	6	1.890	42	0,982301	0,888000	0,932773
0.65	0.75	333	5	1.890	43	0,985207	0,885638	0,932773
0.40	0.50	349	28	1.867	27	0,925729	0,928191	0,926959
0.40	0.45	349	34	1.861	27	0,911227	0,928191	0,919631
0.40	0.40	349	36	1.859	27	0,906494	0,928191	0,917214
0.80	0.80	317	0	1.890	64	1,000000	0,832021	0,908309
0.80	0.85	316	0	1.891	64	1,000000	0,831579	0,908046
0.80	0.90	316	0	1.891	64	1,000000	0,831579	0,908046
0.85	0.80	315	0	1.891	65	1,000000	0,828947	0,906475
0.85	0.85	314	0	1.891	66	1,000000	0,826316	0,904899
0.85	0.90	314	0	1.891	66	1,000000	0,826316	0,904899
0.90	0.80	314	0	1.891	66	1,000000	0,826316	0,904899
0.90	0.85	313	0	1.891	67	1,000000	0,823684	0,903319
0.90	0.90	313	0	1.891	67	1,000000	0,823684	0,903319

Análise de sensibilidade para os limites informados de sensibilidade - TRABALHOS EM EVENTOS								
Similaridade Trabalho	**Similaridade Coautor**	**Verdadeiro Positivo**	**Falso Positivo**	**Verdadeiro Negativo**	**Falso Negativo**	**Precisão**	**Recall**	**F1-Measure**
0.65	**0.55**	**1.192**	**96**	**6784**	**131**	**0,925466**	**0,900983**	**0,913060**
0.55	0.55	1.205	119	6.764	115	0,910121	0,912879	0,911498
0.60	0.55	1.194	108	6.773	128	0,917051	0,903177	0,910061
0.65	0.60	1.181	96	6.784	142	0,924824	0,892668	0,908462
0.50	0.50	1.219	148	6.738	98	0,891734	0,925588	0,908346
0.55	0.60	1.194	118	6.765	126	0,910061	0,904545	0,907295
0.50	0.45	1.220	154	6.731	98	0,887918	0,925645	0,906389
0.50	0.40	1.220	155	6.730	98	0,887273	0,925645	0,906053
0.60	0.60	1.183	107	6.774	139	0,917054	0,894856	0,905819
0.65	0.65	1.168	95	6.784	156	0,924782	0,882175	0,902976
0.55	0.65	1.181	116	6.766	140	0,910563	0,894020	0,902215
0.60	0.65	1.170	106	6.774	153	0,916928	0,884354	0,900346
0.70	0.70	1.154	89	6.786	174	0,928399	0,868976	0,897705
0.75	0.70	1.153	85	6.787	178	0,931341	0,866266	0,897626
0.80	0.70	1.147	85	6.787	184	0,931006	0,861758	0,895045
0.70	0.75	1.131	89	6.788	195	0,927049	0,852941	0,888452
0.75	0.75	1.130	85	6.788	199	0,930041	0,850263	0,888365
0.45	0.50	1.221	215	6.671	96	0,850279	0,927107	0,887032
0.80	0.75	1.124	85	6.789	205	0,929694	0,845749	0,885737
0.45	0.45	1.222	232	6.653	96	0,840440	0,927162	0,881674
0.45	0.40	1.222	245	6.640	96	0,832993	0,927162	0,877558
0.70	0.80	1.088	89	6.788	238	0,924384	0,820513	0,869357
0.75	0.80	1.087	85	6.789	242	0,927474	0,817908	0,869252
0.80	0.80	1.081	85	6.789	248	0,927101	0,813394	0,866533
0.85	0.80	1.071	80	6.790	262	0,930495	0,803451	0,862319
0.80	0.85	1.069	85	6.789	260	0,926343	0,804364	0,861055
0.80	0.90	1.067	85	6.789	262	0,926215	0,802859	0,860137
0.90	0.80	1.060	72	6.790	281	0,936396	0,790455	0,857258
0.85	0.85	1.059	80	6.790	274	0,929763	0,794449	0,856796
0.85	0.90	1.057	80	6.790	276	0,929639	0,792948	0,855870
0.90	0.85	1.048	72	6.790	293	0,935714	0,781506	0,851686
0.90	0.90	1.046	72	6.790	295	0,935599	0,780015	0,850752
0.40	0.50	1.221	381	6.505	96	0,762172	0,927107	0,836588
0.40	0.45	1.222	448	6.437	96	0,731737	0,927162	0,817938
0.40	0.40	1.222	512	6.373	96	0,704729	0,927162	0,800786

ANEXO C – ORGANOGRAMA DA UNIVERSIDADE FEDERAL FLUMINENSE

ANEXO D – TABELAS DE COMPOSIÇÃO DAS UNIDADES DE ENSINO

CAMPUS DO GRAGOATÁ
DEPARTAMENTO DE EDUCAÇÃO FÍSICA
GLE - INSTITUTO DE LETRAS
DEPARTAMENTO SOCIEDADE EDUCAÇÃO E CONHECIMENTO
DEPARTAMENTO DE LETRAS CLÁSSICAS E VERNÁCULAS- UFF
ESCOLA DE SERVIÇO SOCIAL
DEPARTAMENTO DE HISTÓRIA
INSTITUTO DE LETRAS/GCL
CENTRO DE ESTUDOS SOCIAIS APLICADOS
INSTITUTO DE CIÊNCIAS HUMANAS E FILOSOFIA
CENTRO DE ESTUDOS GERAIS
FACULDADE DE EDUCAÇÃO

CAMPUS DO VALONGUINHO
DEPARTAMENTO DE QUÍMICA ORGÂNICA
DEPARTAMENTO DE ODONTÉCNICA
DEPARTAMENTO DE BIOLOGIA MARINHA
INSTITUTO DE MATEMÁTICA - UFF
FACULDADE DE ODONTOLOGIA PUNF
DEPARTAMENTO DE NUTRIÇÃO SOCIAL
DEPARTAMENTO DE ESTATÍSTICA
FACULDADE DE ODONTOLOGIA/LABIOM-R
INSTITUTO DE BIOLOGIA
FACULDADE DE ADMINISTRAÇÃO E CIÊNCIAS CONTÁBEIS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
ESCOLA DE ENGENHARIA

UNIDADES ISOLADAS NITERÓI
DEPARTAMENTO DE SAÚDE E SOCIEDADE
DEPARTAMENTO DE PATOLOGIA E CLÍNICA - MCV
DEPARTAMENTO DE ARTE
DEPARTAMENTO DE MEDICINA CLÍNICA
ESCOLA DE ENFERMAGEM AURORA DE AFONSO COSTA
ESCOLA DE ENFERMAGEM
FACULDADE DE DIREITO
FACULDADE DE ECONOMIA
FACULDADE VETERINÁRIA
INSTITUTO BIOMÉDICO
INSTITUTO DE ARTES E COMUNICAÇÃO SOCIAL
DEPARTAMENTO DE MORFOLOGIA
INSTITUTO DE SAÚDE DA COMUNIDADE
FACULDADE DE FARMÁCIA

CAMPUS PRAIA VERMELHA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
DEPARTAMENTO DE URBANISMO
DEPARTAMENTO DE GEOQUÍMICA
ESCOLA DE ARQUITETURA E URBANISMO
INSTITUTO DE COMPUTAÇÃO
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
DEPARTAMENTO DE ENGENHARIA MECÂNICA
INSTITUTO DE GEOCIÊNCIAS
CENTRO TECNOLÓGICO
INSTITUTO DE FÍSICA

UNIDADES INTERIOR
POLO UNIVERSITÁRIO DE CAMPOS DOS GOYTACAZES
PÓLO UNIVERSITÁRIO DE VOLTA REDONDA ICEX
ESCOLA DE CIÊNCIA HUMANAS E SOCIAIS DE VOLTA REDONDA
INSTITUTO DE EDUCAÇÃO DE ANGRA DOS REIS
POLO UNIVERSITÁRIO DE NOVA FRIBURGO
INSTITUTO DO NOROESTE FLUMINENSE DE EDUCAÇÃO SUPERIOR
INSTITUTO DE CIÊNCIAS DA SOCIEDADE E DESENVOLVIMENTO REGIONAL
POLO UNIVERSITÁRIO DE RIO DAS OSTRAS
ESCOLA DE ENGENHARIA INDUSTRIAL METALÚRGICA DE VOLTA REDONDA
FACULDADE DE ODONTOLOGIA / PÓLO UNIVERSITÁRIO DE NOVA FRIBURGO

ÓRGÃOS SUPLEMENTARES
HOSPITAL UNIVERSITÁRIO ANTÔNIO PEDRO
CENTRO DE CIÊNCIAS MÉDICAS