UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

UMA PLATAFORMA PARA SUPORTE ÀS BIBLIOTECAS DIGITAIS DE EVENTOS CIENTÍFICOS COM FOCO NA EXTRAÇÃO DE INFORMAÇÃO

por

NEIDE FERREIRA ALVES

Tese submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco como parte dos requisitos para a obtenção do grau de Doutor em Engenharia Elétrica.

ORIENTADOR: Prof. Dr. RAFAEL DUEIRE LINS
CO-ORIENTADORA: Profª. Dra. MARIA LENCASTRE

Recife, Agosto de 2013.

Catalogação na fonte
Bibliotecária Margareth Malta, CRB-4 / 1198

A474p Alves, Neide Ferreira.

Uma plataforma para suporte às bibliotecas digitais de eventos científicos com foco na extração de informação / Neide Ferreira Alves. - Recife: O Autor, 2013.

xii, 121 folhas, il., gráfs.

Orientador: Prof. Dr. Rafael Dueire Lins.

Co-orientadora: Profa. Dra. Maria Lencastre.

Tese (Doutorado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia Elétrica, 2013.

Inclui Referências e Apêndices.

1. Engenharia Elétrica. 2. Biblioteca digital. 3. Extração de informação. 4. Processamento de documentos. 5. Processamento em nuvem. I. Lins, Rafael Dueire. (Orientador). II. Lencastre, Maria. (Co-orientadora). III. Título.

UFPE

621.3 CDD (22. ed.) BCTG/2014-051

Universidade Federal de Pernambuco
Pós-Graduação em Engenharia Elétrica

PARECER DA COMISSÃO EXAMINADORA DE DEFESA DE TESE DE DOUTORADO

NEIDE FERREIRA ALVES

TÍTULO

"UMA PLATAFORMA PARA SUPORTE ÀS BIBLIOTECAS DIGITAIS DE EVENTOS CIENTÍFICOS COM FOCO NA EXTRAÇÃO DE INFORMAÇÃO"

A comissão examinadora composta pelos professores: RAFAEL DUEIRE LINS, CIN/UFPE; VALDEMAR CARDOSO DA ROCHA JÚNIOR, DES/UFPE; FERNANDA MARIA RIBEIRO DE ALENCAR, DES/UFPE; MICKAËL COUSTATY, Universidade de La Rochelle; MARIA LENCASTRE PINHEIRO DE MENEZES CRUZ, SC/UPE; DENIS SILVA DA SILVEIRA, DCA/UFPE e DANIEL MARQUES OLIVEIRA, CHEMETCH/BRASIL, sob a presidência do primeiro, consideram a candidata **NEIDE FERREIRA ALVES APROVADA**.

Recife, 16 de agosto de 2013.

CECÍLIO JOSÉ LINS PIMENTEL
Coordenador do PPGEE

RAFAEL DUEIRE LNS
Orientador e Membro Titular Interno

FERNANDA MARIA RIBEIRO DE ALENCAR
Membro Titular Externo

MARIA LENCASTRE PINHEIRO DE MENEZES CRUZ
Co-Orientadora e Membro Titular Externo

DENIS SILVA DA SILVEIRA
Membro Titular Externo

VALDEMAR CARDOSO DA ROCHA JÚNIOR
Membro Titular Interno

DANIEL MARQUES OLIVEIRA
Membro Titular Externo

MICKAËL COUSTATY
Membro Titular Externo

Aos meus pais, Tacisio e Rosimeiry.

À minha avó, Petronilia.

E aos futuros doutores: Pedro Ariosto (meu amado filho),

Karen, Guilherme e Maria Heloisa (queridos sobrinhos).

AGRADECIMENTOS

A Deus, por iluminar o meu caminho.

Ao meu orientador, Rafael Dueire Lins, pela dedicação,paciência econhecimentos transmitidos sem medir esforços para atingirmos o objetivo.

À minha co-orientadora, Maria Lencastre, pelo companheirismo, encorajamento e compartilhamento do saber.

À UEA, pelo apoio, mas como poderia ter sido mais simples, sem tantas mudanças nas regras preestabelecidas.

Aos professores e colaboradores da UFPE, que tanto contribuíram com o Dinter.

À Fucapi, por ter me dado condições de iniciar esta jornada.

Aos colegas de curso, Ednelson, Ernande, Isaac, Jucimar, Raimundo, Ricardo, Rodrigo, Ruben Sicchar, Vitor e Walter, pelo incentivo e ajuda nos momentos mais difíceis.

Aos amigos da UFPE Lizandra e Gabriel França.

À minha família e em especial ao Pedro Ariosto que tanto tentou participar deste trabalho.

A todos que contribuíram direta e indiretamente para realização e conclusão desta tese, minha eterna gratidão.

No meio do caminho tinha uma pedra,
Tinha uma pedra no meio do caminho.
Não: no meio do caminho tinham perdas: Glória e Joaquim

No meio do caminho tinha uma pedra,
Tinha uma pedra no meio do caminho.
Não: no meio do caminho tinha uma gravidez.

No meio do caminho tinha uma pedra,
Tinha uma pedra no meio do caminho.
Não: no meio do caminho tinha o Pedro.
E Pedro ajudou a tirar as pedras do meio do caminho.

(Adaptado de *Carlos Drummond de Andrade*)

Resumo da Tese apresentada à UFPE como parte dos requisitos necessáriospara a obtenção do grau de Doutor em Engenharia Elétrica.

UMA PLATAFORMA PARA SUPORTE ÀS BIBLIOTECAS DIGITAIS DE EVENTOS CIENTÍFICOS COM FOCO NA EXTRAÇÃO DE INFORMAÇÃO

Neide Ferreira Alves

Agosto/2013

Orientador: Prof. Dr. Rafael Dueire Lins
Co-orientadora: Profª. Dra. Maria Lencastre
Área de Concentração: Telecomunicações
Palavras-chave:biblioteca digital, extração de informação, processamento de documentos,computação em nuvem.
Número de Páginas: 130.

A presente tese descreve as especificações e requisitos para o desenvolvimento de Bibliotecas Digitais de documentos textuais, considerando a possibilidade de reuso e a extração de dados. Considerando o imenso volume de informação disponível nesses repositórios, é de grande interesse a construção de sistemas capazes de selecionar automaticamente apenas os dados de interesse do usuário, facilitando assim o acesso, a manipulação e a divulgação dessas informações. O Modelo de Referências de Biblioteca Digital da DELOS foi utilizado para guiar a construção do ambiente, como consequência foi desenvolvida a plataforma pLiveMemory com módulos implementados para *desktop* e *web*, neste último, a infraestrutura da nuvem do Google é utilizada. Entre os módulos desenvolvidos há um específico para identificação e extração de referências bibliográficas, o qual usa, entre outros, o algoritmo de *Naïve Bayes* juntamente com as técnicas de expressões regulares. Também há um módulo para identificação de palavras-chave em arquivos de formato PDF editável. Os resultados obtidos mostraram os ganhos com a utilização das estratégias adotadas nas diversas fases do projeto, como na classificação automática de informação dos textos de artigos científicos.

Abstract of Thesis presented to UFPE as a partial fulfillment of therequirements for the degree of Doctor in Electrical Engineering.

A PLATFORM FOR SUPPORTING THE DEVELOPMENT OF DIGITAL LIBRARIES OF SCIENTIFIC EVENTS WITH FOCUS ON INFORMATION EXTRACTION

Neide Ferreira Alves

August/2013

Supervisors: Prof. Dr. Rafael Dueire Lins and Prof. Dr. Maria Lencastre
Area of Concentration: Telecommunications
Keywords:digital libraries, information extraction,document processing,cloud computing.
Number of Pages: 130.

This thesis describes the specifications and requirements for the development of digital libraries of textual documents, considering the possibility of reuse and data extraction. Considering the huge volume of information available in such repositories is of great interest to build systems that can automatically select only the data of interest to the user, thus facilitating access, manipulation and dissemination of such information. The Digital Library Reference Model of the DELOSwasused to guide the construction of the environment. As a result, we developed the platform pLiveMemory, which provides development modules for desktop and web. In the latter, the Google cloud infrastructure is used. Among the modules developed there is one specific for identification and extraction of bibliographic references, which uses, among others, the Naïve Bayes algorithm along with the techniques of regular expressions. There is also a module for identifying keywords in editable PDF files. The results showed gains with the use of the strategies adopted in the various phases of the project, as in the automatic classification of information from texts of scientific papers.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A popularidade do uso da *Internet* ressalta os avanços da humanidade na busca, disponibilização e alcance aos mais diferentes tipos de informação. O rompimento de barreiras antigas relacionadas à ausência de conhecimento (por dificuldade de acesso, custo, distribuição, ou proibição de manuseamento de documentos) tornou acapacidade de adquirir informação uma atividade mais democrática na sociedade atual.

Neste contexto, as bibliotecas digitais surgem como estruturadoras e facilitadoras de acesso ao conhecimento por diferentes comunidades. Em 1998, a *Digital Library Federation*(DLF) definiu bibliotecas digitais como:

> Organizações que disponibilizam os recursos, incluindo o pessoal especializado, para selecionar, estruturar, oferecer acesso intelectual, interpretar, distribuir, preservar a integridade e garantir a persistência ao longo do tempo de coleções de obras digitais; o objetivo é que estas estejam prontas e economicamente disponíveis para uso por uma comunidade ou um conjunto definido de comunidades (DLF, 1998).

As bibliotecas, conforme descrito por Barker *apud* Marchiori (1997), podem ser classificadas em diferentes tipos: polimídia, eletrônica, digital e virtual. As bibliotecas polimídia são similaresàs bibliotecas convencionais, porém incluem diferentes tipos de meiospara armazenagem da informação.A biblioteca eletrônica implica na utilização de um sistema de *software,* que inclui apoio à construção de índices *on-line*, busca de textos completos e na recuperação e armazenagem de registros, podendoenvolver a digitalização de livros.A biblioteca digital difere das demais porque a informação que ela contém existe apenas na forma digital, podendo residir em meios diferentes de armazenagem; ela não contém livros na forma convencional e a informação pode ser acessada tanto em locais específicos quanto remotamente. Já a biblioteca virtual depende da tecnologia da realidade virtual: um *software* próprio acoplado a um computador sofisticado reproduz o ambiente de uma biblioteca em duas ou três dimensões, criando um ambiente de imersão e interação.

Neste cenário, as bibliotecas digitais são um universo composto por uma estrutura complexa, que inclui o contínuo crescimento e evolução de abordagens, soluções e sistemas, levando à necessidade de estabelecer os fundamentos comuns, capazes de definir a base para uma melhor compreensão coletiva, comunicação e estímulo à evolução da área (DELOS, 2012). Vale ressaltar que o desenvolvimento em torno das bibliotecas digitais

proporcionou resultados suficientes para que padrões emergissem, permitindo o encapsulamento dos esforços realizados. Porém, estes sistemas são muito heterogêneos no espaço e funcionalidade, e sua evolução não segue um único caminho(DELOS, 2012).

A necessidade de uma representação completa encapsulando todas as perspectivas potenciais levou à elaboração do Manifesto da Biblioteca Digital (IFLA/UNESCO Manifesto *for Digital Libraries*). Estemanifesto explora a compreensão coletiva de bibliotecas digitais; ele tem como objetivo estabelecer as bases e identificar os conceitos fundamentais, dentro do universo das bibliotecas digitais, facilitando a integração da pesquisa e propondo as melhores formas de desenvolver sistemas adequadamente.

Lins *et al.* (2009) dividem as pesquisas voltadas para as bibliotecas digitais em três ramos, que possuem grandes áreas de intersecção e que estão classificadas de acordo com o tipo de dispositivo de digitalização utilizado (manual, automático e por meio de câmeras digitais). Assim, a classificação proposta pelos autores, com relação às pesquisas, consiste em:

- Documentos históricos, cujo formato ainda está em papel, exigem uma digitalização manual e criteriosa, uma vez que o envelhecimento do papel requer cuidados para evitar que estes sejam danificados ou destruídos.
- Documentos burocráticos, cujo processo de digitalização pode ser automatizado em lote, desde que o equipamento permita este processo e os documentos não estejam muito danificados.
- Documentos adquiridos por meio de câmeras digitais, que necessitam de processamento específico para a remoção de bordas, correção de distorções de iluminação, brochura, perspectiva etc.

Geralmente, a digitalização dos documentos garante um maior tempo de vida dos mesmos; desde que estes sejam adequadamente armazenados, o seu conteúdo terá uma maior capacidade de permanecer visível e com a mesma aparência por um maior período de tempo. A digitalização automática pode ser efetuada por *scanners* de linha de produção que são capazes de processar, atualmente, cerca de centenas de documentos por dia. Em tais documentos, a aplicação de técnicas de processamento digital de imagem (PDI) (GONZALEZ, 2000)pode possibilitar: uma melhor visualização e transmissão eficiente (via redes de computadores), indexação automática, transcrição semiautomática, compressão, redução no tamanho dos arquivos para armazenamento entre outros.

A partir de documentos digitalizados, a extração de informações pode ser automatizada facilitando processos de busca de conteúdo, criação de estatísticas, e a conexão entre documentos. No entanto, para o sucesso na extração automática de elementos, em documentos digitalizados, é essencial a identificação dos padrões nos quais estes documentos foram escritos. O conhecimento da formatação do documento facilita a identificação e classificação dos elementos existentes, favorecendo um correto tratamento dos mesmos. Neste contexto, existem diversos esforços da comunidade científica relacionados à modelagem, construção e extração de elementos em bibliotecas digitais. Le Bourgeois *et al*. (2007) trabalham com a digitalização e indexação de livros do período da renascença para a extração de dados. O trabalho de Álvarez (2007) extrai informações de artigos, por meio de etiquetagem semiautomática e classes gramaticais. Já Constans (2009) aborda extração de informações em artigos acadêmicos com o uso de expressões regulares[1]. Em Avancini *et al*. (2007) os autores abordam um ambiente colaborativo de biblioteca digital. Em Lins *et al*. (2009) é apresentado o ambiente LiveMemory, voltado para tratamento de imagens de artigos científicos; o projeto contempla as fases iniciais de digitalização até à aplicação de filtros,considerando questões relacionadas a:manuseio, qualidade e acesso de artigos científicos em papel. Este último trabalho, classificado como pesquisa em documentos burocráticos, serviu de motivação para a contextualização do problema a ser tratado nesta tese.

1.1 Motivação

Nos mais diversos congressos científicos, durante o século XX, foram gerados e publicadosmilhares de páginas de documentos, sendo que muitas destas páginas ainda estão, unicamente, em forma impressa. Os efeitos físicos do envelhecimento do papel (tais como, fungos, cupins e umidade) podem destruir as informações contidas nos documentos;os artigos podem desaparecer se não houver um gerenciamento e tratamentoapropriado dos mesmos, sem contar a necessidade de se manter a história sobre as conferências(LINS *et al*., 2009).

[1]Expressão Regular - é um método formal de se especificar um padrão sintático. É uma composição de símbolos, caracteres com funções especiais, que, agrupados entre si e com caracteres literais, formam uma sequência, uma expressão. Essa expressão é interpretada como uma regra, que indicará sucesso se uma entrada de dados qualquer casar com essa regra, ou seja, obedecer exatamente a todas as suas condições (JARGAS, 2009).

A digitalização, o tratamento, e a classificação desses documentos permitegarantir o seu maior tempo de vida, além de facilitar a divulgação e distribuição das suas informações, construídas ao longo de vários anos de pesquisa pela comunidade científica. Outro aspecto relevante é a viabilização de processos de busca e indexação dosconteúdos de artigos como, por exemplo, de referências entre artigos científicos.O uso de ferramentas para Reconhecimento de Padrões e Recuperação de Informação, juntamente com técnicas da Engenharia de Documentos, podem facilitar a estruturação, rotulação e classificação dos conteúdos dos artigos, aumentando a possibilidade de sucesso na correta extração de dados e, consequentemente, na disseminação do conhecimento científico presente em bibliotecas digitais.

1.2 Contextualização e Identificação do Problema

A dificuldade para manter, conservar, gerenciar e acessar documentos científicos, tanto em formato de papel quanto em formato digital, de forma integrada, está presente no dia a dia dos membros de diferentes instituições de pesquisa. Essa questão tem origem em diversas causas/razões relacionadas à complexidade do domínio.

AFigura 1contempla, de forma sintetizada, as principais razões que originam o problema relacionado à complexidade no gerenciamento e manutenção de bibliotecas digitais de documentos científicos.A figura contextualiza a questão de pesquisa desta tese, que é voltada para o estabelecimento de estratégias que viabilizem a construção deplataformas que efetivamente proporcionem a solução de problemas intrínsecos ao gerenciamento, acesso e distribuição, ao grande legado de conteúdos de artigos científicos, considerando diferentes representações dos mesmos: papel, imagens ePDF (*Portable Document Format* ou Formato de Documento Portátil), padrão criado pela empresa Adobe (2012).

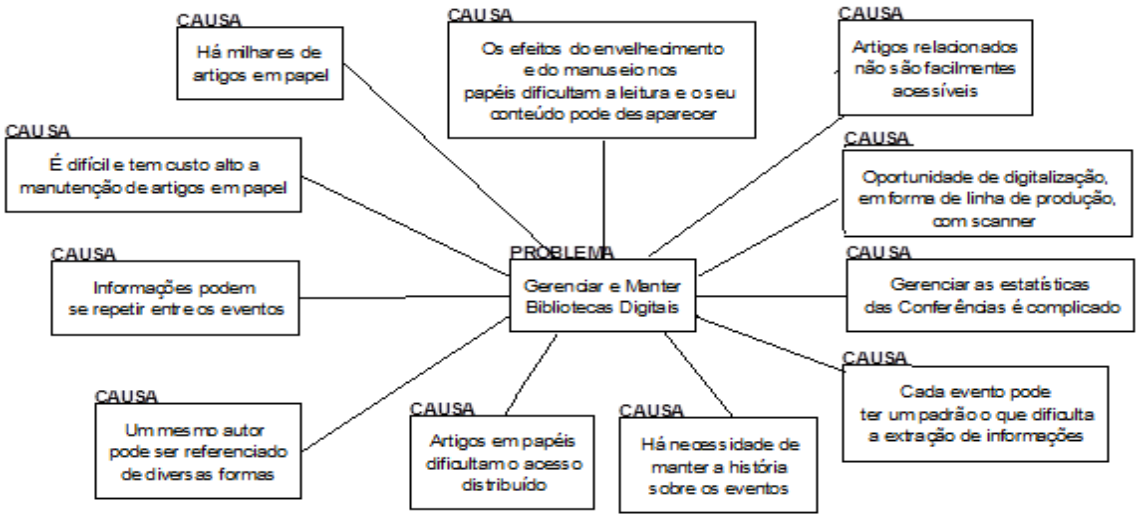


Figura 1 – *Diagrama Problema-Causas em Documentos de Conferências Científicas.*

Pode-se classificar o problema tratado nesta tese como direcionado à busca e proposição de soluções relacionadas a quatro questões básicas de suporte ao gerenciamento e manutenção de bibliotecas digitais,voltadas para documentos científicos, verFigura 2.

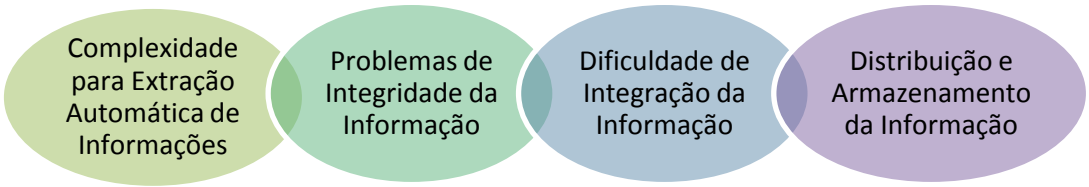


Figura 2 – *Subproblemas relacionados ao Gerenciamento e Manutenção de Bibliotecas Digitais.*

A extração de informações em artigos científicos, em especial, se depara com dificuldades inerentes à falta de aderência efetiva dos textos aos padrões estabelecidos pelas conferências. Entre as dificuldades tem-se: incompletude de informação (ausência de resumos ou palavras-chave, entre outros); referências com formatações diversas e incompletas; falta de integridade da informação, como por exemplo, mesmo autor ou título de artigo referenciado de diferentes maneiras. Outro complicador é a diversificação de padrões entre os eventos científicos, podendo haver para mesma conferência edições com formatos diferentes.

Existem ainda problemas relacionados àintegridade da informação (ex: mesmo autor referenciado dediferentes formas); integração de artigos científicos (ex: estabelecimento de ligações entre referências e artigos, e entre eventos); além de questões relativas à distribuição e armazenamento de grande volume de informação.

Até o momento não foi identificada solução para o problema de extração de dados no contexto das bibliotecas digitais de artigos científicos, com tratamento de imagens para osarquivos, bem como para arquivo no formato PDF. Apenas se encontrou trabalhos que se aproximam propondo uma abordagem para extração em contextos específicos; ou ambientes para bibliotecas digitais, que partem do pressuposto que as informações dos artigos já estão em um formato pronto para ser disponibilizado, ou seja, não lidam com as complexidades e necessidades identificadas.

Assim, a questão de pesquisa desta tese concentra-se na definição de estratégias para a extração automática de informações em artigos científicos. Considera-se que, a partir de uma boa solução de extração, pode-se mais facilmente garantir a integridade de informação, a integração de informações, assim como o correto armazenamento das mesmas.

1.3 Objetivos da Tese

Os objetivos da tese estão descritos inicialmente por meio do seu objetivo geral, sendo detalhados em seguida seus objetivos específicos.

1.3.1 Geral

Proposta de uma plataforma, apoiada por um sistema, para criar e gerenciar bibliotecas digitais integradas a partir de documentos científicos não digitalizados, com foco nos problemas de extração de informação.

1.3.2 Específicos

- Definir uma arquitetura geral e procedimentos específicos, relacionados a questões ainda em aberto na extração de informação de artigos de documentos, para montar uma biblioteca digital de artigos científicos;
- Considerar padrões existentes relacionados a bibliotecas digitais, visando garantir uma contextualização com relação a padrões já utilizados.
- Desenvolver um ambiente para digitalização, transcrição e extração de dados de artigos científicos;
- Desenvolver um ambiente *web* para disponibilização distribuída de dados, considerando o grande volume de informação armazenado.

Assim, esta tese propõe uma solução para os problemas abordados, usando extração e classificação de dados, uso de modelo de referência para geração de bibliotecas digitais, além da infraestrutura da nuvem para divulgar e disponibilizar os dados. A abordagem definida e adotada mostrou-se viável, principalmente na extração, pois os padrões utilizados alcançaram resultados satisfatórios, com índices de mais de 70% de acertos, utilizando cobertura como medida.

1.4 Principais Contribuições

Para abordar as questões de extração em documentos científicos, ainda não resolvidas na literatura, foram exploradas técnicas para identificação de padrões nos documentos para posterior classificação. Um vetor de características foi criado, os documentos foram selecionados para uma primeira etapa de treino e em seguida para a etapa de teste. Os documentos foram classificados com o uso de Expressões Regulares, em um primeiro momento, e em seguida técnicas de classificação automática, como o algoritmo K-NN, juntamente com Distância Euclidiana e Similaridade do Cosseno e por fim foi utilizado o algoritmo *Naïves Bayes*. O título e a introdução dos artigos foram utilizadas para extração de palavras-chave, dos artigos que não as possuíam.

Apesar do foco da tese ser em extração de informações em documentos científicos, o trabalho considerou estratégias para resolução de diferentes problemas,de forma integrada,de acordo com a Tabela 1.

Tabela 1 – *Problemas Relacionadose Soluções.*

Problema	**Solução**
Complexidade na extração das informações.Dificuldade para identificar automaticamente os elementos de busca, devido aos diversos padrões de formatação em eventos.	Definição de padrões com o uso de expressões regulares e criação de um vetor de características.
Artigos em papel sofrem desgastescom o tempo, e mesmo os digitais (em mídias como CD/DVD) têm uma vida útil de 5 a 10 anos. O custo de manutenção em papelé alto.	Uso de técnicas baseadas em LINS*et al*. (2009) de forma integrada.
Falta de padronização para o gerenciamento dos arquivos em formato de papel/digital, e de seus relacionamentos.	Uso da abordagem do LiveMemory (digitalização/filtragem) para documentos em papel, definição de base de dados para informações extraídas e armazenamento em nuvem.

Problema	Solução
Cada evento requer a extraçãoespecífica de suas informações que permita uma posterior análise e geração de relatórios estatísticos.	Extração e classificação de dados específicos para busca em base de dados usando diferentes elementos dos artigos (título, autor, ano de publicação, local etc.).
Dificuldade em manter as informações consistentes.	Identificação de informações reutilizadas entre os artigos.
Garantia de consistência.	Armazenamento em base de dados.
Artigos em diversos idiomas.	Aplicação de algoritmo para identificar idiomas.

Como apoio às estratégias desenvolvidas nesta tese,para extração e geração de uma plataforma - denominada pLiveMemory -foram consideradasduas abordagens principaisde embasamento, verFigura 3:

- Digitalização e processamento de imagem, ondeforam adotadas e reimplementadas algumas soluções do LiveMemory (Lins*et al.*,2009), descritas na seção 2.1, para a criação de umaplataforma integrada;
- Padronização de Bibliotecas Digitais, como as propostas pela *Network of Excellence for Digital Libraries* (DELOS).

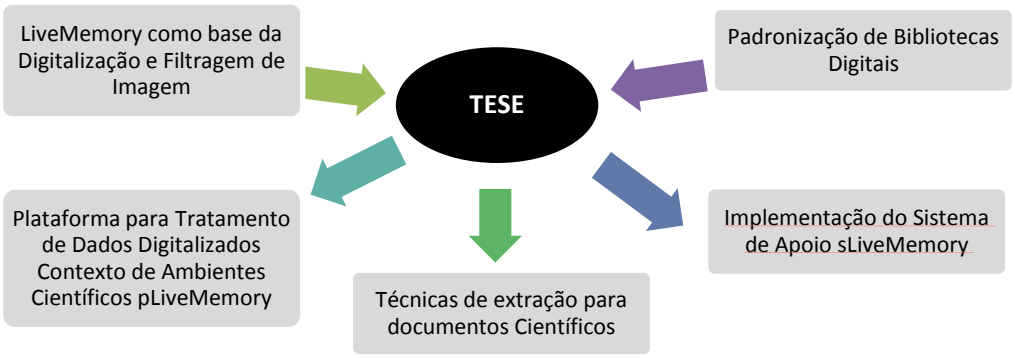


Figura 3 – *Base para geração da Plataforma pLiveMemory.*

NaFigura 4 é apresentado um esquema da solução proposta.Este esquema contempla a seleção dos documentos a serem tratados, arquivos em papel ou no formato PDF (i); os documentos em papel são digitalizados (ii); gerando, deste modo, arquivos no formato de imagem (iii);estes passam por um OCR (iv); para geração do arquivo em formato TXT (v), da mesma forma os documentos, em formato PDF (vi), são enviados ao

software PDFBox (vii) para geração do arquivo TXT (v); deste arquivo são extraídas as informações a serem armazenadas na base de dados (viii); os arquivos PDF são enviados ao Google *Drive*©(ix) e por fim as informações são disponibilizadas, via web, no Google *Sites*©(x). No capítulo 3, este esquema será detalhado com o acréscimo de outras etapas.

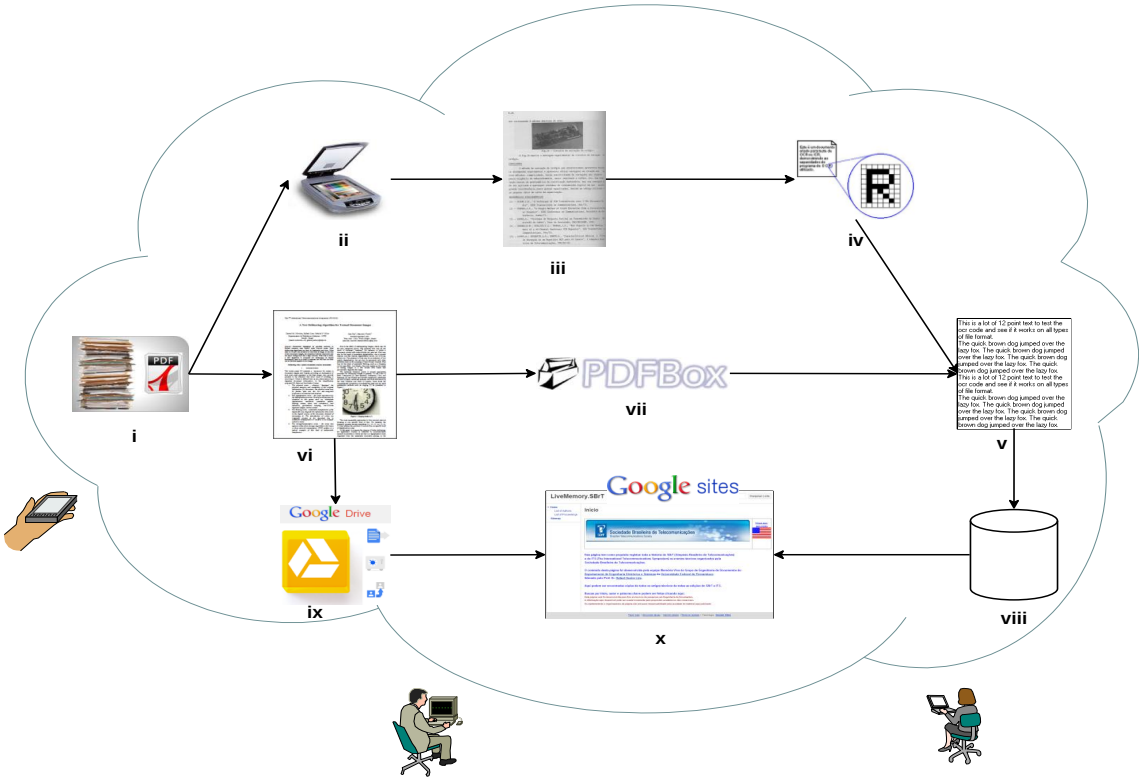


Figura 4 – *Base para geração da Plataforma pLiveMemory.*

1.5 Procedimentos Metodológicos

Este trabalho foi baseado no tipo de pesquisa bibliográfica e experimental, uma vez que para alcançar os seus objetivosfoi necessário um bom embasamento teórico, para posteriormente serem iniciados os respectivos experimentos. A abordagem foi quantitativa, sendo necessário mensurar os resultados obtidos. O método adotado foi o indutivo, pois este foi realizado em três etapas: observação dos fenômenos, descoberta da relação entre eles e generalização da relação, de acordo com Lakatos (2003).

As etapas da pesquisa foram divididas conforme descrição a seguir:

- Proposta inicial da metodologia para tratamento de bibliotecas digitais;
- Definição de estratégia para extração de informações;

- Instanciação da plataforma proposta a partir do Modelo de Referências da DELOS;
- Avaliação do reuso na montagem de novas bibliotecas digitais;
- Implementação dos módulos de processamento de imagem e extração de informação;
- Desenvolvimento de estratégia e módulo para identificar palavras-chave;
- Carga e teste de aproximadamente 3000 artigos;
- Disponibilização da biblioteca digital em nuvem (*web*).

Conforme pode ser visto naFigura 5, a estratégia metodológica aplicada para validação da solução proposta foi dividida em três fases principais. A primeira fase foi voltada para a reimplementação de técnicas de processamento e compressão de imagens, gerando documentos monocromáticos de boa qualidade a partir de documentos digitalizados em grande volume. Na segunda fase, os dados foram extraídos e armazenados em um banco de dados. Houve o foco no tratamento de problemas relacionados à extração de informações dos artigos, em formato PDF ou de imagem. Na terceira fase, foi feito o envio e disponibilização dos dados para a nuvem, permitindo um armazenamento mais independente, e via *web*.

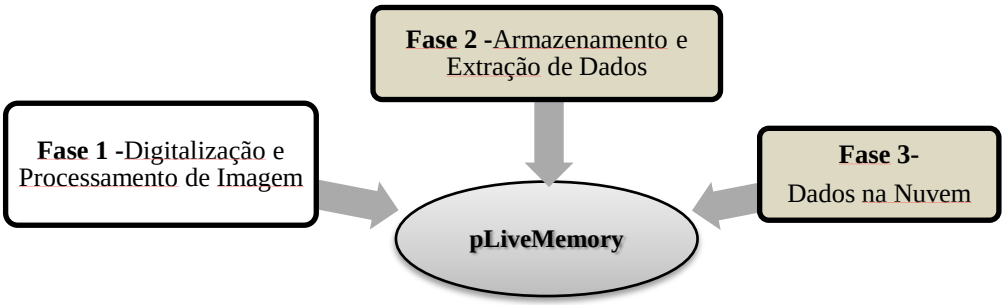


Figura 5 – *Fases do pLiveMemory.*

Os procedimentos adotados consideraram a implementação dos ambientes no Microsoft© Visual C#, os dados armazenados no banco de dados MySql (2013) e divulgados na *web*, por meio da nuvem do Google©. Os resultados foram analisados com o intuito de propor melhorias na classificação de documentos, por fim os resultados alcançados foram publicados em congressos de renome.

1.6 Estrutura desta Tese

Além deste capítulo introdutório,esta teseé compostade cinco outros capítulos, conforme detalhamento feito a seguir:

- Capítulo 2 – Trabalhos Relacionados: apresenta trabalhos relacionados a ambientes para tratamento de imagens de documentos, bibliotecas digitais, extração de dados e modelos de referências;
- Capítulo 3 – Geração de Biblioteca Digital para Anais de Eventos Científicos: apresenta o modelo proposto para a construção da PlataformapLiveMemory;
- Capítulo 4 – PlataformapLiveMemory: descreve os ambientes implementados para *desktop*, como também os métodos utilizados para reconhecimento e extração de informações, como referências e palavras-chave, do título e/ou introdução de artigos científicos;
- Capítulo 5 – Plataforma pLiveMemory na Nuvem: descreve o ambiente desenvolvido para*web*, focando nas características do ambiente na nuvem do Google©;
- Capítulo 6 – Conclusão: apresenta uma análise crítica relativa aos resultados esperados e aos obtidos no experimento, destacando as principais contribuições, assim como menciona as possibilidades de continuação deste trabalho, apresentando ideias de expansão, melhoria ou mesmo correção de falhas nos métodos utilizados.

2 TRABALHOS RELACIONADOS

Este capítulo apresenta trabalhos relacionados aos principais assuntos abordados nesta tese. Na seção 2.1 são apresentadas6 abordagens relacionadas ao apoio no desenvolvimento de Bibliotecas Digitais. Em seguida, na seção 2.2, são apresentados 5 trabalhos sobre Extração de Dados. Por fim, na seção 2.3é descrito o Modelo de Referência para Bibliotecas Digitais da DELOS, usado como base para a abordagem proposta nesta tese.

2.1 Abordagens de apoio às Bibliotecas Digitais

Com relação às Bibliotecas Digitais são apresentadasabordagens que apoiam, de diferentes formas, a construção destes ambientes. Entre elas tem-se: LiveMemory(LINS *et al*., 2009), Ambiente Cyclades(AVANCINI *et al*., 2007), *Framework*[2]GARDI(GURRIN *et al*., 2009), Biblioteca DML-C(SOJKA *et al*., 2010), Ferramenta DEBORA(LE BOURGEOIS *et al*., 2007) e *Framework* dLibra(MAZUREK *et al*., 2005).

A Tabela 2 apresenta uma síntese do principal foco e características de cada uma, destacando aquelas que são *frameworks* e as que possuem uma etapa para digitalizar documentos ou extrair dados.

Tabela 2 – *Síntese de Abordagens para apoio a Bibliotecas Digitais.*

Bibliotecas Digitais	Principal Foco e Características	*Framework*	Digitalização	Extração de Dados
LiveMemory (LINS *et al*., 2009)	Tratamento de imagens de artigos científicos da SBrT	N	S	N
Cyclades (AVANCINI *et al*., 2007)	Biblioteca digital com ambiente colaborativo	N	N	N
GARDI (GURRIN*et al*., 2009)	Memórias digitais humanas	S	N	N
DML-CZ (SOJKA*et al*., 2010)	Publicações matemáticas	N	S	S
DEBORA (LE BOURGEOIS*et al*., 2007)	Livros da renascença	N	S	S
dLibra (MAZUREK*et al*., 2005)	Biblioteca digital polonesa: herança cultural, materiais regionais, materiais educativos e notas musicais	S	N	N

[2]*Framework* é um esqueleto de um sistema que pode ser personalizado para uma determinada aplicação. (GAMMA*et al.*, 1995).

2.1.1 LiveMemory - Abordagens para Tratamento de Imagem de Documentos

O LiveMemory é um ambiente de gerenciamentode digitalização de documentos, que permite o armazenamento em CD/DVDe utiliza módulos de tratamentode imagem para melhorar as imperfeições da fase digitalização.

Com o objetivo de manter a história da Sociedade Brasileira de Telecomunicações (SBrT) foi desenvolvido, em meados de 2007, o Projeto LiveMemory, cujo objetivo contemplou a disponibilização, em mídia, de todo o acervo publicado nos congressos da SBrT e do *International Telecommunications Symposium* (ITS), uma publicação internacional da SBrT, que acontece a cada 4 anos (LINS *et al.*, 2009).

Os documentos mais antigos da SBrT foram digitalizados (LINS *et al*, 2009) e passaram por tratamentos para extração de imperfeições oriundas do processo de digitalização. Neste caso a ferramenta BigBatch (descrita a seguir) foi utilizada por meio de técnicas de processamento e compressão de imagens, as quais geraram documentos monocromáticos a partir dos documentos digitalizados, em grande volume. Após a digitalização, os artigos foram disponibilizados no formato PDF editável, ou seja, os arquivos não eram de imagens e desta forma os softwares de leitura de PDF, poderiam lê-los. Em 2008,foi feito o lançamento de um DVD com os 25 anos de história dos anais da SBrT (LINS *et al*., 2009).

Os artigos foram digitalizados com uma resolução de 200 dpi e armazenados em formato não comprimido (BMP). As imagens digitalizadas foramguardadas em um diretório que corresponde ao ano do evento. Ressalta-se que o LiveMemory é destinadoa usuários não especialistas em processamento de imagens, assim, a parte de processamento de imagem não pede nenhum parâmetro de entrada.

A maioria dos artigos tratados no LiveMemory foram impressos em pretoebranco e muitas vezes incorporavam elementos gráficos como fotos, figuras e gráficos que são impressos usando técnicas de *dithering*[3]. As imagens em preto e branco ou monocromáticas ocupam menos espaço do que as coloridas, seu carregamento para visualização também é mais rápido e elas precisam de menos tinta para impressão.Deste modo, optou-se por padronizar as imagens para este formato.

A ferramenta BigBatch, usada no LiveMemory,foi projetada para processar, automaticamente, milhares de imagens monocromáticas de documentos gerados por

[3]*Dithering* - Técnica que sobrepõe duas imagens com cores diferentes para dar a impressão de que é apenas uma imagem, simulando por meio da sobreposição uma terceira cor.

scanners de linha. Entre suas funções estão a remoção de bordas ruidosas, checagem e correção da orientação do papel, cálculo e compensação do ângulo de inclinação, corte de imagem para padronização dos tamanhos dos documentos e compressão, de acordo com o formato de arquivo definido pelo usuário (LINS *et al*, 2006).A ferramenta pode funcionar tanto em modo autônomo quanto em modo assistido pelo operador. No modo autônomo, ela é capaz de processar em *clusters* de estações de trabalho ou em *grids* e para estes um lote completo de documentos, colocado em um diretório, é processado sequencialmente. O modo assistido pelo operador permite que este utilize ferramentas de processamento de imagens para aplicar nas imagens dos documentos.

A Figura 6 exemplifica a remoção de borda e correção de inclinação das imagens de documentos (a) e (b), os resultados estão, respectivamente, nas imagens (c) e (d), na primeira a inclinação foi corrigida e as bordas removidas e na segunda a borda também foi excluída e a imagem rotacionada em 90º para esquerda.

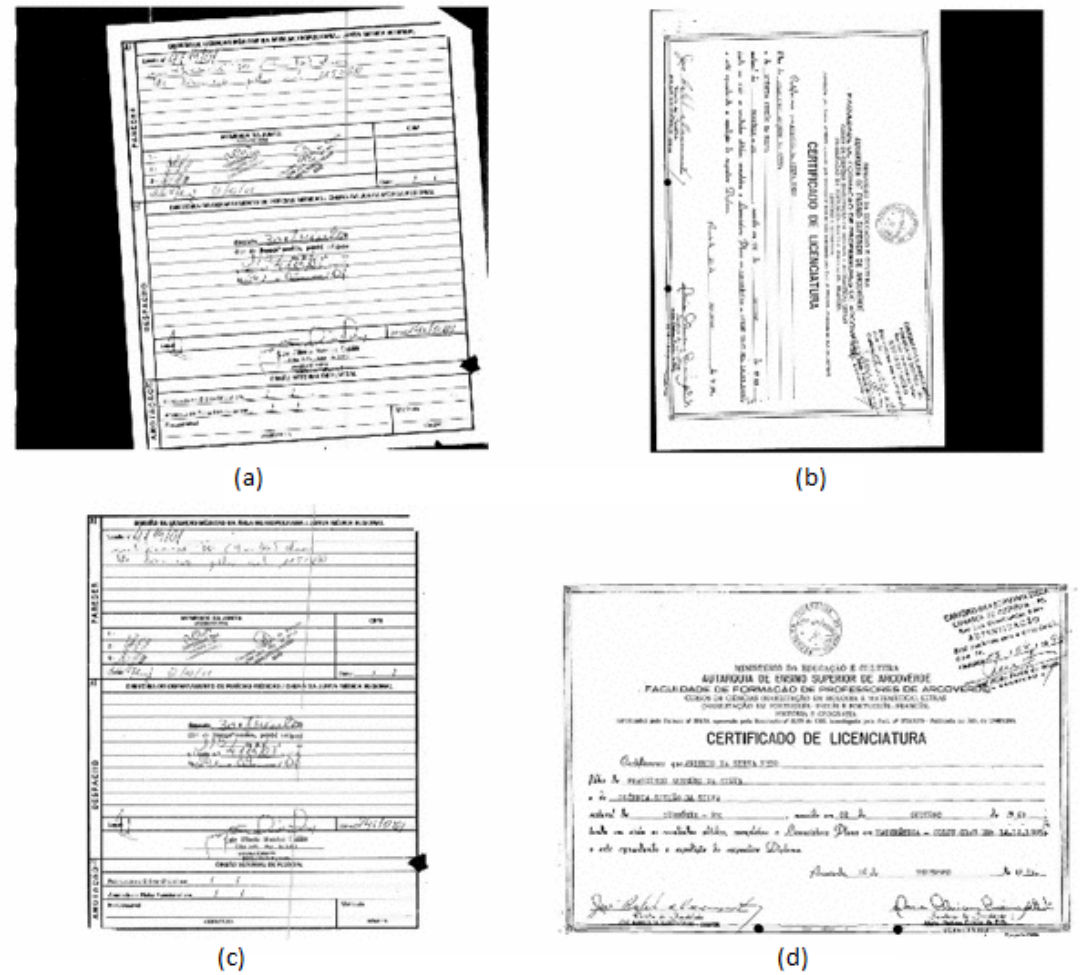


Figura 6 – *Exemplo de Processamento de Imagens de Documento com o BigBatch.*

Como LiveMemory faz uso de algumas das funcionalidades do BigBatch, a sua interface de processamento de imagem pode funcionar nos modos usuário ou lote, como já mencionado. A Figura 7apresenta uma captura de tela do LiveMemory trabalhando em modo usuário. Este modo permite ao usuário aplicar filtros, tais como: binarização, remoção de borda, ajuste de rotação, remoção de *pixels* isolados (ruído sal e pimenta), nas imagens dos documentos. Vale ressaltar que alguns filtros só funcionam nas imagens em binário (preto e branco). O usuário também pode retornar a imagem processada para a imagem original ou levá-la para livremente processar a imagem na ferramenta *ImageJ*, programa de processamento de imagem em Java, (ImageJ, 2012).

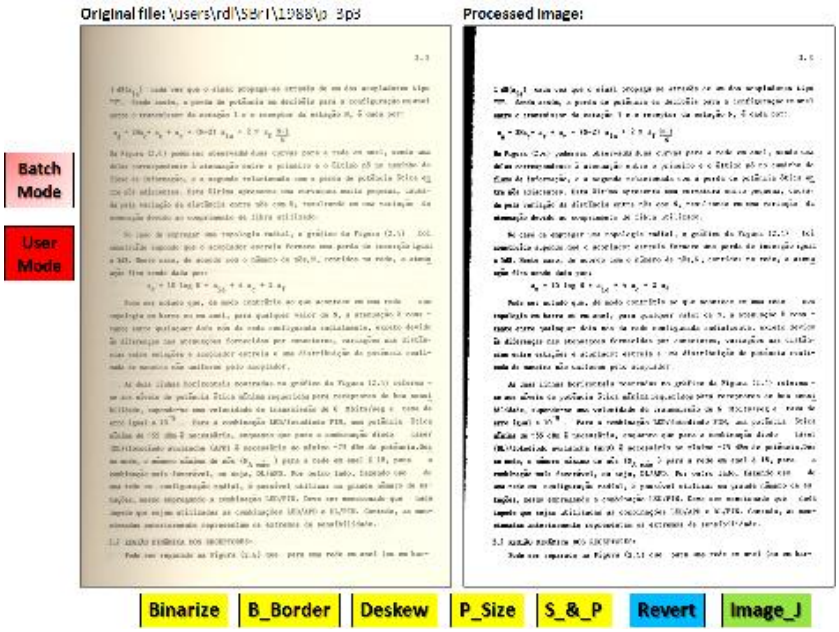


Figura 7 – *Interface de Processamento de Imagem do LiveMemory.*

2.1.2 Ambiente Cyclades

CYCLADES (AVANCINI*et al.*, 2007) é um ambiente de biblioteca digital colaborativo que disponibiliza ao usuário funcionalidades organizadas em 4 categorias: (i) busca de informação; (ii) organização do espaço de informação; (iii) colaboração com outros usuários que compartilham interesses semelhantes; e (iv) obtenção de recomendações.

Neste ambiente de colaboração, com espaço para reuniões, os usuários podem tirar proveito do conhecimento uns dos outros por meio da partilha de informações, opiniões e experiências. Além disso, o usuário pode organizar o seu espaço de informação de acordo

com seu gosto pessoal e uso;a partir dessas informaçõeso sistema busca inferir propriedades sobre os interesses dos usuários, as relações entre usuários e comunidades de usuários, bem como fazer recomendações, com base em padrões de preferência. Os experimentos realizados pelos autores nesta plataforma focaram no módulo de recomendação, e mostraram a eficácia dos algoritmos adotados neste ambiente. Vale ressaltar que o ambiente Cyclades foi criado usando o *framework* fornecido pela DELOS(CANDELA *et al.*, 2008).

O objetivo da Cyclades é fornecer um ambiente integrado para os usuários e grupos de usuários (comunidades) que desejam utilizar, de forma personalizada e flexível, arquivos abertos, ou seja, arquivos eletrônicos de documentos compatíveis com o padrão na *Open Archives Initiative* (OAI[4]). OAI é um acordo entre os vários prestadores de arquivos digitais, a fim de fornecer um nível mínimo de interoperabilidade. Cyclades permite o acesso aos metadados fornecidos por estes arquivos, pois reúne esses registros e por meio deles fornece acesso aos documentos completos (se existirem e se seu acesso for permitido). Sobre eles, Cyclades atua como um provedor de serviços OAI, conforme Figura 8.

[4] OAI - tem por objetivo desenvolver e promover padrões de interoperabilidade que visam facilitar a disseminação de conteúdo. A OAI desenvolveu um código partilhado para *tags* de metadados. Os textos completos dos documentos podem estar em diferentes formatos e localizações, mas se usarem as mesmas *tags* de metadados tornam-se interoperáveis. Os seus metadados podem ser colhidos e todos os documentos podem ser procurados e recuperados como se estivessem em uma única base de dados (OAI, 2012).

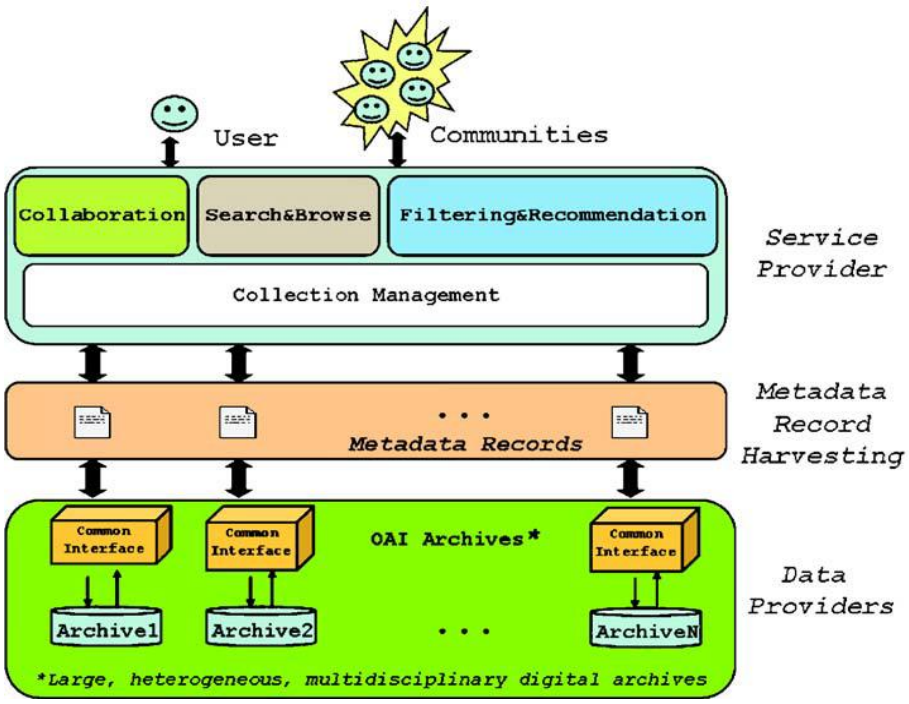


Figura 8 – *Visão Lógica das Funcionalidades de Cyclades. (AVANCINI et al., 2007)*

2.1.3 Framework GARDI

Gurrin em um dos seus trabalhos (Gurrin *et al*., 2009) apresenta um *framework* para biblioteca digital, denominado GARDI, cujo domínio são as memórias digitais humanas; estas contemplam inúmeras fotos, vídeos e áudios que precisam de um gerenciamento complexo, principalmente pela diversidade do armazenamento (celular, câmeras, *tablet*, computadores etc.), durante todo o seu ciclo de vida. Por intermédio de um conjunto de regras de transformação, o GARDI automaticamente faz a otimização do armazenamento e a recuperação de grandes quantidades de dados digitais pessoais e heterogêneos.

Os objetos digitais, ao invés de serem mantidos em estado fixo, da publicação até à eliminação, são postos em um ciclo de estado completo. O objeto digital pode ser modificado (transformado) por meio de um certo número de ciclos. Um bom exemplo é um vídeo CCTV (*Closed-Circuit Television*), com dados redundantes, sem movimento, que podem ser removidos e consequentemente reduzidos para uma sequência de imagens mais relevantes, ou mesmo reduzidos aos metadados que identificam os eventos de segurança mais relevantes. Assim, uma hora de vídeo pode ser substituída por um número de objetos menores, que exigem uma menor capacidade de armazenamento, mas mantêm o mesmo valor semântico.

A partir da Figura 9, observa-se que no *framework* GARDI os objetos digitais são capturados e importados, indicado em (1), para o armazenamento (2), onde são publicados

na biblioteca digital. Enquanto estão no estado de publicação, os objetos são observados (3), podendo ser desencadeado, por transformação, e transformados pelo mecanismo de transformação (4), utilizando um conjunto de regras de transformação (5), ao final são apresentados ao usuário (6). Normalmente, as regras de transformação devem ser configuradas para terem um impacto mínimo sobre a semântica dos dados, uma vez que essas transformações vão ser persistentes, ou seja, terão um efeito permanente sobre os objetos digitais.

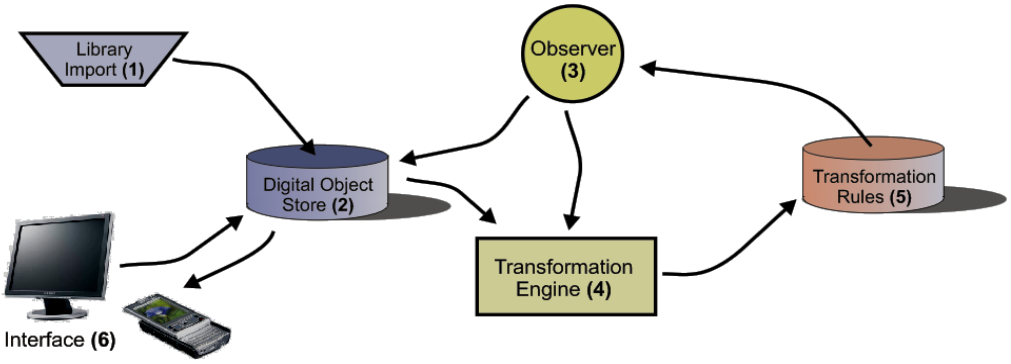


Figura 9 – *Visão de Alto Nível do Framework GARDI. (GURRIN et al., 2009)*

2.1.4 Biblioteca DML-CZ

O trabalho de (SOJKA*et al.*, 2010) relata as experiências adquiridas, as lições aprendidas e as ferramentas desenvolvidas para a Biblioteca Digital Tcheca de Matemática (*The Czech Digital Mathematical Library* – DML-CZ), a qual consiste na digitalização das publicações matemáticas da República Tcheca. Atualmente, as ferramentas estão sendo adaptadas para a Biblioteca Digital Europeia de Matemática (*The European Digital Mathematics Library* – EUDML).

O objetivo do projeto DML-CZ é a digitalização da literatura matemática, mais relevante, publicada em terras checas. A biblioteca é composta por periódicos, monografias e anais de conferências do século XIX e publicações recentes, o acervo disponibiliza ao público quase 30.000 artigos ou 300.000 páginas.O foco do trabalho são arquivos em formato PDF, pois para este foi desenvolvido uma ferramenta de recompressão. O projeto descreve abordagens e ferramentas para criação de bibliotecas digitais, tais como: assinatura digital, métricas de similaridade entre documentos, validação de dados e algumas ferramentas de OCR para símbolos matemáticos.

No caso dos arquivos de entrada estarem em papel, o trabalho propõe a sua digitalização e aplicação do processo de identificação de símbolos matemáticos via OCR; o texto gerado é então transformado para PDF, e em seguida há uma etapa de gerenciamento das informações contidas nos arquivos. O *workflow* exibido naFigura 10 mostra os vários tipos de dados de entrada suportados pela ferramenta: impresso, tiff, pdf, ps, txt. Ao final do processo os usuários podem acessar a biblioteca digital.

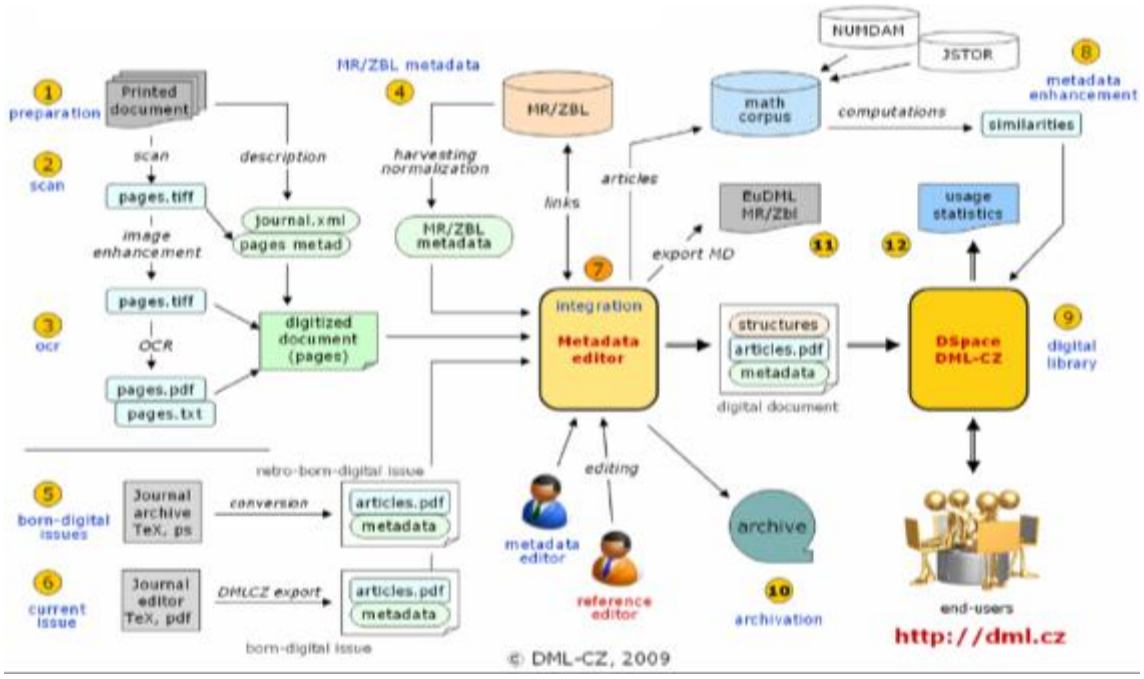


Figura 10 – *Workflow de nível superior do DML-CZ. (SOJKA et al., 2010)*

2.1.5 Ferramenta DEBORA

Em (LE BOURGEOIS *et al*., 2007) os autores descrevem a ferramenta DEBORA (*Digital accEss to BOoks of the RenAissance*), a qual trabalha com a digitalização e indexação de livros do período da renascença, permitindo assim aos usuários acesso aos livros raros do século XVI.

A fase de preparação para extração das informações é o que há de mais relevante neste trabalho. Como há inúmeros objetos decorativos nas páginas dos livros, o processo de segmentação e extração de conteúdo exige o uso de várias técnicas, como a decomposição da imagem em dois níveis: (1) primeiro plano com os caracteres e (2) o plano de fundo, a folha do papel.

Há também o processo de retroconversão de documentos históricos, o qual consiste na conversão de imagens originais para um formulário eletrônico reutilizável, adequado às necessidades dos usuários que inclui: captura de imagem, extração de metadados,

armazenamento e indexação de imagens, conversão automática em formulário eletrônico reutilizável, publicação na Internet, e compressão de dados para acesso mais rápido.

O esquema geral do sistema DEBORA é descrito na Figura 11.Neste é possível observar que as imagens passam por um processo de transcrição que, junto com os metadados, servirão de base para as consultas.

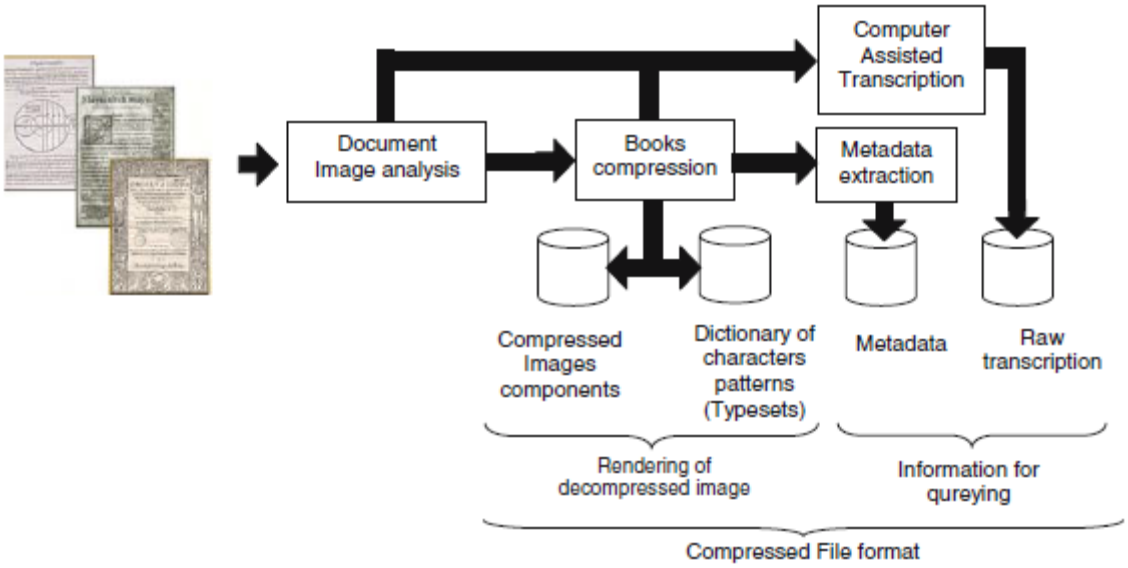


Figura 11 – *Visão geral do esquema DEBORA (LE BOURGEOIS et al., 2007).*

2.1.6 Framework dLibra

Em (MAZUREK*et al.*, 2005) os autores apresentam o *framework* dLibra, que originou a primeira Biblioteca Digital Polonesa. Neste trabalho é abordado o ciclo de vida de objetos digitais, baseando-se em experiências da biblioteca digital WBC (*Wielkopolska Biblioteka Cyfrowa*). O ambiente possui mais de 4000 publicações agrupadas em quatro coleções: herança cultural, materiais regionais, materiais educativos e notas de música. O intuito é mostrar como um ciclo de vida bem construído pode estender as funcionalidades das bibliotecas digitais, tanto para leitores de bibliotecas quanto para bibliotecários, os quais são os responsáveis pela manutenção do conteúdo.

dLibra é uma estrutura de biblioteca digital distribuída e portátil, criada para ser configurável com uma arquitetura em várias camadas. No centro da arquitetura está o servidor dLibra, conforme Figura 12, separado por seis serviços (servidores): metadados, conteúdo, usuário, busca, busca distribuída e evento.

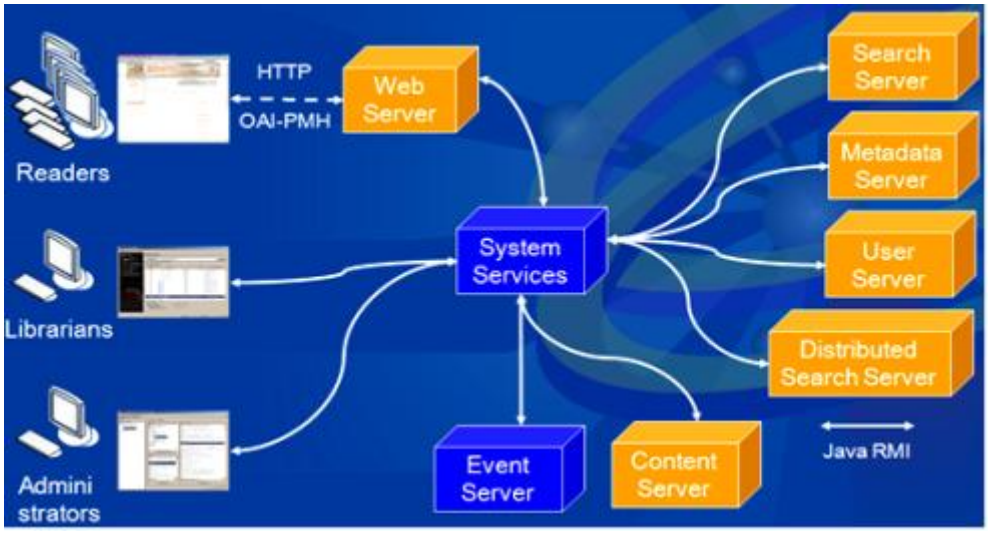


Figura 12 – *Arquitetura dLibra (MAZUREK et al., 2005).*

2.2 Abordagens para Extração de Dados

Com relação à extração de dados, há várias pesquisas que extraem informações de diversas partes de documentos científicos. Em alguns trabalhos as informações são extraídas das imagens, após a aplicação de um OCR, ou seja, diretamente de textos. Porém, dependendo de como está a imagem, o OCR pode não ser tão específico na geração do texto, deste modo outras técnicas podem ser utilizadas, como a segmentação das áreas de interesse e aplicação de filtros para melhorar a aparência da imagem ou mesmo para excluir objetos indesejados como bordas. Em outros casos, a extração acontece em arquivos no formato PDF e/ou HTML, para este último os dados já estão semiestruturados, pois as páginas exibem as informações após um pré-cadastro.

Esta seção apresenta 5 trabalhos relacionadosàextração de dados. A Tabela 3sintetiza as principais características de cada, salientando: os tipos de arquivos utilizados para a entrada de dados, as técnicas utilizadas na extração, e os dados que são extraídos.

Tabela 3 – *Síntese das Abordagens para Extração em Artigos Científicos.*

Ambiente de Extração	Arquivo de Entrada	Técnica	Dados Extraídos
Ferramenta FIP (ÁLVAREZ, 2007)	PDF, PS	Indução de regras de extração	Título, autores, afiliação, resumo, palavras-chave e referências.
Extração em Artigos Científicos Japoneses (OHTA, 2008)	Imagem	Campo Aleatório Condicional para rotulação	Título, autor, *abstract* e referências.
Extração de Autores de Artigos Científicos (CONSTANS, 2009)	HTML	ER; Identificação de letras maiúsculas e quebras de linhas.	Extração do nome de autores
Extração de Referências(SILVA, 2004)	HTML	Técnicas Classificação(K-NN, Naïves Bayes e PART) eHMM	Extração de referências: título, autor, editora, num pág, mês etc.
Contexto de Citações (ALJABER, 2010)	LaTeX e HTML	Clusters; algoritmo K-Means	Palavras ao redor de citações para versimilaridade entre artigos

2.2.1 Ferramentade Extração FIP

O trabalho de Álvarez (ÁLVAREZ, 2007) aborda o módulo, da ferramenta FIP, para extração de informações de artigos científicos. Neles são considerados: o título, autores, afiliação, resumo, palavras-chave e as referências bibliográficas.

Nesta ferramenta, os autores realizaram experimentos em duas etapas. Na primeira etapa foi considerado o corpo do artigoe na segundaas referências bibliográficas. A extração é feita por meio da indução de um conjunto de regras de extração, geradas automaticamente para as referências e manualmente para o corpo dos artigos, por intermédio da análise dos artigos, com a finalidade de descobrir padrões sintáticos existentes.

Na extração das referências bibliográficas, o trabalho aplica o método de etiquetagem semiautomática (rotulação de elementos), para posteriormente, fazer a classificação dos itens que compõem as referências, aproximadamente 30 *tags*. O mapeamento consiste em etiquetar todos os termos do documento selecionado, com um conjunto pré-definido de etiquetas e, posteriormente, combinar e extrair as informações de etiquetas correspondentes. Para o sucesso na extração, informações de natureza gramatical são consideradas.

Os arquivos, no formato PDF e PS, são convertidos para TXT, com a ferramenta *pdftotext*[5]. De posse dos textos os dados são extraídos e o resultado é armazenado no formato XML.

2.2.2 Extração em Artigos em Japonês

Na abordagem proposta em Ohta (2008) a extração de informações é feita em textos de artigos científicos, em idioma japonês, oriundos de um OCR. Primeiramente, são rotulados blocos de texto, como elementos bibliográficos, pré-determinados, depois são rotulados caracteres,em cada bloco de texto.O método usa Campo Aleatório Condicional *(Conditional Random Field* - CRF[6]) para a rotulagem de ambos, os blocos de textos e os caracteres.Os experimentos mostraram que a proposta de rotulagem de um bloco de texto, extraiu todos os elementos bibliográficos, pré-definidos (título do artigo, autor, *abstract*), em mais de 97% dos artigos. A rotulagem proposta para caracteres também extraiu

[5]pdftotext - é uma ferramenta de código aberto, com linha de comando que converte arquivos PDF para textos simples (PDFTOTEXT, 2013).

[6] Campo Aleatório Condicional - é um modelo estocástico utilizado para etiquetar e segmentar sequências de dados e extrair informação de documentos (LAFFERTY, 2001).

corretamente todos os caracteres dos nomes de autores, com mais de 99% dos blocos de texto em japonês.

2.2.3 Extração de Autores de Artigos

Constans (2009) aborda extração de nomes dos autores, de artigos científicos. A extração é feita com o uso de Expressões Regulares. Há a identificação de letras maiúsculas e quebras de linhas para posterior rotulação de caracteres e aplicação das Expressões Regulares.

A relevância do trabalho está nas observações que os textos que começam com letra maiúscula, se não são os nomes dos autores, são títulos ou início de texto e, normalmente, vem acompanhadas das palavras, no caso em inglês, como: *of*, *the*, *and*, *in*, *for*, *from*, *with*, *to* e *on*. Aproximadamente, cinquenta prefixos foram suficientes para extrair, corretamente, todos os autores de um conjunto de 2350 artigos.

2.2.4 Extração de Referências

Em Silva (2004) é proposto um sistema para extrair referências, por meio de uma abordagem híbrida, que combina o uso de técnicas de classificação de textos com os Modelos de Markov Escondidos (*Hidden Markov Model* - HMM). Os experimentos foram divididos em duas fase.

Na primeira fase as referências são divididas em fragmentos ao encontrar vírgula e ponto, para este último, desde que o texto não possua apenas uma letra, como no caso da abreviação de autores.Em seguida, cada fragmento preenche diversos vetores de características, com características definidas manualmente ou automaticamente, há casos em que o vetor possui mais de 100 características; finalmente os vetores são classificados, em mais de 10 campos, tais como: título, autor, editora, número de página, mês etc. As técnicas de classificação de texto utilizadas, nesta primeira fase foram: K-NN, *Naïves Bayes* e PART[7].

A segunda fase, utiliza a saída da primeira fase e com um HMM reclassifica os dados. Nos experimentos foi possível observar que, na maioria dos campos do formulário de saída, houve uma precisão maior quando a segunda fase foi aplicada, com ganho de até 87%.

[7] Algoritmo PART – é um indutor de modelos baseados em regras de decisão; constrói árvores de decisão parciais a cada iteração e transforma a melhor folha da árvores atual em uma regra (FRANK, 1998).

2.2.5 Contexto de Citações

O trabalho de Aljaber (2010) utiliza o contexto de citação (texto ao redor dos marcadores "[]" de referência, usados para se referir a outros trabalhos científicos)para verificar a similaridade entre textos. A janela de busca está limitada ao tamanho de 50 palavras ao redor da citação, a partir de ambos os lados do marcador de citação [*], limitada ao tamanho de um parágrafo.

Para um determinado documento, foi construído um vetor de expressão, que consiste de termos mencionados no texto original. O grau de frequência,ou limiar, de termos definido é igual a 3. A solução adotada usa análise de grupos (*clusters*), como o algoritmo K-Means (TAN *et al*., 2009).Os experimentos trabalharam com, aproximadamente, 6000 documentos de duas coleções científicas: Física de Altas Energias (formato LaTeX) e Genômica (formato HTML).

Os resultados experimentais indicam que o uso de contextos de citação, quando combinado com o vocabulário no texto completo do documento, é um meio alternativo e promissor de capturar temas críticos abrangidos por artigos científicos.

2.3 Modelo de Referência de Biblioteca Digital da DELOS

O Modelo de Referência de Bibliotecas Digitais desenvolvido pela *Network of Excellence for Digital Libraries*(DELOS) apresenta uma estrutura conceitual (*framework*) que visa capturar entidades significativas e suas relações no universo específico das bibliotecas digitais, caracterizando a essência deste universo. Por meio do modelo proposto, podem-se desenvolver modelos mais concretos e padronizados, uma vez que ele fornece um roteiro que permite compartilhar experiências.

Vale ressaltar, que a DELOS destaca a importância da evolução contínua deste modelo de referência, que tem como um de seus financiadores a Comissão Europeia, no quadro do programa *Information Society Technologies* (IST). Os principais objetivos do IST são pesquisas, cujos resultados são de domínio público, e transferência de tecnologia, por meio de acordos de cooperação com as partes interessadas (DELOS, 2012).

Segundo DELOS, uma Biblioteca Digital é uma organização em evolução, que passa a existir a partir de uma série de passos que reúnem todos os componentes necessários. A DELOS considera que o processo de evolução de uma biblioteca digital inclui a noção de três diferentes sistemas que, juntos, formam um *framework* que considera

três níveis de conceitualização do universo desse domínio, que serão descritos na seção 2.3.1 deste capítulo.

Ainda de acordo com DELOS, apesar da grande variedade e diversidade de bibliotecas digitais existentes, há um pequeno conjunto de conceitos fundamentais que constituem a base de todos os sistemas. Eles servem como ponto inicial para qualquer pesquisador que queira estudar e compreender a área, para qualquer projetista e desenvolvedor de sistema construir uma biblioteca digital, e para qualquer provedor de conteúdo que busque expor seu conteúdo por meio de tecnologias de bibliotecas digitais.A seção 2.3.2 deste capítulo apresenta esses conceitos.Para descrever a operação completa da organização da Biblioteca Digital e a forma como se espera entregar os serviços, foram previstos atores interagindo com a biblioteca digital, tendo papéis em três categorias diferentes e complementares. Os mesmos serão descritos na seção 2.3.3.

2.3.1 Framework para Biblioteca Digital

Conforme descrito em Candela *et al.* (2008), o *framework* proposto para bibliotecas digitais, inclui três tipos de "sistemas", os quais formam três camadas distintas. Estes sistemas correspondem aos diferentes níveis de conceitualização do universo de bibliotecas digitais:

- Biblioteca Digital – uma organização abstrata (virtual) com um objetivo social bem definido;ela coleta, gerencia e preserva, por um longo tempo, um conteúdo digital importante, oferecendo funcionalidades (para a comunidade de usuários especializados) sobre esse conteúdo; a sua qualidade é mensurável e ela tem políticas bem definidas;
- Sistema de Biblioteca Digital – um sistema de *software* implantado que se baseia numa arquitetura definida, possivelmente distribuída, e fornece todas as funcionalidades exigidas por uma determinada biblioteca digital. Os usuários interagem com uma biblioteca digital por meio do sistema de biblioteca digital correspondente. Uma biblioteca Digital existe devido a este sistema;
- Sistema de Gerenciamento de Biblioteca Digital – um sistema de *software* genérico que suporta a infraestrutura de *software* apropriada para (i) a produção e administração do Sistema de Biblioteca Digital, incorporando o conjunto de facilidades consideradas fundamentais (ii) a integração de *software* adicional,

oferecendo mais refinamento, instalações especializadas ou avançadas. Ele suporta o ciclo de vida de um ou mais Sistemas de Bibliotecas Digitais.

2.3.2 Conceitos Fundamentais de Biblioteca Digital

O modelo DELOS define seis conceitos fundamentais que são a base para as bibliotecas digitais e que influenciam a sua estrutura (Figura 13): conteúdo, usuário, qualidade, funcionalidade e política aparecem na própria definição de biblioteca digital, já o sexto surge na definição de sistema de biblioteca digital que é a arquitetura.

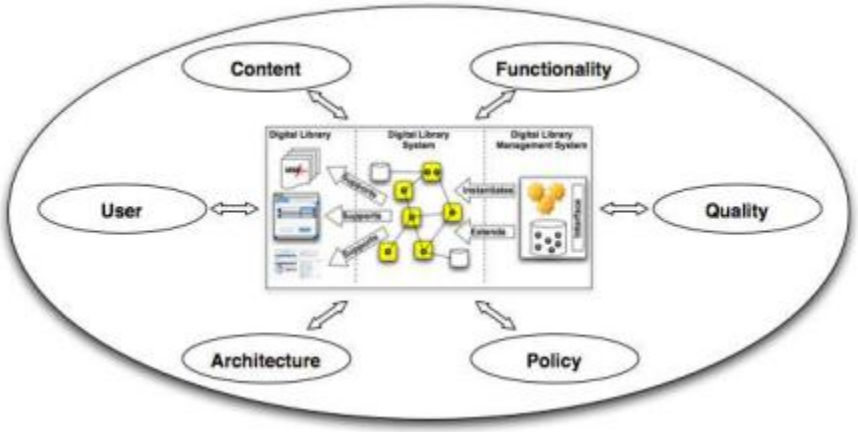


Figura 13 – *O Universo da Biblioteca Digital: principais conceitos (CANDELA et al., 2008).*

A definição de cada um desses conceitos, conforme Candela *et al.* (2008), é descrita a seguir:

- Conteúdo– representa as informações que a biblioteca manipula e disponibiliza aos usuários;
- Usuários –são os atores (humanos e/ou máquinas) que interagem com o sistema;
- Funcionalidade– representa os serviços que a biblioteca digital oferece aos seus usuários; no mínimo eles incluem: inclusão/registro de novos objetos, pesquisa e navegação;
- Qualidade–são os parâmetros que podem ser usados para caracterizar e avaliar o conteúdo e comportamento de uma biblioteca digital. A qualidade pode ser associada não só com classes de conteúdo ou funcionalidade, mas também com objetos de informação ou serviços específicos;

- Política–representa as regras e condições, incluindo os direitos digitais, que regem o funcionamento como um todo. As políticas pertencem a diferentes classes, por exemplo, nem todas as políticas são definidas dentro da biblioteca digital ou na gestão da organização; mais detalhes podem ser vistos no Apêndice B, que mostra entre outros o mapa conceitual de Política de Uso e Acesso (*Policy Domain Conceptual Map*);
- Arquitetura–mapeamento da funcionalidade dos conteúdos oferecidos por uma biblioteca digital sobre os componentes de *hardware* e *software*.

O modelo de referência também define de quatro formas diferentes e complementares os usuários, conformeFigura 14:

- Usuários Finais: clientes finais que a biblioteca digital vai servir, são os consumidores do conteúdo;
- Projetistas: organizadores e orquestradores do ponto de vista da aplicação da biblioteca digital, são eles que caracterizam os serviços;
- Administradores do Sistema: organizadores e orquestradores do ponto de vista físico;
- Desenvolvedores de Aplicativos: implementadores dos *softwares* necessários para criar a biblioteca digital.

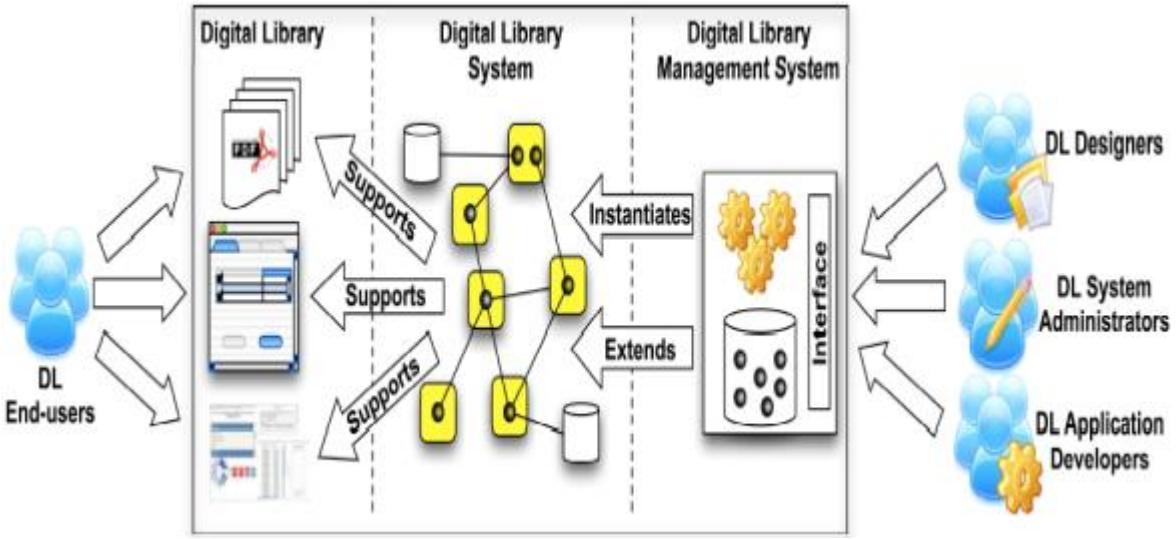


Figura 14 – *Principais atores versus a estrutura de três camadas (CANDELA et al., 2008).*

2.3.3 O Domínio das Bibliotecas Digitais

Segundo DELOS, o universo das bibliotecas digitais pode ser representado em uma hierarquia de domínios, por meio de mapas conceituais como demonstrado na Figura 15. A Biblioteca Digital, o Sistema de Biblioteca, juntamente com o Sistema de Gerenciamento de Biblioteca Digital são todos definidos pelo Domínio de Biblioteca Digital. Este último contém o Domínio dos Recursos da Biblioteca Digital e um Domínio Complementar.

O Domínio dos Recursos (*Resource*) inclui os seguintes domínios: Conteúdo, Usuário, Funcionalidade, Política, Qualidade e Arquitetura. Já o Domínio Complementar contém todos os outros domínios, que, embora não constituam ofoco das bibliotecas digitais, são necessários para representar os sistemas, tais como: Domínio do Tempo (períodos de tempo e intervalos); Domínio do Espaço (regiões e locais); Domínio do Idioma(aspectosdométododecomunicação, falado ou escrito); etc. (CANDELA *et al*., 2008).

O Domínio dosRecursos da Biblioteca Digital é dedicado a capturar os itens comuns de todos os "elementos de primeira classe" do universo das bibliotecas digitais. Ele representa todas as entidades e relações que são gerenciadas no universo da Biblioteca Digital, conforme apresentado naFigura 16.

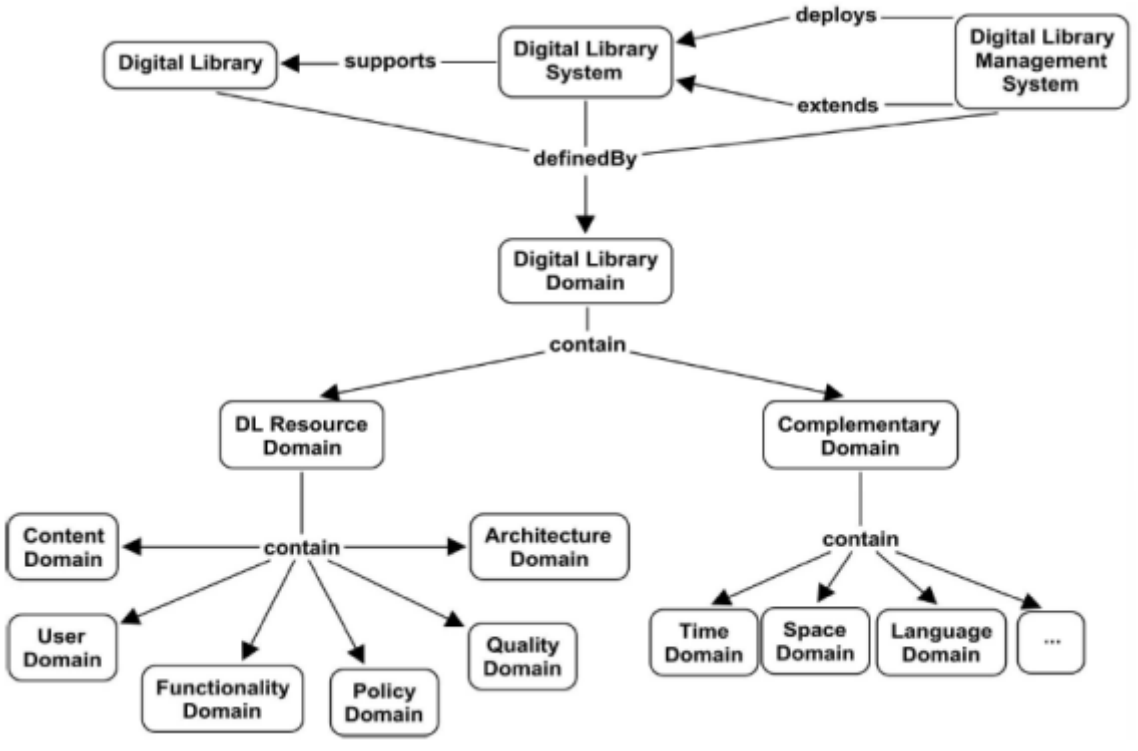


Figura 15 – *Mapa Conceitual do Universo das Bibliotecas Digitais (CANDELA et al., 2008).*

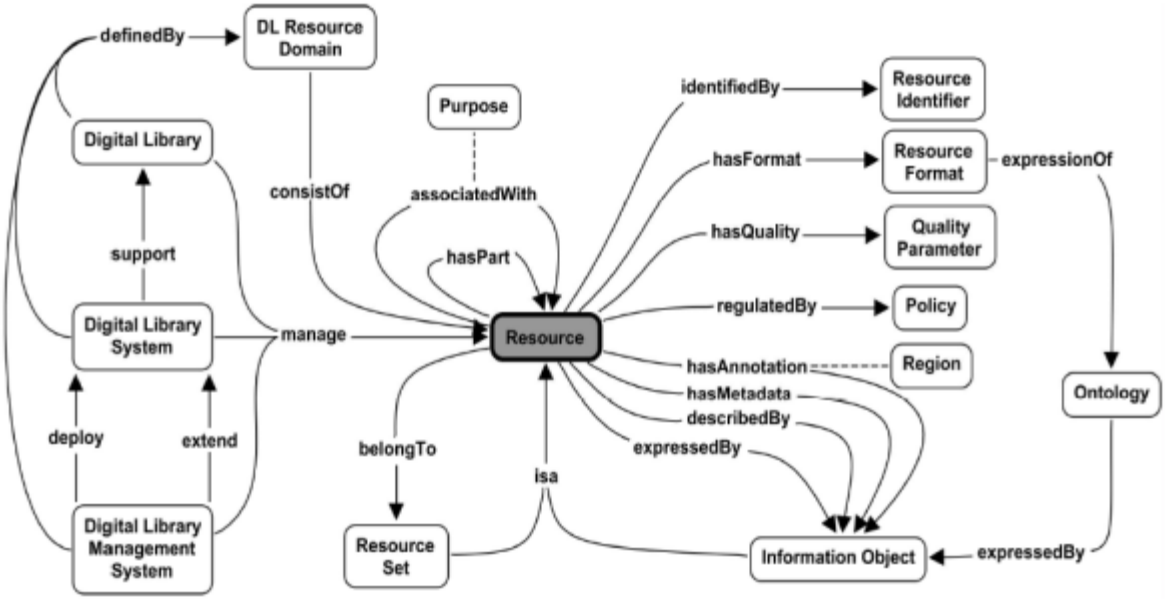


Figura 16-*Mapa Conceitual do Universo da Biblioteca Digital: recursos (CANDELA et al., 2008).*

O conceito mais geral do Domínio dosRecursos da Biblioteca Digitalé o Recurso, que captura as características de qualquer entidade de uma Biblioteca Digital. Um recurso pode possuir partes, e podetambém estar associado a outro recurso.

Os recursos podem ser instanciados por Objetos de Informação **(***Information Objects***)** sou por Conjunto de Recursos (*Resource Set*). Os recursos são regulados por políticas, e possuem: formatos, parâmetros de qualidade, identificadores.Cada recurso representa o conceito principal em seu respectivo domínio, portanto cada domínio contém Recursos, e os recursos são os blocos de construção de todos os domínios da Biblioteca Digital.

Os Objetos de Informação podem ser: documentos, imagens, vídeos, objetos compostos de multimídia, fluxo, bancos de dados, coleções, consultas e seus conjuntos de resultados, atores (humanos e entidades inanimadas) funções e componentes de arquitetura. Um Objeto de Informação pode ser também: uma anotação relacionada a um recurso, definir os metadados de um recurso, descrever ou expressar um recurso ou ontologia.

Como já mencionado, o Apêndice B apresenta outros mapas conceituais relevantesno Universo da Biblioteca Digital, como os exibidos nas Figuras 57, específico para Recursos, Figura 58, abrange o Domínio Ator, Figura 59, Domínio dos Parâmetros de Qualidade, Figura 60, Domínio de Política, Figura 61, Domínio do Sistema de Política e

nas Figuras 62 e 63 as quais mostram mapas conceituais das Políticas e dos Parâmetros de Qualidade do pLiveMemory, respectivamente.

Na Figura 17 pode-se observar que no topo háo Modelo de Referênciamais abstrato, que orienta Arquitetura de Referência mais específica e a Arquitetura Concreta mais abaixo. Estes devem restringir desenvolvimento e implementação de qualquer sistema efetivo. Também consideram elementos de entrada e trabalhos relacionados, os quais incluem:

- Entradas– objetivos, requisitos, motivações e,em geral, o mercado de Biblioteca Digital, como apresentado no lado esquerdo da figura; e
- Trabalhos Relacionados– as boas práticas, pesquisas relevantes, padrões, componentes disponíveis, e protocolos (mostrado no lado direito da mesma figura) que serão usados na Implementação.

Quando essas estruturas são adotadas e seguidas pela comunidade, os sistemas resultantes serão amplamente compatíveis uns com os outros, a interoperabilidade, assim, proporcionada vai abrir novos horizontes, significativos, para pesquisa (CANDELA*et al.*, 2008).

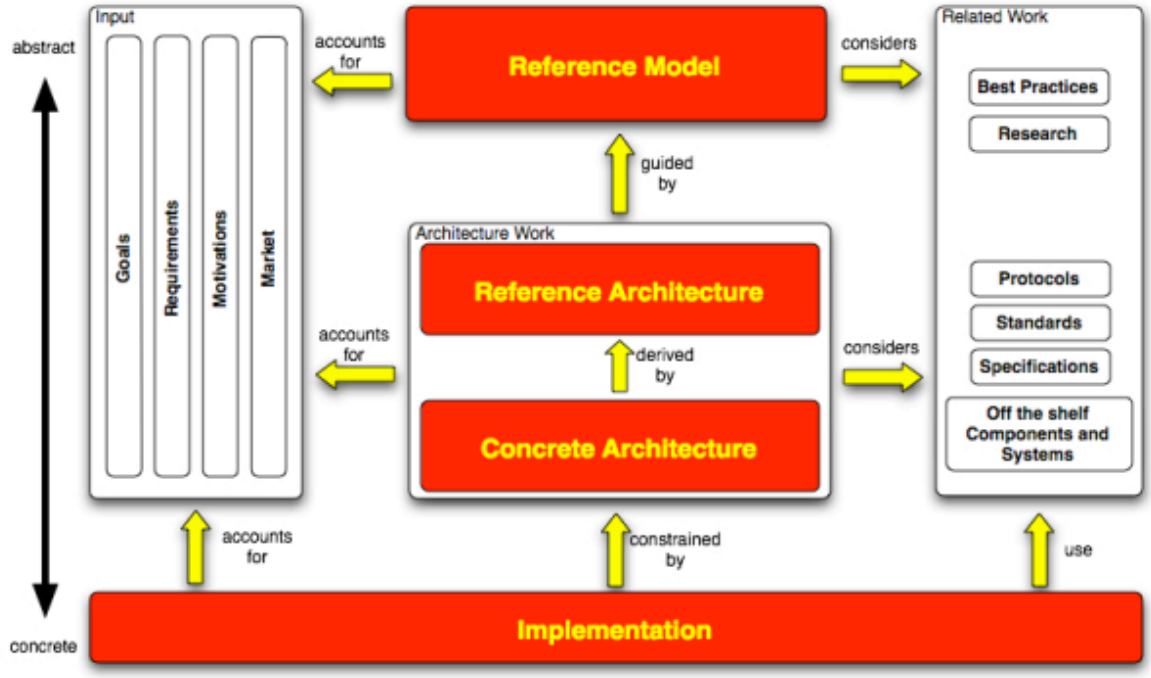


Figura 17 – *Framework para Desenvolvimento de Biblioteca Digital (CANDELA et al., 2008).*

2.4 Considerações

Com relação às abordagens de bibliotecas digitais, observou-se a existência de ambientes que: definem *frameworks*, trabalham com digitalização, e que fazem extração de dados. Entre os objetivos identificados nesses trabalhos estão: disponibilização de informações sobre artigos científicos e publicações com conteúdo específico (matemática, livros da renascença, dados culturais, informações pessoais relacionadas a memórias digitais humanas), e suporte a ambientes colaborativos, como o Cyclades.

Porém, não foi encontrada uma proposta integrada que tratasse desde a geração de bibliotecas digitais, passando pelo tratamento de imagens, extração de dados, como elementos da capa, referências e palavras-chave.

Como esta tese propõe a plataforma pLiveMemory para bibliotecas digitais, considerou-se relevante analisar a aderência do que foi proposto com relação a um padrão existente na comunidade das bibliotecas digitais – o Candela *et al*. (2008).

Apesar de Candela *et al*. (2008) não contemplem etapas como digitalização e filtragem de documentos, mas considerou-se a sua padronização para construção da biblioteca. Essa estratégia permite garantir que a plataforma pLiveMemory faz uso de conceitos e *frameworks* comuns, estimulando uma maior comunicação e interação dentro da comunidade, além de uma descrição mais completa, seguindo os elementos e requisitos previstos pela comunidade (no caso da DELOS descritos por intermédio de mapas conceituais).

No *framework* apresentado pela DELOS, para o desenvolvimento das bibliotecas digitais, existem diferentes níveis de abstração conforme apresentado na Figura 17. No topo, está o Modelo de Referência, o nível mais abstrato que consolida a diversidade de abordagens existentes em um todo coeso e consistente, oferecendo um mecanismo para permitir a comparação de diferentes sistemas de bibliotecas digitais, fornecendo uma base comum para a comunicação dentro da comunidade da biblioteca digital.

O Modelo de Referência por sua vez vai orientar o nível seguinte, que contempla uma Arquitetura de Referência, representando padrão de projeto arquitetônico que indica uma solução abstrata que implementa os conceitos e relações identificadas no Modelo de Referência.

Nesta tese, este modelo será representado pela extensão/instanciação do modelo da DELOS visando atender às demandas relacionadas com a Plataforma pLiveMemory. Por fim, segue uma Arquitetura Concreta para a Plataforma pLiveMemory, que é realizada

substituindo os mecanismos previstos na arquitetura de referência com os padrões concretos e especificações.

Considerando especificamente as abordagens de extração, alguns trabalhos buscam extrair informações sobre artigos científicos, sejam para um domínio específico (artigos em japonês) ou para domínios mais abrangentes. As técnicas são diversas, incluindo expressões regulares, rotulação de dados e indução de regras de extração.

No caso da extração de referências bibliográficas várias particularidades são observadas, como: as técnicas utilizadas (rotulação de elementos, expressões regulares, HMM, K-NN, *NaïvesBayes*) e a quantidade de dados a serem extraídos (título, autores, referências, instituições, *e-mails* etc.).

Observou-se que não existia um ambiente integrado, com solução única para as diversas extrações, pretendidas neste trabalho, dados da primeira página do artigo, referências e palavras-chave.

O trabalho de Álvarez (2007) não é totalmente automático e trabalha com várias itens das referências. Em Ohta *et al*. (2008) a extração é especifica para artigos científicos, em idioma japonês, e foca na extração de blocos (título, autor, resumo) para posterior extração, considerando caracteres, dos autores. Constans (2009) aborda extração, apenas, de nomes dos autores, de artigos científicos, nomes estes presentes no início dos artigos. A abordagem de Silva (2004) se aproxima da definida neste trabalho, porém utiliza duas etapas, tornando o processo mais complexo, sem contar que o vetor de características possui mais de 100 características e são extraídos mais de 10 campos para cada referência. Por fim, o trabalho de Aljaber *et al*. (2010) utiliza o contexto de citação para verificar a similaridade entre textos, o resultado se aproxima da abordagem para extração de palavras-chave, porém extrai itens com o intuito similar das palavras-chaves, ou seja, encontrar trabalhos relacionados.

3 GERAÇÃO DE BIBLIOTECA DIGITAL PARA ANAIS DE EVENTOS CIENTÍFICOS

Os modelos de referência podem ser usados como uma forma de orientar a construção denovos *softwares*. Nesses modelos são usadas técnicas de reuso para definir, *frameworks*, conceitos,padrões de funcionamento, entre outros. Estecapítulo descreve a Plataforma pLiveMemory, num escopo que envolve documentos impressos ou eletrônicos, considerando não só o modelo de Referência da DELOS, mas também a experiência adquirida em(ALVES *et al.*, 2011), e no desenvolvimento das plataformas: *Academus*(LINS *et al.*, 2011) e *Thanatos*(ALMEIDA *et al.*, 2011).

Este capítulo, inicialmente,abordaquestões de reuso em Engenharia de Software, uma vez que se parte deste conceito para fundamentar a montagem de toda a plataforma. Em seguida, a Plataforma pLiveMemory é descrita usando como base o modelo de Referência da DELOS.

3.1 Reuso em Engenharia de *Software*

O conceito de reutilização de *software*, segundo Krueger(1992), é fundamental naEngenharia de *Software* moderna, ele consiste em empregar artefatos de *softwares* desenvolvidos, anteriormente,para a geração de um novo sistema. A reutilização de*software* pode ser aplicada em qualquer produto que faz parte do ciclo de vidado *software*, não apenas em fragmentos de código. Isto significa que os desenvolvedores podem fazer a reutilização de documentos de requisitos, especificações, arquitetura de projetos, ou qualquer outro artefato no processo de desenvolvimento de *software*(BARNES*et al.*, 1991).

Muitas técnicas foram desenvolvidas para ajudar no reuso de *software*, com o objetivo de diminuir os custos e esforços, além de melhorar a qualidade e a confiabilidade do *software* (FAYAD*et al.*,1999). Entre as abordagens de reuso destacam-se:

- Padrões (*Patterns*):Alexander *et al*. (1977) definiu padrão como uma entidade que descreve um problema que ocorre repetidamente em um ambiente e então descreve a essência de uma solução para este problema, de tal forma que se possa usar essa solução milhões de vezes, não obrigatoriamente do mesmo modo. Existem vários níveis onde os padrões podem ser aplicados na Engenharia de *Software*, tais como: projeto, análise, programação etc.

- *Frameworks* (FAYAD*et al.*, 1999) são usados quando há funcionalidades comuns para diferentes aplicações; como citado no início do Capítulo 2, de acordo com Gamma *et al.*(1995)um *framework* é um esqueleto de um sistema que pode ser personalizado para uma determinada aplicação, sendo um guia para o projeto do reuso das classes envolvidas na definição de suas responsabilidades e colaborações;
- Modelos de Referência são formados por um conjunto mínimo para unificar conceitos, axiomas e relacionamentos dentro do domínio de um problema particular (CANDELA*et al*., 2008);
- Mapa Conceitual é a representação gráfica, em duas ou mais dimensões, de um conjunto de conceitos construídos de tal forma que as relações entre eles sejam evidentes (NOVAK, 1998);
- Ontologias são usadas para estabelecer um entendimento comum sobre um domínio de interesse, funcionando como uma linguagem de intercâmbio para a comunicação entre aplicações e organizações. São assim muitas vezes usadas como modelos de referência (GUIZZARDI, 2005).
- Linhas de Produto de *Software* (LPS) focam no desenvolvimento, considerando como base uma família de aplicações (CLEMENTS *et al.*, 2002); elas representam conjunto de produtos de *software* relacionados que são gerados a partir de ativos reutilizáveis. Produtos estão relacionados no sentido de que eles compartilham funcionalidades comuns.
- Modelos de *Features*descrevem um domínio por meio da representação das características comuns e variáveis de uma LPS. É entendido que os relacionamentos e restrições de um modelo de*features* representam uma fórmula proposicional cuja instância corresponde a membros de uma LPS (BATORY, 2005).

Todos esses conceitos relacionados ao reuso vêm sendo usados nos mais diversos domínios de desenvolvimento de *softwares*, facilitando a implementação de novas aplicações, quer repetindo soluções prévias, comparando-as, customizando *frameworks*, invocando ou integrando componentes, entre outros.

3.2 Objetivo da PlataformapLiveMemory

O foco da plataforma pLiveMemory está na construção/gerenciamento de uma biblioteca digital que, além de divulgar informações de artigos, também considera a digitalizaçãode documentos no formato de papel, trata as imagens, extraidados e disponibiliza-os via *web*. Porém, nem sempre haverá a necessidade de todas essas etapas, como a digitalização e consequentemente o tratamento das imagens, quando se trabalha com documentos no formato PDF.

O objetivoéportanto ser uma plataforma para o Gerenciamento e Manutenção de Bibliotecas Digitais para DocumentosCientíficos, que permite:

- Manter diferentes entidades/elementos relacionadas aos artigos (conferências, edições, configurações, instituições, pesquisadores, entre outros);
- Identificar/selecionar artigos em vários formatos(papel, PDF ou JPEG/TIFF/PNG), para posterior digitalização/tratamentoe visualização;
- Digitalizar artigos em papel;
- Aplicar filtrosnos artigos em formato de imagem, visando aumentar a qualidade dos mesmos em termos de visibilidade dos caracteres do texto;
- Extrair texto de imagem e/ou PDF: extrair informações específicas dos arquivos, de acordo com a formatação existente;
- Extrair características/elementos dos arquivos (tais como,autores, título, referência, entre outros);
- Realizar consultas sobre os conteúdos dos artigos, tais como: lista de autores, palavras-chave mais usada, autores mais referenciados etc.;
- Gerarrelatórios quepossibilitem a visualização organizada dos dados armazenados no banco de dados.

Para o desenvolvimento da plataforma foram consideradas melhores práticas,pesquisas relevantes, protocolos, padrões, especificações e componentes.Destaca-se as ferramentas desenvolvidas para extração e identificação de conteúdo dos artigos, tais como, palavras-chave e referências, exibidas nas Figuras 18 e 19, como componentes que se relacionam com o ambiente para *desktop* e que serão detalhadas na seções 3.3 e 3.4, respectivamente.

3.3 Instanciação doModelo de Referência da Delos

O Modelo de Referência da DELOS, que contempla um *framework* dividido em 3 camadas, foi usado para definir a estrutura geral da proposta. Conforme pode ser observado na Figura 18, o Sistema de Gerenciamento da Biblioteca Digital proposto pela DELOS érepresentado aqui como a Plataformado Sistema de Gerenciamento pLiveMemory responsável por manter toda infraestrutura e o ciclo de vida dos sistemas a serem criados. Já o Sistema da Biblioteca Digital corresponde aqui ao Sistema do LiveMemory (sLiveMemory), responsável por manter a biblioteca em si disponível. Por fim, a Biblioteca Digital, vista pelos usuários finais, é ilustrada como a Biblioteca da SBrT, usada nesta tese como estudo de caso.

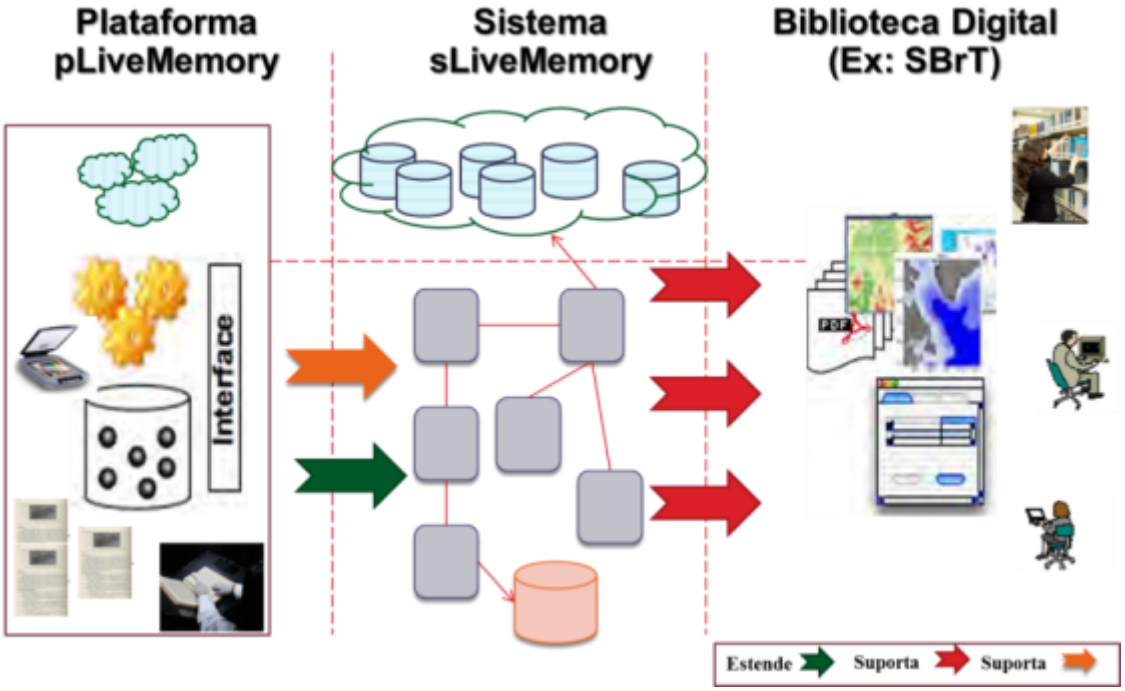


Figura 18 – *A estrutura proposta representa pelo Framework do Modelo de Referência da Delos.*

O esquema proposto para a pLiveMemory,que tratae mantém informações de documentos é retratado na Figura 19. Ele contempla desde a fase de digitalização de documentos, até à disponibilização da biblioteca digital completa nas nuvens ou em *desktop*. Pode-se observar que diversos *softwares* prontos podem ser usados para auxiliar a montagem do funcionamento da plataforma, como OCRTesseract (2011), PDFBox (2011), Abbyy FineReader (ABBYY, 2011) e BigBatch (LINS*et al.*, 2006), assim como *hardware*

para apoio à digitalização – o *scanner* – e plataformas de dados em nuvem, por meio dos servidores de dados da nuvem, como por exemplo o Google©.

A execução do esquema pode ser descrita em várias etapas, conforme detalhamento a seguir:

- (i)o documento em formato de papel é selecionado para digitalização;
- (ii) o documento é enviado ao *scanner*;
- (iii) como resultado tem-se um documento digitalizado, em formato de imagem;
- (iv)o sistema *BigBatch*, que trata de imagens é utilizado;
- (v) o documento tratado pelo *BigBatch*é gerado;
- (vi) o OCR (*OCR-Tesseract*) é aplicado;
- (vii) as páginas do documento são armazenadas em um diretório específico;
- (viii) documento em formato PDF são selecionados;
- (ix) um extrator de informação, por exemplo, o *PDFBox* é executado;
- (x) o resultando das etapas (vi) e (ix) é um documento no formato TXT;
- (xi) os dados são extraídos do TXT;
- (xii) os dados são armazenados no banco de dados;
- (xiii) Servidores na Internet, como o *Google Drive*©, são usados para o envio dos documentos no formato PDF;
- (xiv) o arquivo, em formato de imagem, é transformado para o formato PDF e o *softwareAbbyy FineReader*(2011) pode ser utilizado;
- (xv) as informações dos documentos são publicadas em um *site* público, como o Google Sites©.

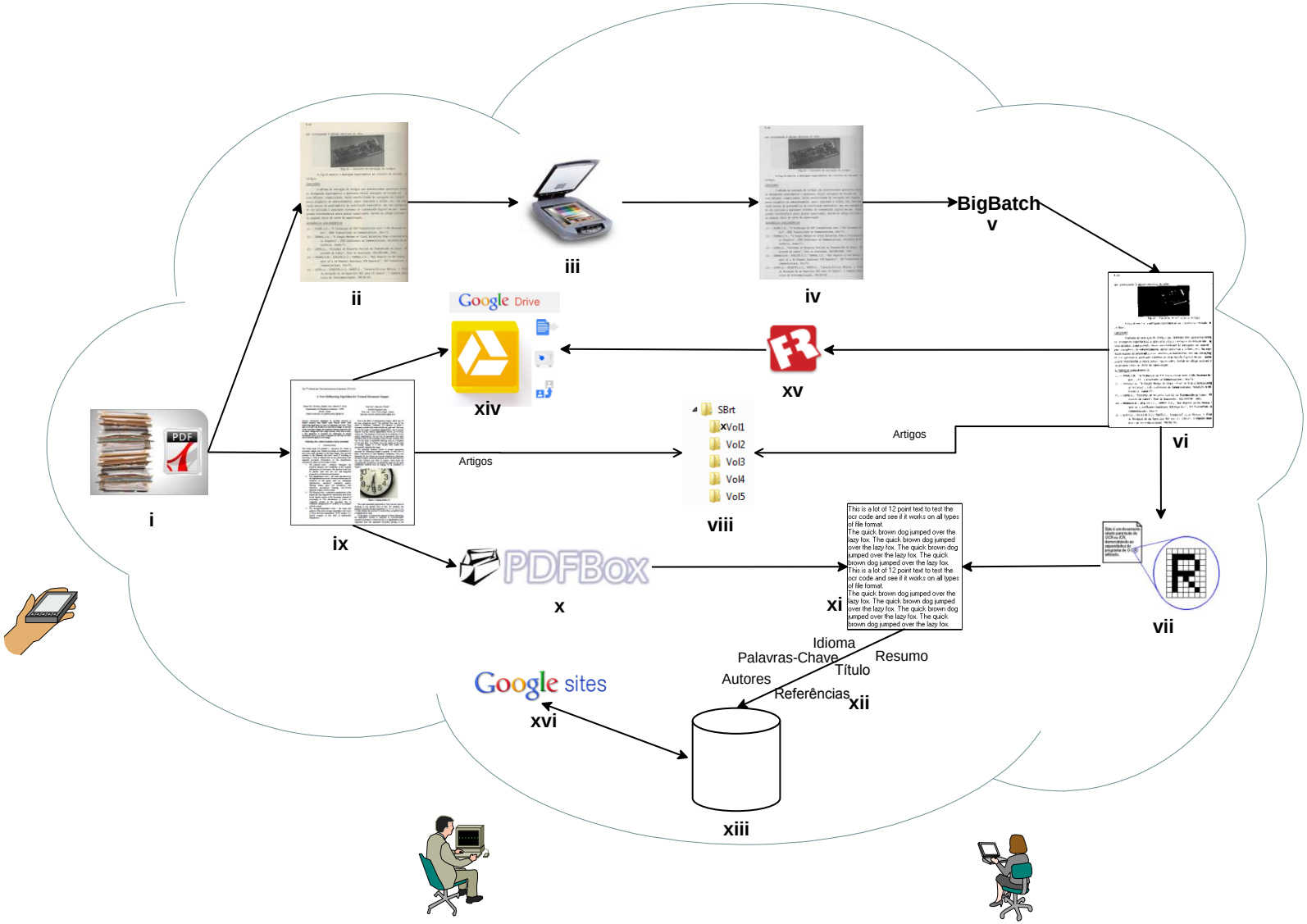


Figura 19 – *Esquema para tratar e manter informações de documentos.*

3.4 Arquitetura da PlataformapLiveMemory

A arquitetura da Plataforma pLiveMemory é baseada no modelo em três camadas:Apresentação, Negócio e Dados. Essas camadas são detalhadas a seguir, de acordo com aFigura 20:

- A camada de Apresentação está dividida em duas partes independentes, a que é disponível via *desktop* e a que está disponível na Nuvem do Google©;
- A camada de Negócio é executada nos servidores da UFPE. Ela contempla engenhos, como Extração de Referências e Busca de Palavras-Chave;
- A camada de Dados, implementada numa base de dados do MySql (2013), está localizada na UFPE com os dados que foram extraídos dos artigos. Nesta camada também estão os arquivos, em formato TXT, com os endereços dos arquivos, em formato PDF,os quais foram enviados ao Google Drive©.

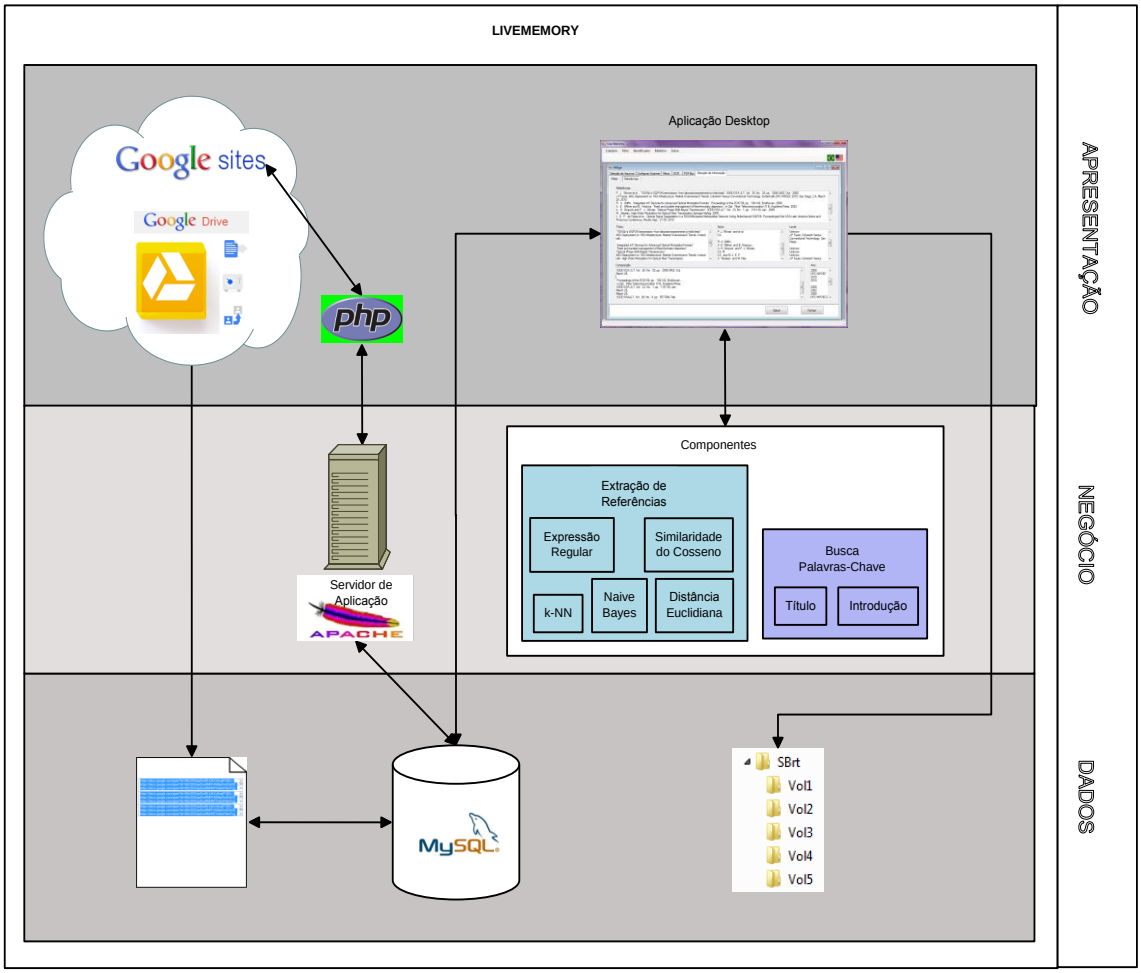


Figura 20 – *Arquitetura em três camadas dapLiveMemory.*

3.5 Descrição dos Conceitos napLiveMemory

Alguns dos conceitos presentes na pLiveMemory são descritos a seguir, por intermédio de uma mapa conceitual. Na Figura 21 se observa que a pLiveMemory implanta e estende o Sistema LiveMemory (sLiveMemory), que por sua vez dá suporte à Biblioteca Digital, no caso ilustrado aqui a Biblioteca da SBrT (que inclui os artigos dos anais de cada edição da conferência SBrT). Os documentos, de acordo com Candela *et al*. (2008), são considerados Objetos de Informação, os quais são agrupados em Conjuntos (*Resource set*) que, por sua vez, representam cada edição da conferência. O conjunto das edições da conferência forma o Histórico dos Anais da SBrT, onde cada edição pode ter uma formatação específica relacionada aos artigos. Outros objetos que podem ser visualizados são os atores, as consultas a serem realizadas, os artigos em PDF e o próprio banco de dados.

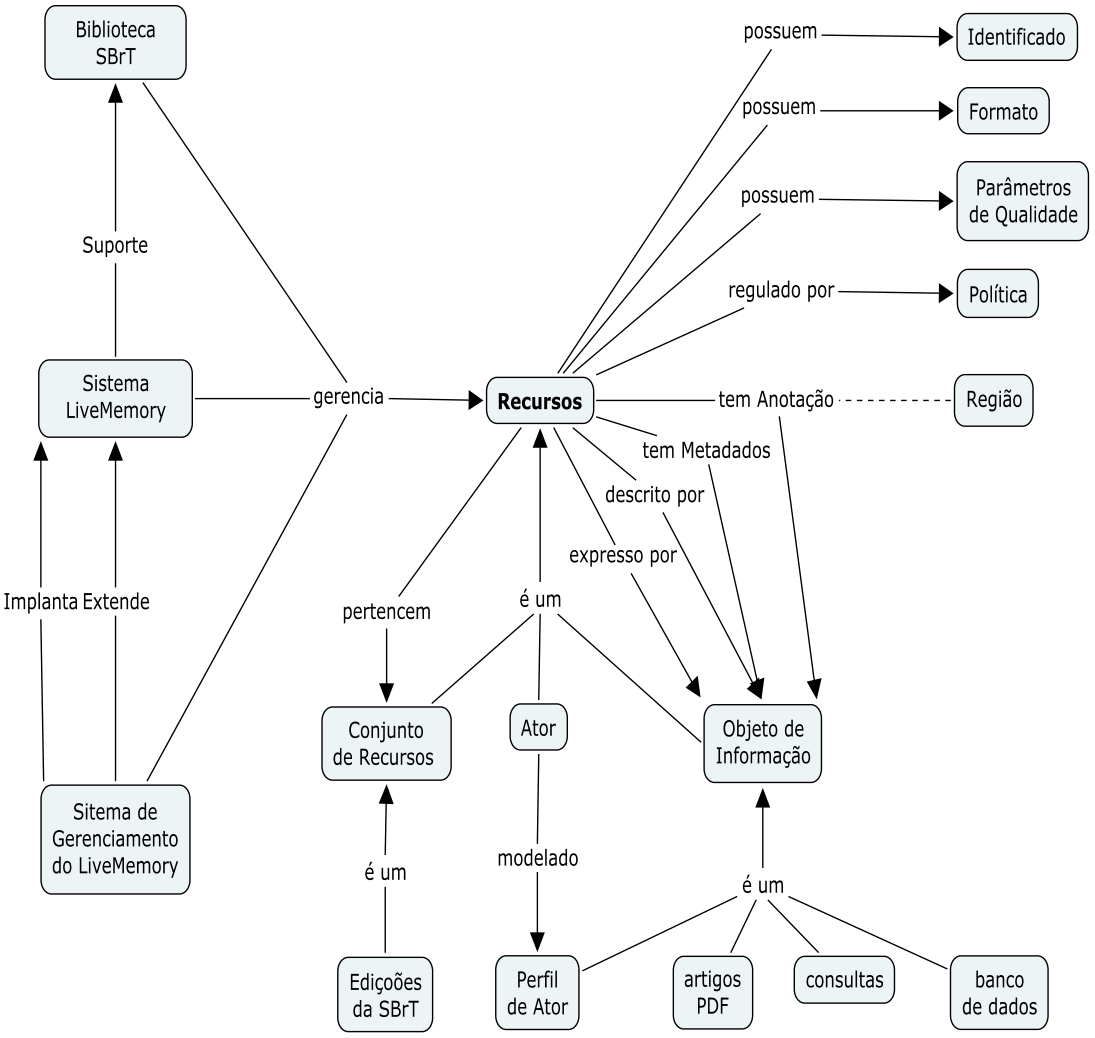


Figura 21 – *Mapa Conceitual dos Principais Recursos da pLiveMemory.*

Os recursos instanciados (expressos ou descritos) por objeto de informação também possuem:

- Identificadores únicos – código definido pelo sistema, mas há o identificador para acessar os arquivos, em PDF, que foram enviados ao Google Drive©;
- Formato dos documentos – os arquivos, acessados via *web*, estão no formato PDF, porém, no ambiente para *desktop*, os formatos em papel e imagem também são aceitos;
- Parâmetros de qualidade – são evidenciados no tempo de acesso aos documentos, na qualidade de visualização dos textos (artigos antigos, entre outros). O fato dos documentos estarem na nuvem (no caso Google)permite que estes estejam mais acessíveis;
- Políticas – o usuário não pode copiar o conteúdo dos PDFs, nem fazer *downloads*, alterar informações da base de dados etc.;
- Anotação dos recursos – são os*hiperlinks* dos arquivos;
- Metadados – são identificados por meio das informações sobre a característica dos arquivos (tamanho, formato etc.), bem como as informações sobre a base de dados.

3.6 Atores da pLiveMemory

Os atores também podem ser vistos como objetos de informação. ApLiveMemory considera que existem oito tipos de atores, conforme Diagrama de Caso de Uso descrito na Figura 56, no Apêndice A (Descrição dos Casos de Uso) desta tese.Entre eles tem-se o Usuário, papel que realiza: Consultas, Cadastro, Avaliação do Sistema (é um dos parâmetros de Qualidade). Mesmo sendo uma biblioteca não há o papel, explícitode um bibliotecário, este é desempenhado pela equipe técnica, que é o ator Usuário. Outros papéis são desempenhados por: *Scanner*; *software* BigBatch, sistema que realiza a etapa de filtragem; *software* OCR; *software* PDFBox; e sistema Classificador que é o módulo que extrai e classifica as informações automaticamente.

No diagrama de caso de uso não há necessidade de explicitar os atores que desempenham papéis para construção e manutenção da biblioteca digital, mas o modelo de referência os recomenda, como descrito na Figura 22, entre os papéis tem-se: usuário final, projetista, administrador e desenvolvedor. O usuário pode ser: criador de conteúdo e este é instanciado pelo operador do sistema que é o responsável por realizar a digitalização dos

documentos, que estão em papel; o consumidor de conteúdo é instanciado pelos pesquisadores; o administrador tem como função administrar o sistema.

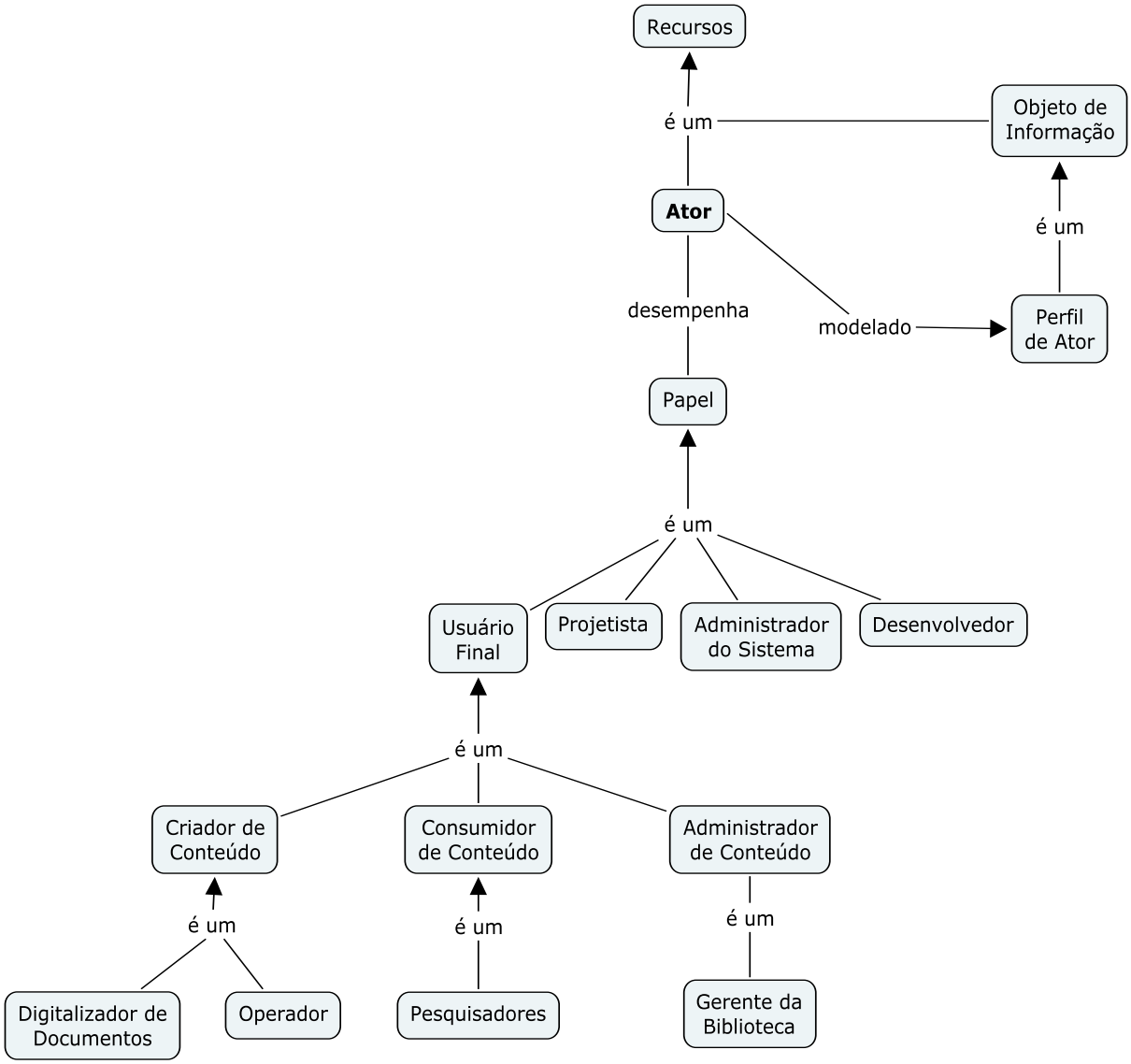


Figura 22 – *Mapa Conceitual dos Atores do pLiveMemory.*

3.7 Modelagem dos Requisitos e Modelo Conceitual dapLiveMemory

A descrição dos Requisitos é apresentada no Apêndice Apor meio do diagrama e especificações de Caso de Uso,com o objetivo de demonstrar, de maneira objetiva, o que o sistema faz e quais as possíveis interações dos usuários com o sistema.

Quanto ao modelo conceitual o pLiveMemory é composto por 17entidades, de acordo com a Figura 23, a saber:Artigo (Documento Científico), Instituição, Autor,

Pesquisador, Chair, Usuário, Referência, Edição, Conferência, Área, SubÁrea, Tópico, Configuração, Palavra-Chave, Resumo, Idioma e Classe de Palavras.

As entidades podem ser divididas em três grandes grupos: o registro do artigo em si, dados da edição e dados auxiliares. Para o registro dos Artigos é necessário o cadastro de informações sobre: Autor, Instituição, Referência, Pesquisador, Palavra-Chave e Resumo. Quanto aos dados da Edição são necessárias as informações sobre: Edição, Conferência, Chair, Área, SubÁrea e Tópico. As informações auxiliares são compostas pelas demais entidades: Configuração, Idioma, Classe de Palavras e Usuário.

Vale ressaltar, que a classe pesquisador é usada para representar pessoas da academia que são: autores de um documento científico, como artigo, ou autores de referências bibliográficas, bem como podem participar da edição como *chair* (responsável pela seleção dos artigos). Assim, na classe "Pesquisador" a pessoa é única, mas isso não é verdade na classe "Autor", pois em um documento científico a pessoa Autor pode estar relacionada a diferentes instituições.O modelo lógico do banco de dados está descrito no Apêndice C, desta tese.

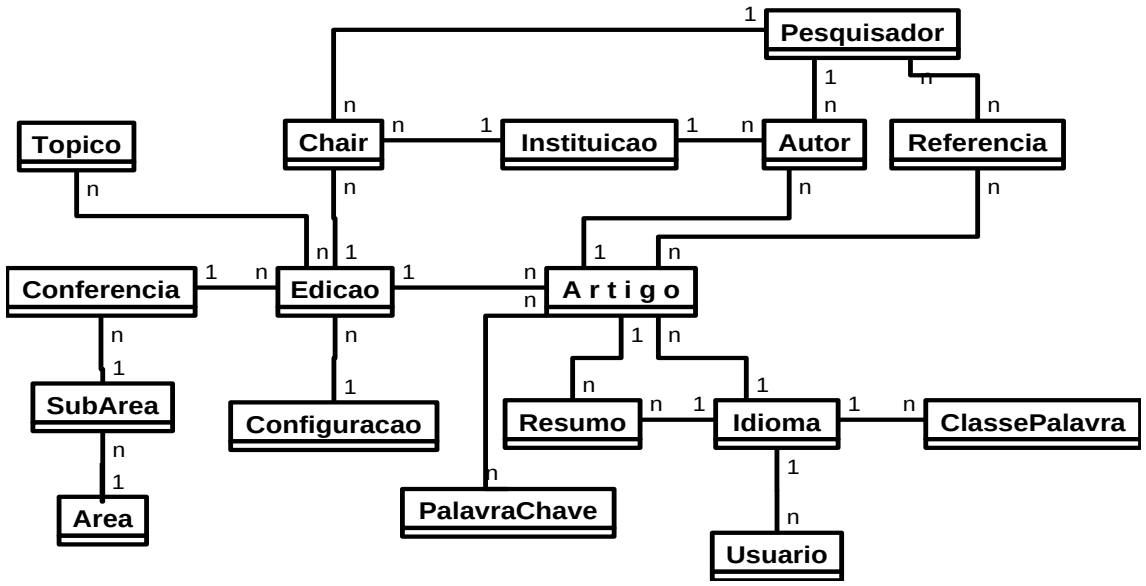


Figura 23 – Diagrama de Classe do pLiveMemory.

3.8 Fluxo para a Geração de Biblioteca Digital de Evento Científico

O fluxo geral de informação dapLiveMemory para geração de uma Biblioteca Digital está descrito por intermédio do Diagrama de Atividades, verFigura 24, conforme os seguintes passos:

- Registrar as informações da conferência, se isso não foi feito anteriormente, e sua edição;
- Indicar se os artigos da edição foram impressos ou não. Se impresso informar o diretório das páginas digitalizadas que serão renomeadas durante o "*upload*" para o sistema;
- Identificar o tipo de arquivo (JPEG, PNG, TIFF). O usuário pode aplicar filtros para melhorar a qualidade das imagens e, em seguida, aplicar *software*OCR ou o *software*PDFBox, se os arquivos estiverem em formato PDF;
- Extrair informações sobre o artigo (título, autores, resumo, palavras-chave e referências);
- Após a extração das informações cada parte extraída é enviada a uma base de dados.

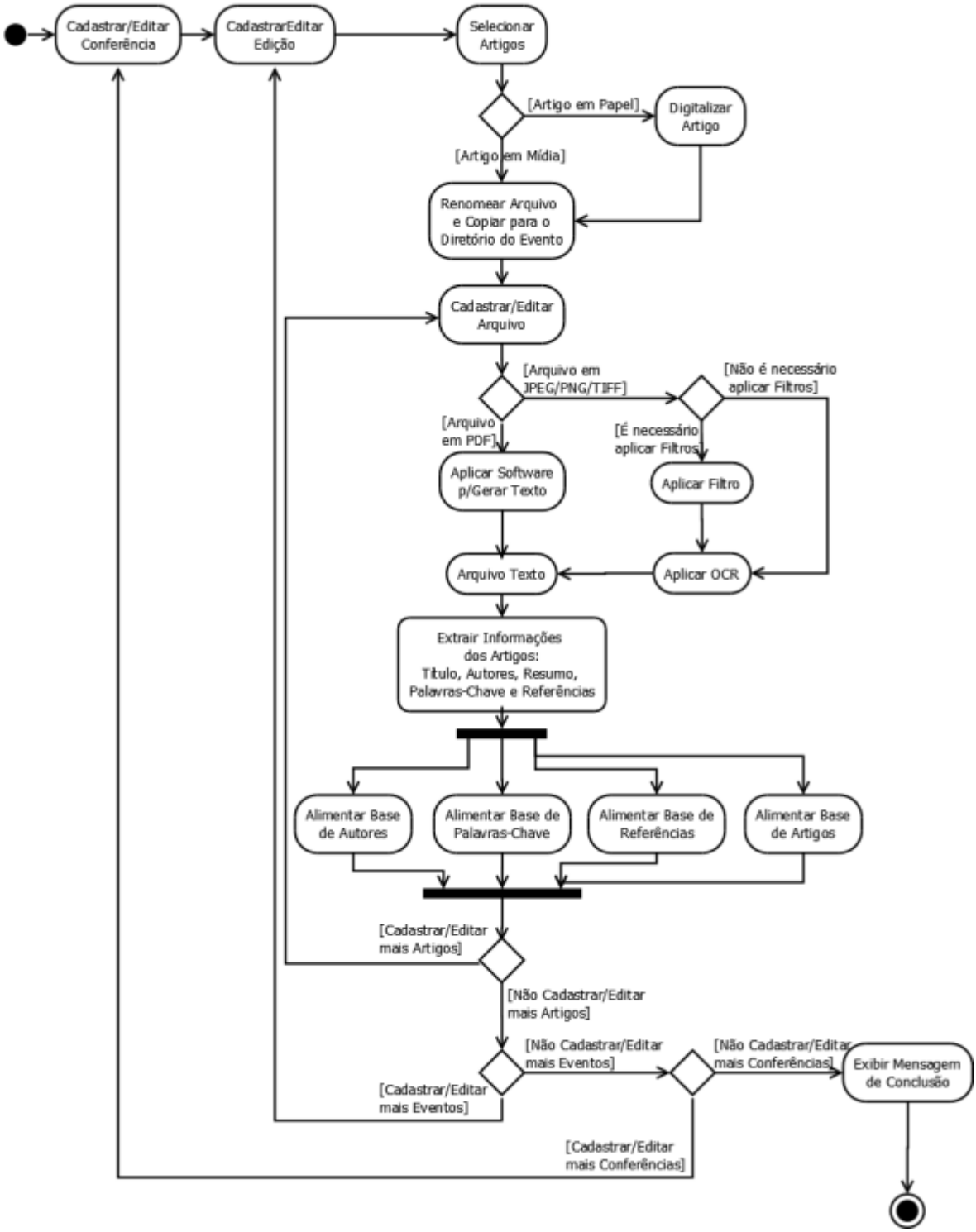


Figura 24 – *Diagrama de Atividades dapLiveMemory.*

De posse dos conceitos de reuso, considerou-se não apenas o tratamento de bibliotecas digitais para ambientes científicos, mas também a experiência com a montagem de outras bibliotecas como,*Academus* e *Thanatos* dentro do mesmo contexto de formato

dos documentos. Gerando a proposta de um padrão para reuso de informação na montagem de novas bibliotecas digitais (ALVES *et al*., 2012b).

A plataforma *Academus* cria bibliotecas digitais de dissertações e teses. São armazenados, além dos arquivos no formato PDF, informações, tais como: título, autor, orientador, co-orientador, área de concentração, sumário, *abstract* e palavras-chave.

Já o *Thanatos* é uma plataforma projetada para extrair informações dos documentos civis (certidões de nascimento e de óbito) do Estado de Pernambuco, ou seja, é formada por uma coleção de livros “escritos à mão” mantidos pelas autoridades locais a partir da primeira metade do século 20.

A árvore de *features*, apresentada na Figura 25, descreve as propriedades/características associadas ao gerenciamento de documentos de uma forma geral. A partir dessas plataformas pode-se considerar aspectos recorrentes que permitem reuso de vários artefatos comuns identificados durante o levantamento dos requisitos. Assim, a especificação dos artefatos envolvidos representa uma espécie de *framework* que inclui: esquema geral, problemas identificados, requisitos, restrições, funcionalidades e dados persistentes. Os detalhes são apresentados de tal forma que a informação e conhecimento representados pelos artefatos, obtidos a partir de experiências anteriores, ajudamna construção de novas bibliotecas digitais e também servem para melhorar a sua qualidade.

Na árvore de *features*, os recursos obrigatórios são indicados com um círculo preenchido, tais como: tipo de texto, formato, extração e geração de texto. Os demais recursos, representados por círculos vazios,são opcionais como, por exemplo, digitalização e filtro (que só serão necessários se os documentos estiverem em formato de papel e após a digitalização necessitarem de tratamento para melhoria da imagem ou mesmo segmentar partes da imagem). Também existem algumas características que são classificados como alternativas, tais como: formato, que podem ser papel, imagem ou PDF, e a geração de texto, a qual pode, alternativamente, via OCR ou extrator de texto como o *software*PDFBox.

A árvore de *features* da Figura 25, centra-se no processo de extração de informações, mais especificamente, de documentos científicos, os quais incluem atributos, tais como: título, autor, *email* etc. A Tabela 4 apresenta uma visão geral das principais características, considerando o gerenciamento de documentos descritos na árvore de *features* e das plataformas analisadas: pLiveMemory, *Academus* e *Thanatos*.

Considerando as características obrigatórias (tipo de texto, formato, extração e geração de texto) todas as três plataformas as possuem. As três plataformas aceitam como entrada os documento em formatosde papel, porém *Academus*não aceita o formato de imagem, pois precisa ser digitalizado, e *Thanatos* não aceita a entrada do formato PDF. No caso da digitalização, apesar de todas as três plataformas incluírem esse recurso, pode-se argumentar que nem sempre é obrigatória, da mesma forma como o Filtro já que as imagens consideradas podem ter boa qualidade no momento da transformação para o formato de texto. *Thanatos* considera fonte de texto mista, manuscritas e impressas. *Thanatos* também não trabalha com a extração no formato PDF, pois como mencionado, anteriormente, não aceita este formato.

Tabela 4 – *Principais características das plataformas.*

Características		**pLiveMemory**	***Academus***	***Thanatos***
Tipo de Texto	Manuscrito	N	N	S
	Impresso	S	S	S
Formato	Papel	S	S	S
	Imagem	S	N	N
	PDF	S	S	N
Digitalização	Manual	S	S	S
	Lote	S	N	N
Filtro	Segmentação	S	S	S
	Preto e Branco / Cinza	S	S	S
Extração	Artigos Científicos	S	N	N
	Dissertação e Tese	N	S	N
	Certidão de Óbito	N	N	S
Geração de Texto	OCR	S	S	S
	PDFBox	S	S	N

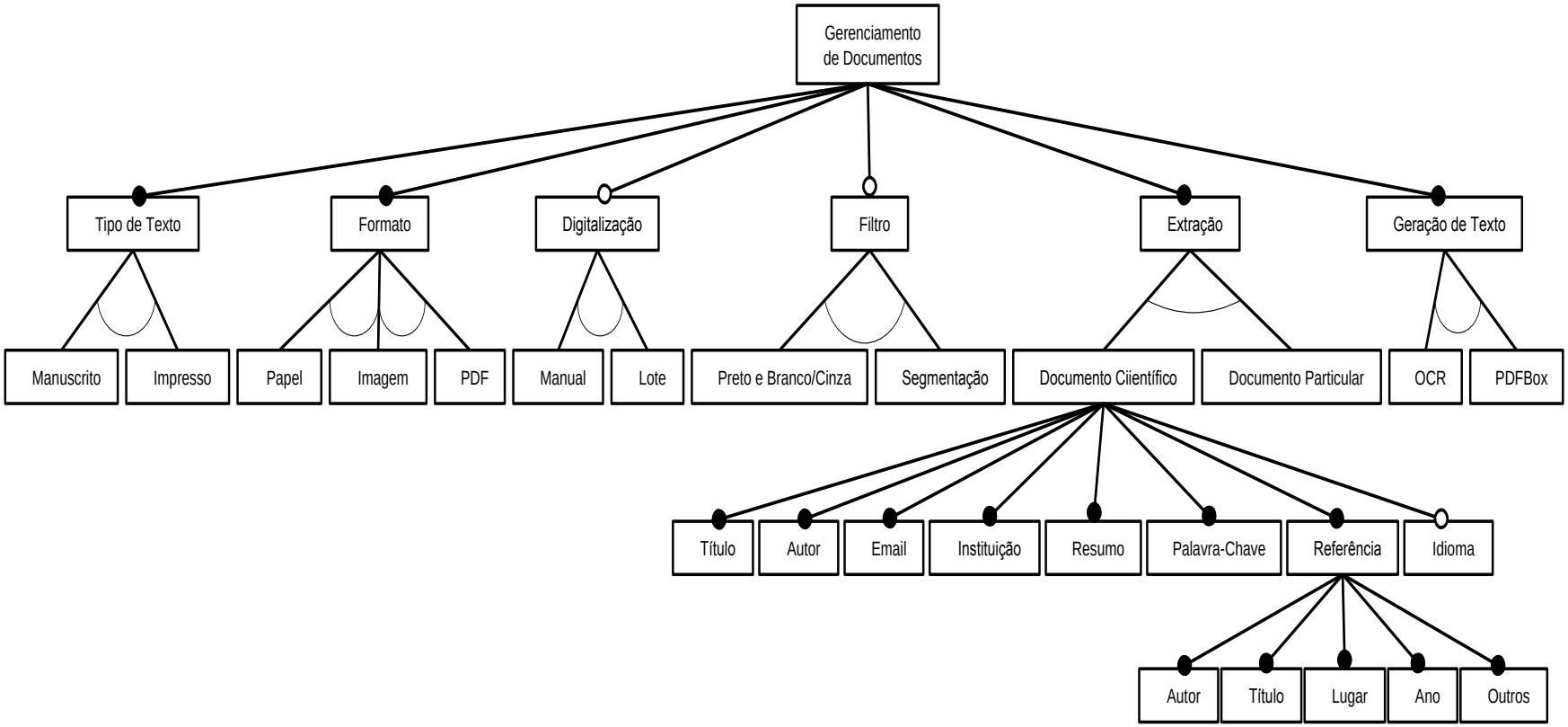


Figura 25 – *Diagrama Árvore de Features para Gerenciamento de Documentos.*

4 PLATAFORMApLIVEMEMORY

Neste capítulo serão descritasas estratégias desenvolvidas para geração dos principais componentes, para *desktop*, daPlataforma pLiveMemory. Estas incluem:

- Processamento de imagem: onde os documentos mais antigos da SBrT foram digitalizados (LINS *et al*., 2009) e passaram por tratamentos para extração de imperfeições oriundas do processo e, neste caso, os filtros da ferramenta BigBatchforam utilizados;
- Modelagem dos dados relacionados à Biblioteca Digital da SBrT e seu armazenamento em base de dados (LINS *et al.*, 2010);
- Extração automática de informações dos documentos da Biblioteca Digital (ALVES *et al.*, 2011);
- Extração de referências bibliográficas dos documentos da Biblioteca Digital (ALVES *et al.*, 2012a);
- Identificação e extração de palavras-chave dos títulos e/ou da introdução dos artigos;
- Desenvolvimento de um *site* para divulgar e possibilitar a consulta dos artigosda SBrT, em formato PDF, este ambiente será descrito no próximo capítulo.

4.1 Processamento de Imagem

Segundo Gonzalez e Woods (2000), o interesse em métodos de processamento de imagens digitais decorre de duas áreas principais de aplicação: (i) melhoria de informação visual para a interpretação humana e (ii) o processamento de dados de cenas para percepção automática por meio de máquinas.

Como o processamento digital de imagens pode ser utilizado para melhorar a informação para interpretação por máquinas, é possível além de melhorar as imagens,auxiliar no processo de classificação, automaticamente,por intermédio do computador; a área que pode auxiliar nesta fase é a da Inteligência Artificial. Esta surgiu como uma área para provir inteligência às máquinas e estas por sua vez surgiram para automatizar as atividades praticadas por humanos. Dentre as várias áreas dentro da inteligência artificial pode-se destacar o Reconhecimento de Padrões, quebusca identificar padrões para seu posterior reconhecimento.

O LiveMemory (LINS*et al.*, 2009), tomado como base na definição de algumas estratégias da plataforma pLiveMemory,é bastante simplificado. Ele dá acesso apenas aos volumes/edições da SBrT, mas já faz uso de algumas funcionalidades do BigBatch; a interface de processamento de imagens funcionanos modos usuário ou lote, conforme já descritona seção 2.1.1. AFigura 26 apresenta a interface desse sistema.

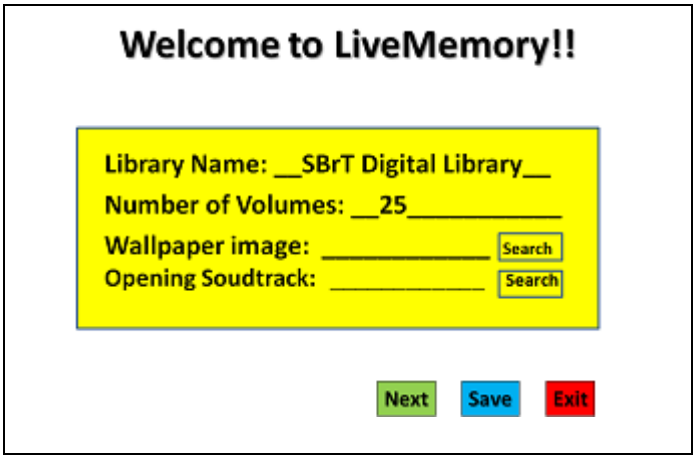


Figura 26–*Interface doLiveMemory (Lins, 2009).*

A plataformapLiveMemory, objeto definido e estudadonesta tese,reimplementa assimtécnicas de (LINS*et al.*,2009) e adicionanovos filtros à sua interface (redefinida e ampliada no contexto do sLiveMemory). A base de documentos do SBrT, digitalizada no LiveMemory, foi utilizada para ilustrar verificar e montar a base real hoje dopLiveMemory. AFigura 27exemplifica a aplicação de um filtro que transforma a imagem para tons de cinza e naFigura 28 a imagem gerada foi transformada para preto e branco com o uso do algoritmo de limiarização.

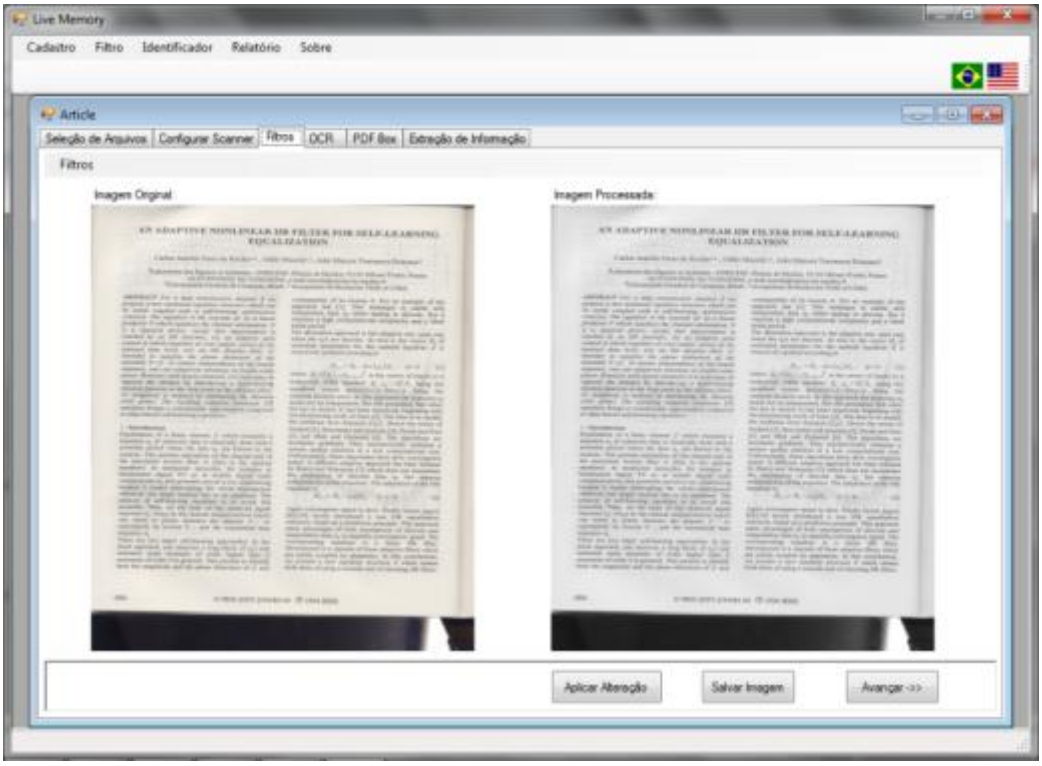


Figura 27–*sLiveMemory: aplicação de filtro para transformar em tons de cinza.*

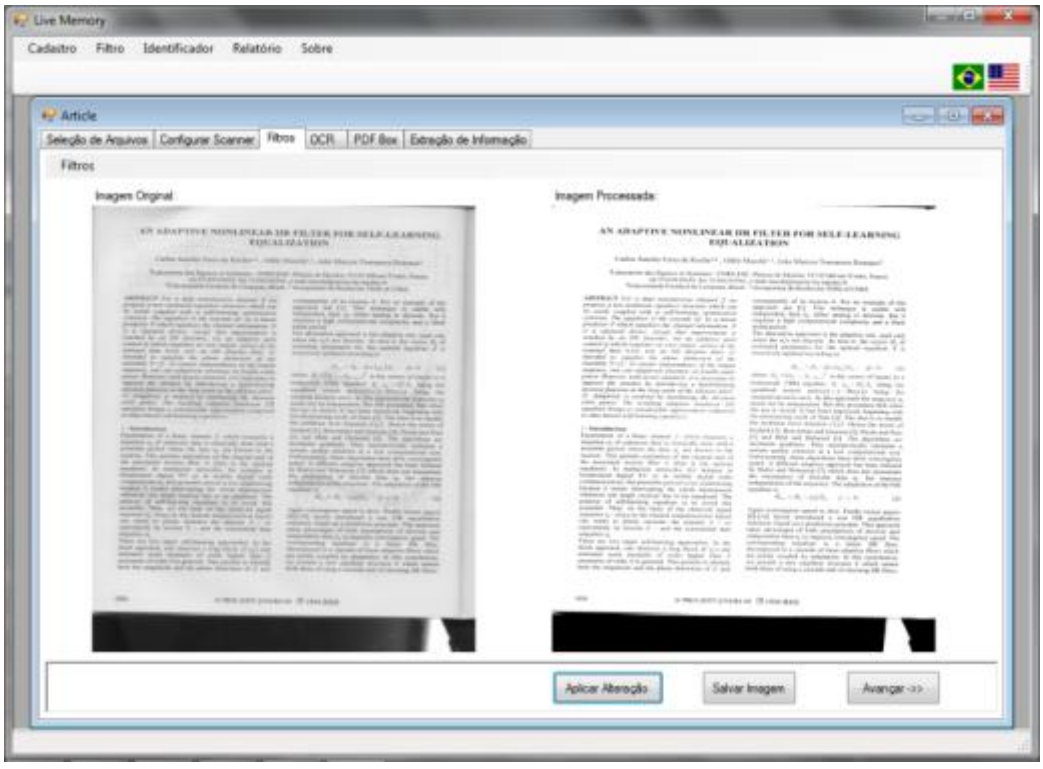


Figura 28–*sLiveMemory: aplicação de filtro para transformar para preto e branco.*

4.2 Modelagem de Dados e Registro de Documentos

O desenvolvimento da plataforma pLiveMemory permitiu,além do gerenciamento dos documentos das bibliotecas, um mecanismo de indexação automática e armazenamento em base de dados dos seguintes tópicos: títulos dos documentos, ano de publicação, autores e palavras-chave. Para o desenvolvimento deste ambiente foi utilizado o Microsoft© Visual C# juntamente com o banco de dados MySql.

No sistema sLiveMemory, para cadastrar uma biblioteca o usuário deve, primeiramente, informar o nome da biblioteca, a quantidade de volumes que a compõem, a imagem que será utilizada como papel de parede e por fim o arquivo de música. Este arquivo deverá ser executado ao inicializar o CD ou DVD, caso o usuário queira gravar um CD/DVD com os artigos de uma determinada edição de uma conferência e distribuí-los, como vem fazendo a SBrT em seus últimos eventos. A Figura 29 exibe a tela do sLiveMemory para cadastrar informações de bibliotecas, como por exemplo, a biblioteca da SBrT.

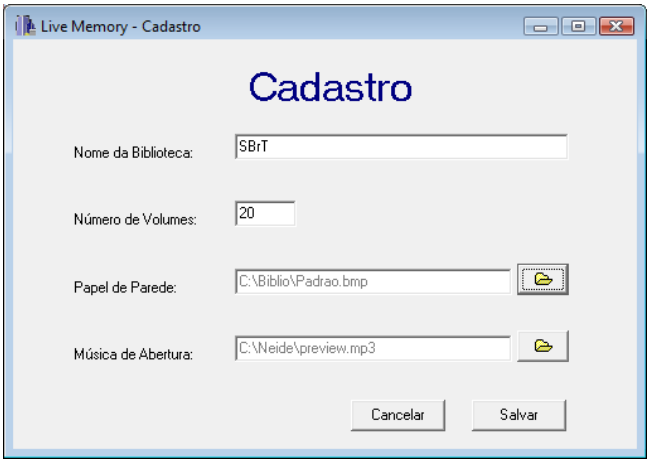


Figura 29– *Tela do sLiveMemory para Cadastrar Informações de Biblioteca.*

Após alguns experimentos a interface para *desktop*foi redefinida e ampliada.Foram acrescidos mais módulos de gerenciamento de bibliotecas.

Ainterface do sLiveMemory é dividida em trêsgrandes categorias:

- Pré-cadastro;
- Cadastro e extração de informações de artigos;
- Ferramentas auxiliares.

O pré-cadastro é composto pelos cadastros que deverão ser feitos antes do cadastrodos artigos, tais como: área, subárea, configuração dos *templates*, conferência, edição etc. O cadastro dos artigos consiste de tratamento para extração e armazenamento das informações que compõem os artigos. A Figura 30, exemplifica a interface principal do sistema sLiveMemory.O que foi denominado de ferramentas auxiliares é formado por: relatórios, identificador de idioma e filtros de processamento de imagem.

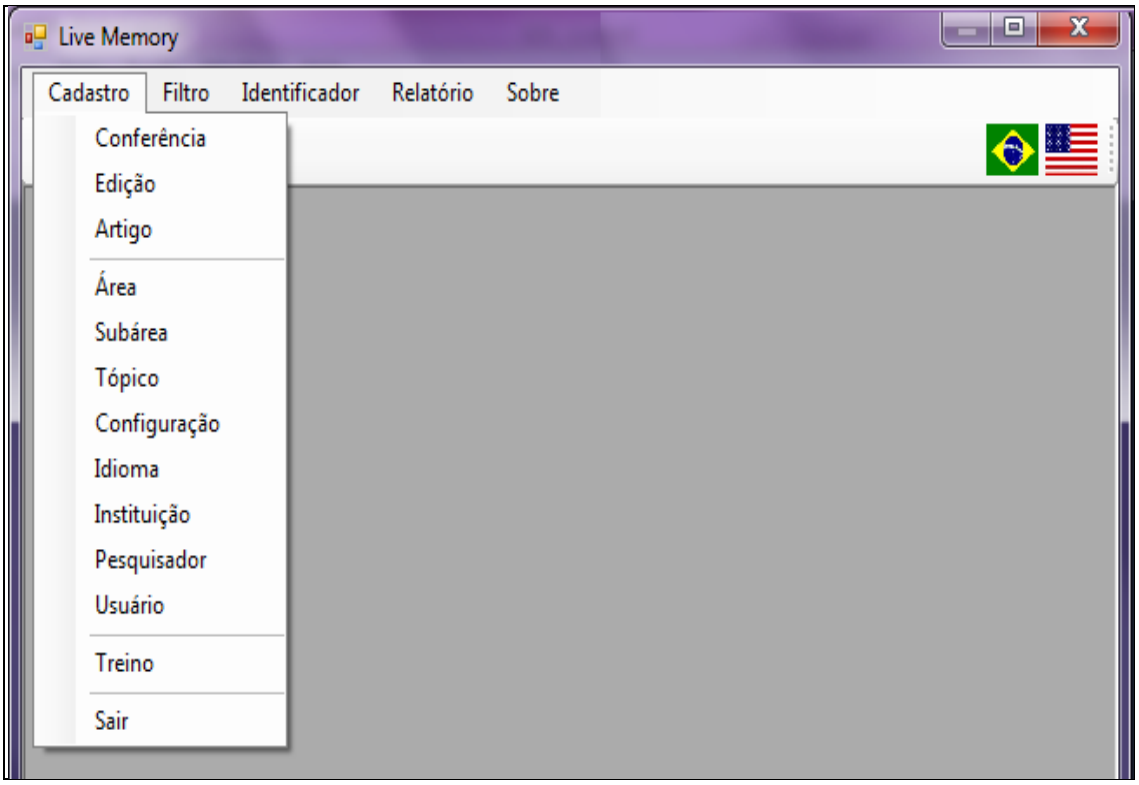


Figura 30 – *Interface Principal do sLiveMemory*.

No *menu* Identificador o usuário pode selecionar o ambiente para cadastrar as classes de palavras, veja Figura 31, utilizadas para identificação do idioma dos artigos e estas também são utilizadas no processo de identificação de palavras-chave, conforme descrito na seção 4.4. Outra opção deste *menu* é o ambiente de configuração e identificação do idioma, a variante do algoritmo de Lins*et al.* (2004) descrito em (CABRAL *et al.*, 2012) foi usado para identificação do idioma o que resultou em 100% de acerto. A Figura 32 exibe a referida interface.

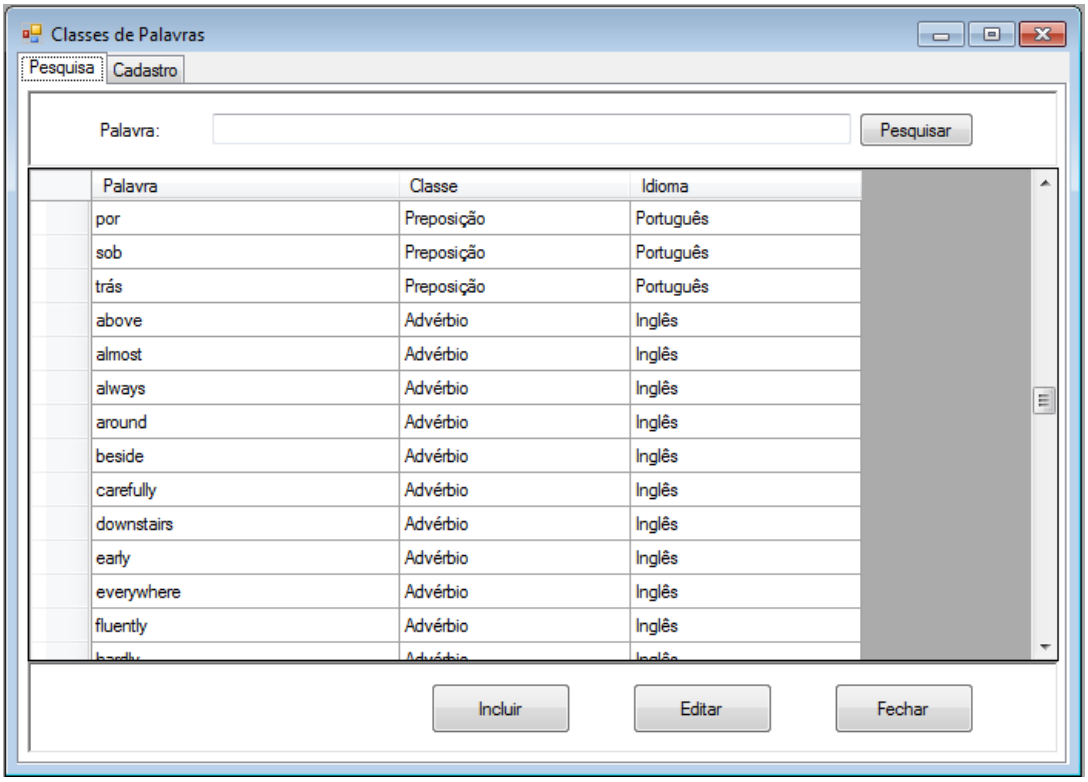


Figura 31– *Interface para cadastrar Classe de Palavras.*

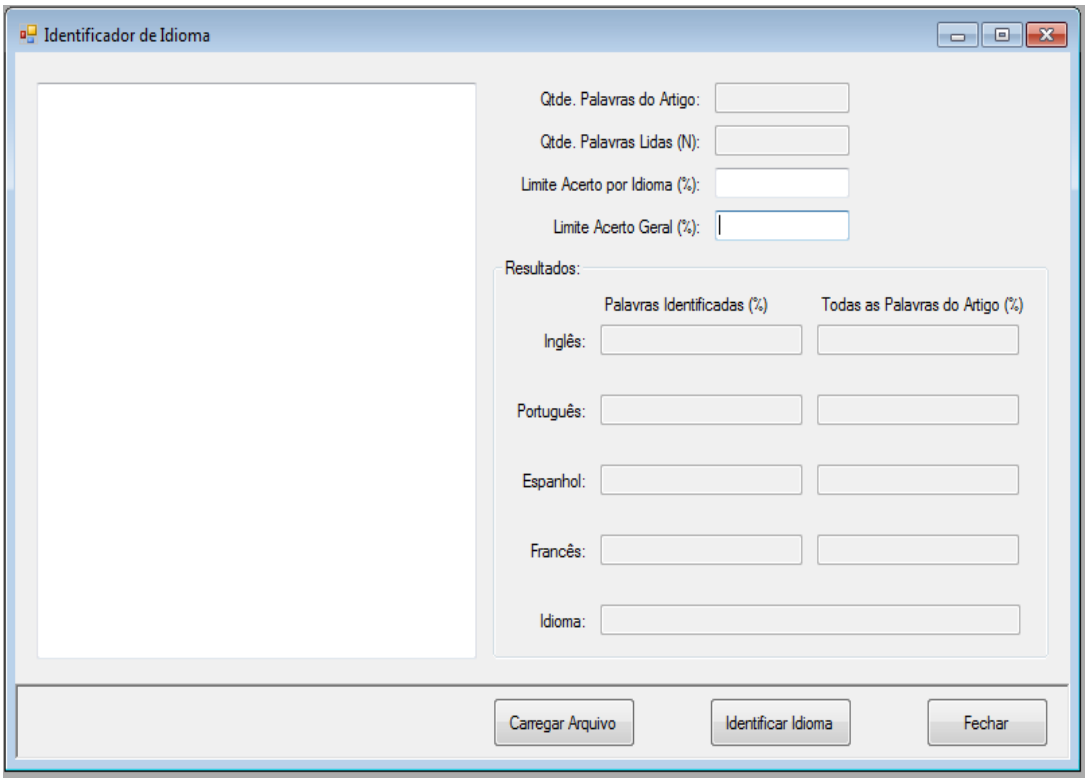


Figura 32– *Interface para Configurar e Identificar Idioma.*

Após os cadastros e configurações, o ambiente está preparado para o processo de leitura dos artigos e extração dos seus dados, para posterior armazenamento na base de dados.No processo de extração das informações o sistema sLiveMemory está preparado para trabalhar tanto com arquivos no formato de imagem (JPEG, PNG e TIFF) quanto no formato PDF.

Durante os experimentos foi utilizado para geração dos textos: (1) nas imagens o OCR Tesseract (2011), uma ferramenta de código aberto originalmente desenvolvida pela HP e distribuída pelo Google; (2) nos s arquivos em PDFfoi aplicada o *software* PDFBox(2011), uma ferramenta Java de código aberto que trabalha com documentos PDF, permitindo a criação de novos documentos, manipulação de documentos existentes eextração de conteúdo.

Quando os arquivos estavam no formato PDF os resultados da extração foram mais significativos, ou seja, houve menos interferência da ferramenta de conversão para texto, assim como foi possível obter as informações referentes ao formato do texto, então optou-se por transformar todas as imagens para PDF editável e assim foram extraídas informações do PDF, como o tamanho dos caracteres em cada linha, isto facilitou a identificação do título do artigo, por exemplo.

De posse do texto extraído o sistema busca as informações para armazenar na base de dados. As informações podem ser divididas em dados da primeira página do artigo e as referências bibliográficas.As Figuras33 e 34 ilustram as telas da interface do sistema sLiveMemory para extração de informações de um artigo, de uma edição específica.

A Figura 33exibe, especificamente,a aba artigo que apresenta a extração de informações de um artigo. Os números em destaque na figura indicam:

- 1 –área com informações gerais (código, edição, idioma e nome do arquivo);
- 2 – informações sobre primeira e última página do artigo;
- 3 –título do artigo;
- 4 –relação de todos os autores do artigo;
- 5 –relação das instituições dos autores;
- 6 –relação dos *emails* dos autores;
- 7 –resumo em Inglês e/ou Português;
- 8 –palavras-chave em Inglês e/ou Português;
- 9 –botões para cancelar ou salvar, no banco de dados, as informações que foramextraídas.

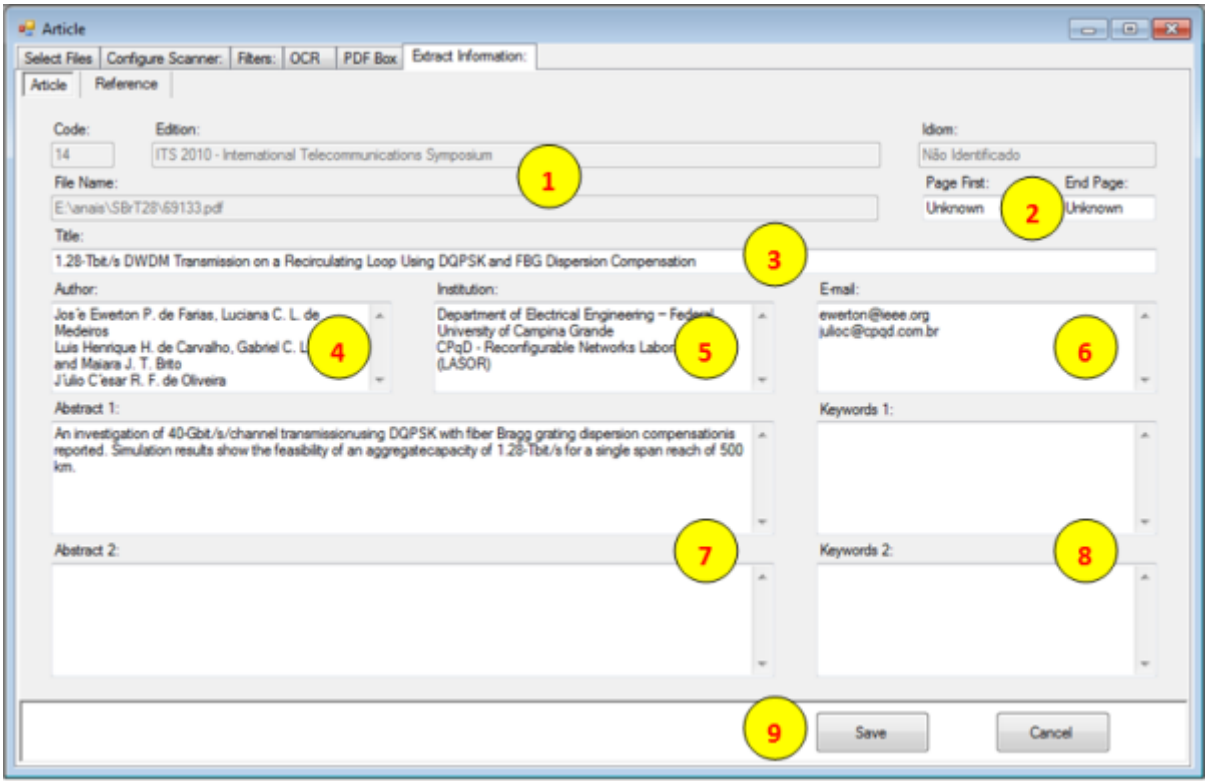


Figura 33 – *Tela de extração de informações: aba Artigo.*

Para classificar as referências utilizadas nos artigos, optou-se por dividi-las em cincopartes, a saber:

- Título – indica o título da referência, o qual pode estar em negrito, itálico, entre aspas ou sem destaque, ou seja, em formato normal;
- Autor – indica os autores das referências;
- Ano – indica o ano de publicação da referência;
- Local – indica o local onde a referência foi publicada ou editada;
- Outros – indicam as demais informações sobre referências, no caso de artigo, este campo pode conter o nome do evento, número DOI etc.

A Figura 34exibe a aba referência, na qual são extraídas e classificadas as referências bibliográficas dos artigos. Os números em destaque na figura indicam:

- 1 – a relação de todas as referências, de um determinado artigo, a serem tratadas;
- 2 –todos os títulos encontrados;
- 3 – todos os autores encontrados;
- 4 – todos os locais encontrados;

- 5 – todas as informações classificadas para o campo Outros;
- 6– todos os anos encontrados;
- 7 – botões para cancelar ou salvar, no banco de dados, as referências que foram extraídas.

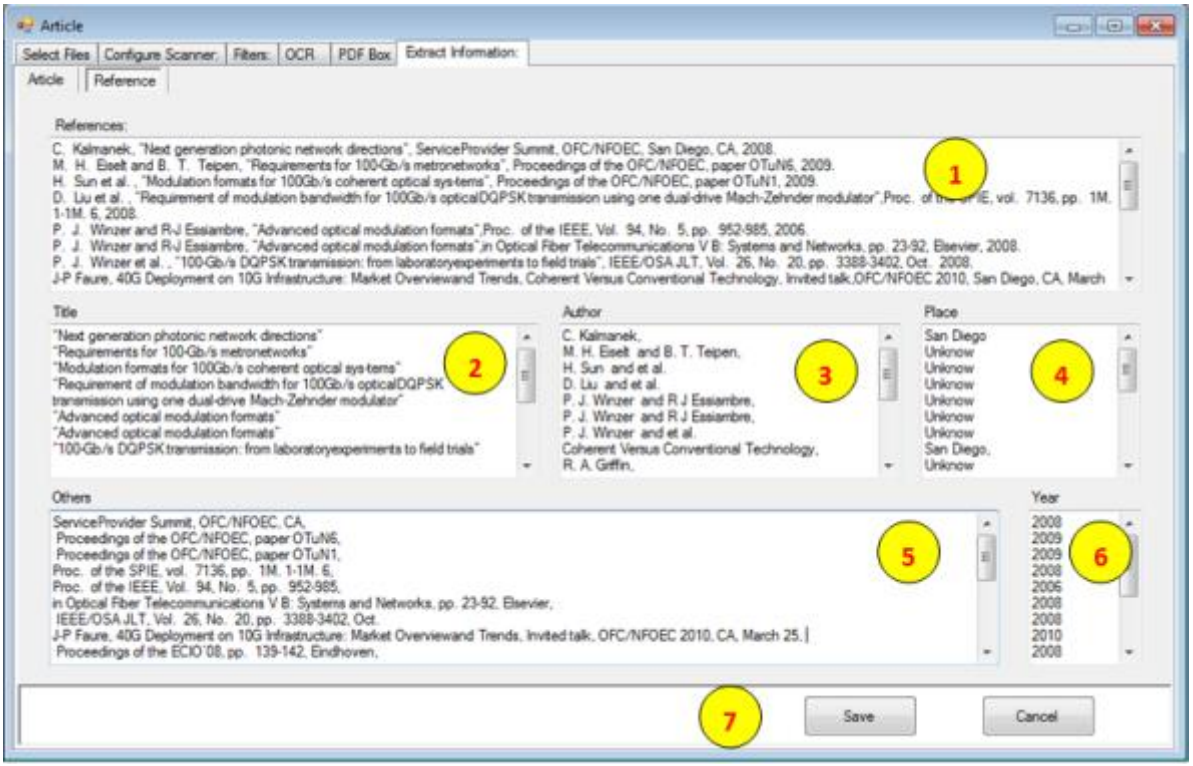


Figura 34 – *Tela de extração de informações: aba Referências.*

Para o processo de identificação e extração dos dados algumas estratégias foram adotadas. Primeiramente, foi utilizadaa técnica de Expressões Regulares, como as exemplificadas na Figura 5, juntamente com as informações cadastradas no ambiente de configuração dos *templates* (quantidade de página por artigo, início da linha do título etc.).

Os resultados para extrair as informações da primeira página do artigo foram bons, mas houve a necessidade de ferramentas mais elaboradas para a extração das informações das referências bibliográficas. Devido as suas peculiaridades a extração das referências está descrita em detalhes na próxima seção.

Após a extração das informações os dados foram armazenados no banco de dados MySql, o que possibilita a consulta das informações via relatório ou via página do Google Sites©, este ambiente está descrito no Capítulo 5.

4.3 Aprendizagem e Extração de Referências Bibliográficas

As referências bibliográficas auxiliam na contextualização de um dado assunto, pois permitem a visualização dos "caminhos" de uma dada pesquisa. Os artigos da SBrT, por exemplo, estão repletos de referências e autorreferências aos eventos anteriores da própria SBrT, desta forma, a extração das referências foi mais que necessária. Então, estas foram extraídas e divididas em cinco partes: título, autor(es), ano, local (cidade, estado ou país) e outros (os demais textos).

A decisão em dividir as informações extraídas em apenas 5 partes, deve-se ao fato de acreditamos que as informações primordiais são *título* e *autor*; também considerou-se que *ano* e *local* são informações que podem auxiliar e o campo *outros* contém as demais informações. Portanto, as estratégias utilizadas, para extrair as referências bibliográficas,podem ser divididas em duas:

- 1 –Expressões Regulares como em Álvarez (2007) e Ohta*et al*. (2008); e
- 2 – Técnicas de ClassificaçãoAutomática: algoritmo K-NN juntamente com Distância Euclidiana e Similaridade do Cosseno, algoritmo *Naïve Bayes*.

Para a extração optou-se, inicialmente, pelas expressões regulares porque estas podem extrair dados, desde que estes apresentem uma certa regularidade no formato. A princípio, as referências possuem padrões, porém ao longo dos testes percebeu-se que nem sempre os padrões são seguidos e que há variedade nos padrões. Então decidiu-se usar técnica de classificação automática.

Dentre as técnicas de classificação automática existentes foram testadas o algoritmo K-NN e o *Naïve Bayes*, pois ambos precisam de uma base para treinamento e trabalham com probabilidades.

O processo de extração é apresentado na Figura 35, a entrada do processo inclui arquivos em formato PDF editável e a saída é o conjunto de todas as referências presentes no artigo.

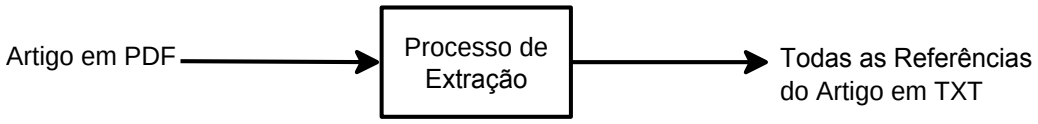


Figura 35–*Processo de Extração das Referências.*

O início do processo tanto para estratégia 1 quanto para estratégia 2, inclui os seguintes passos:

- Transformar os arquivos no formato PDF para o formato TXT, por meio do *software*PDFBox;
- Identificar no texto o local das referências:
  - Primeiramente, busca-se a palavra “reference” ou as combinações desta, tanto em português quanto em inglês, considerando a palavra no singular ou plural. Ao encontrar a palavra, caso esta não esteja no padrão, a mesma é substituída pelo padrão “reference”,
  - Análise para determinação do início das referências: varre-se todo o texto, identificando o local onde a palavra padrão “reference” está verificando se é início de frase e se a mesma está sozinha ou acompanhada de mais uma palavra, como no caso de referências bibliográficas;
- Aplicar, a partir da área que contém as referências, técnicas de expressão regular para extrair todas as referências existentes, usando como padrão os termos:
  - [número],
  - frases que contenham texto entre aspas, ou
  - textos que finalizem com um ano.

Após a preparação do texto é possível aplicar as estratégias e estas estão descritas nas próximas subseções.

4.3.1 Estratégia de Extração com Expressões Regulares

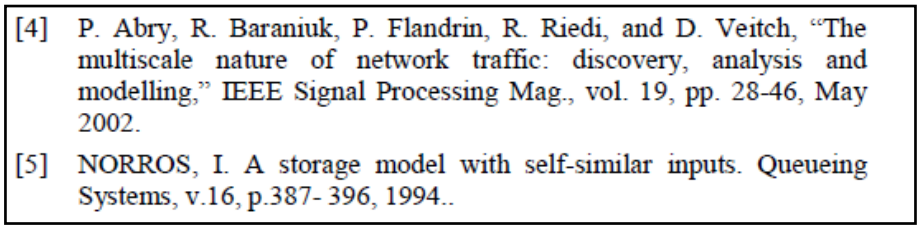
A primeira estratégia é baseada em expressões regulares, os testes mostraram que quando o título de uma referência bibliográfica é encontrado, as outras características são, normalmente, identificadas com sucesso.

Para cada referência identificada, é feita a extração dos seguintes elementos: autor, título, local, ano e outros elementos. O processo de extração com as expressões regulares é o seguinte:

- Primeiro, há uma busca pelo título entre aspas ou a separação entre pontos, neste caso o texto de maior tamanho é considerado o título. A Figura 36 exemplifica dois casos de sucesso, sendo que na primeira referência o autor

utiliza o padrão entre aspas e para a segunda, o mesmo autor, utiliza o padrão entre pontos para separar os termos da referência;

- Após a identificação do título o texto que vier antes deste é considerado o texto que contêm o autor.Como o campo autor geralmente tem mais de um autor, este campo deve ser fragmentado, considerando: nome e sobrenome. O padrão buscado pode conter:
  - "sobrenome + letra + ponto", ou
  - "letra + ponto + sobrenome", ou
  - "nomes sem abreviação";
- As palavras que vêm depois do título serão classificadas como "outras informações", e deste grupo será extraído ano e local. Para o local busca-se o padrão em uma tabela de cidades, estados e países. O texto restante é considerado "outros elementos".


[4] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch, "The multiscale nature of network traffic: discovery, analysis and modelling," IEEE Signal Processing Mag., vol. 19, pp. 28-46, May 2002.

[5] NORROS, I. A storage model with self-similar inputs. Queueing Systems, v.16, p.387- 396, 1994..

Figura 36 – *Identificação de Títulos usando aspas e pontos.*

ATabela 5exibe três exemplos de expressões regulares. O primeiro é utilizado para verificar se um determinado texto está entre aspas, neste caso,desde que o tamanho possua mais de 10 caracteres,o texto será classificado como o título de uma referência. O segundo exemplo serve para identificar anos dentro das referências. O terceiro e último exemplo é utilizado para dividir o texto, quando há: ponto, vírgula ou ponto e vírgula. Para este último exemplo, os textos fragmentados servirão para a fase de classificação automática.

Tabela 5 – *Exemplos de Expressões Regulares.*

Expressão Regular	Significado
["“]+[a-zà-úA-ZÀ-Ú0-9.,;:/&?()– -]+["”]	Verifica se o texto está entre aspas, ou seja, se é título de uma referência
[12][0-9]{3}	Identifica o ano
[.,;]	Utilizado para quebrar texto em fragmentos

A Figura 37 exemplifica um caso de insucesso na identificação do título, pois este é separado por vírgulas e no próprio título também há uma vírgula, ocasionando uma fragmentação no título.

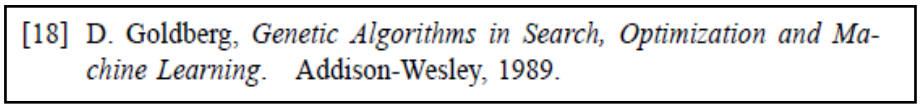

[18] D. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.

Figura 37 – *Referências e Título separados por vírgulas.*

Para evitar redundâncias no armazenamento, caso o título já exista na base de dados, os demais campos são verificados, haja vista que há referências com distinção somente no ano da publicação.

4.3.2 Estratégia de Extração com Classificação Automática

A segunda estratégia é baseada em classificação automática, para essa estratégia as referências são divididas em fragmentos, ou seja, a cada ponto (.), vírgula (,) ou ponto e vírgula (;) o texto é separado e este é chamado de "fragmento de texto".

Cada fragmento de texto é classificado, de acordo com a classificação adotada (título, autor, outros), pelo usuário na etapa de treino ou pelo sistema na etapa de teste. Optou-se por usar apenas três campos nesta fase, sendo que, após a identificação, é possível extrair do campo "outros" as informações referentes ao ano da publicação, assim como o local do evento, por meio da identificação da cidade, estado ou país. Como informado, anteriormente, há uma tabela na base de dados com o registro dos locais de eventos, neste são registrados cidades, estados e nome de países, em diversos idiomas como *Brasil* que também foi registrado com a grafia *z*, ou seja, *Brazil*.

De posse dos fragmentos, foi montado um vetor de características, baseado em Silva (2004), o qual possui 24 campos como descrito naTabela 6. O sistema ao avaliar cada fragmento verifica quais das 24 características o fragmento possui e preenche com valor *1*. Caso a característica avaliada não pertença ao fragmento o campo é preenchido com valor *0*, porém os campos 23 e 24 identificam o comprimento/quantidade de caracteres do fragmento e a posição do fragmento, respectivamente, e, desta forma, não seguem a regra de serem preenchidos com valores *0* ou *1*, ou seja, recebem os valores de acordo com a quantidade de caracteres e a posição no texto.

Tabela 6 – *Vetor de Características.*

Seq.	Campo	Seq.	Campo
1	Número	13	And
2	Ano	14	Mês
3	Vírgula	15	País
4	Ponto e Vírgula	16	Evento: conferência, simpósio
5	Ponto	17	Palavra: available, doc, ps etc.
6	Traço ou Dois Pontos	18	Preposição
7	1st, 2nd, 3rd, 4th, primeiro, segundo, terceiro, quarto	19	Artigo (a, o, as, os, um, uma, an, the etc.)
8	ed., eds., editado	20	Todas Maiúsculas
9	p., pp., pag., pagina, página, page, pg	21	Todas Minúsculas
10	v., vol., volume	22	Iniciais Maiúsculas
11	n., num., no., numero, número	23	Tamanho do Fragmento
12	et al	24	Posição do Fragmento

A

Tabela 7 exemplifica o preenchimento do vetor de característica, onde o fragmento possui apenas as características:

- vírgula (3);
- traço ou dois pontos (6);
- preposição (18);
- tamanho do fragmento (23)- foi preenchido com o valor 53, pois o fragmento possui 53 caracteres;
- posição do fragmento (24) - foi preenchido com o valor 13, pois o início do fragmento, na referência,está na coluna 13.

Os demais campos foram preenchidos com zero, pois o fragmento em análise não possui essas características. Ao final do processo, o usuário classificou o fragmento como “Título”, lembrando que este exemplo foi realizado na etapa de treino e nesta é o usuário que faz a classificação.

Tabela 7 – *Exemplo de preenchimento do Vetor de Características.*

Fragmento					
High-Order Modulation for Optical Fiber Transmission,					
Vetor de Características					
Seq.	**Campo**	**Valor**	**Seq.**	**Campo**	**Valor**
1	Número	0	13	And	0
2	Ano	0	14	Mês	0
3	Vírgula	**1**	15	País	0
4	Ponto e Vírgula	0	16	Evento: conferência, simpósio	0
5	Ponto	0	17	Palavra: available, doc, ps etc.	0
6	Traço ou Dois Pontos	**1**	18	Preposição	**1**
7	1st, 2nd, 3rd, 4th, primeiro, segundo, terceiro, quarto	0	19	Artigo (a, o, as, os, um, uma, an, the etc.)	0
8	ed., eds., editado	0	20	Todas Maiúsculas	0
9	p., pp., pag., pagina, página, page, pg	0	21	Todas Minúsculas	0
10	v., vol., volume	0	22	Iniciais Maiúsculas	0
11	n., num., no., numero, número	0	23	Tamanho do Fragmento	53
12	et al	0	24	Posição do Fragmento	13
Classificação					
Título					

Após o preenchimento do vetor de características o sistema tem mais duas fases, a saber: Treinamento e Testes.

Fase de Treinamento

Para o treinamento, o usuário informa, ao final do preenchimento do vetor de características, o tipo de classificação de cada fragmento. Os seguintes passos são necessários:

- O usuário insere as referências a serem treinadas;
- As partições das referências ou fragmentos são geradas após o sistema encontrar os sinais de pontuação: ponto, vírgula, ponto e vírgula;
- Cada partição ou fragmento é analisadopelo sistema e este preenche o vetor de características;
- Em seguida, o usuário classifica cada fragmento com um dos seguintes rótulos: título, autor ou outras informações;

- Finalmente, o sistema analisa a base de treinamento gerando um vetor de média de cada elemento (título, autor e outras características), ou seja, três vetores são gerados, estes serão utilizados na fase de testes.

A Figura 38exibe um esquema da fase de treinamento, desde a entrada das referências, passando pela fragmentação, preenchimento do vetor de características, rotulação de cada vetor, cálculo egeração dos vetores de média, estes servirãode entrada para a fase de testes.

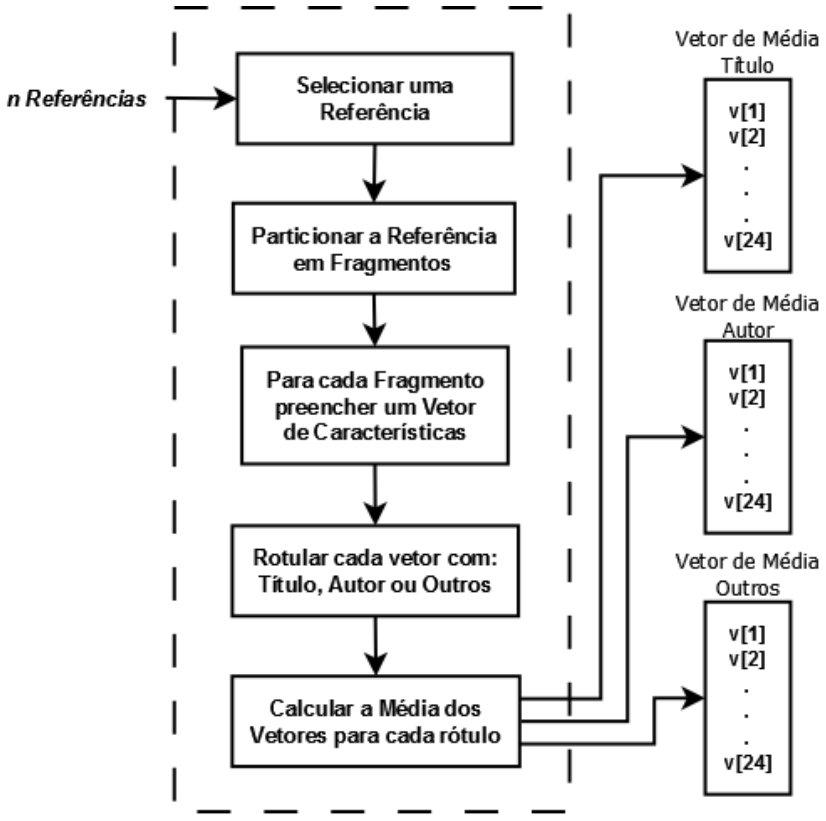


Figura 38 – *Fase de Treinamento e Geração dos Vetores de Média.*

A Figura 39 mostra a interface de treinamento, vale salientar que esta pode ser considerada como um módulo a parte do sLiveMemory. Os números em destaque na figura indicam:

- Item 1 –todas as referências a serem tratadas;
- Item 2 –a referência que está sendo tratada;
- Item 3 –elemento que está sendo avaliado, neste caso, o décimo segundo;
- Item 4 –o fragmento avaliado;

- Item 5 –a classificação informada pelo usuário do sistema, neste exemplo, *Ano*;
- Item 6 –o vetor de 24 características do fragmento a ser tratado;
- Item 7 –botões.

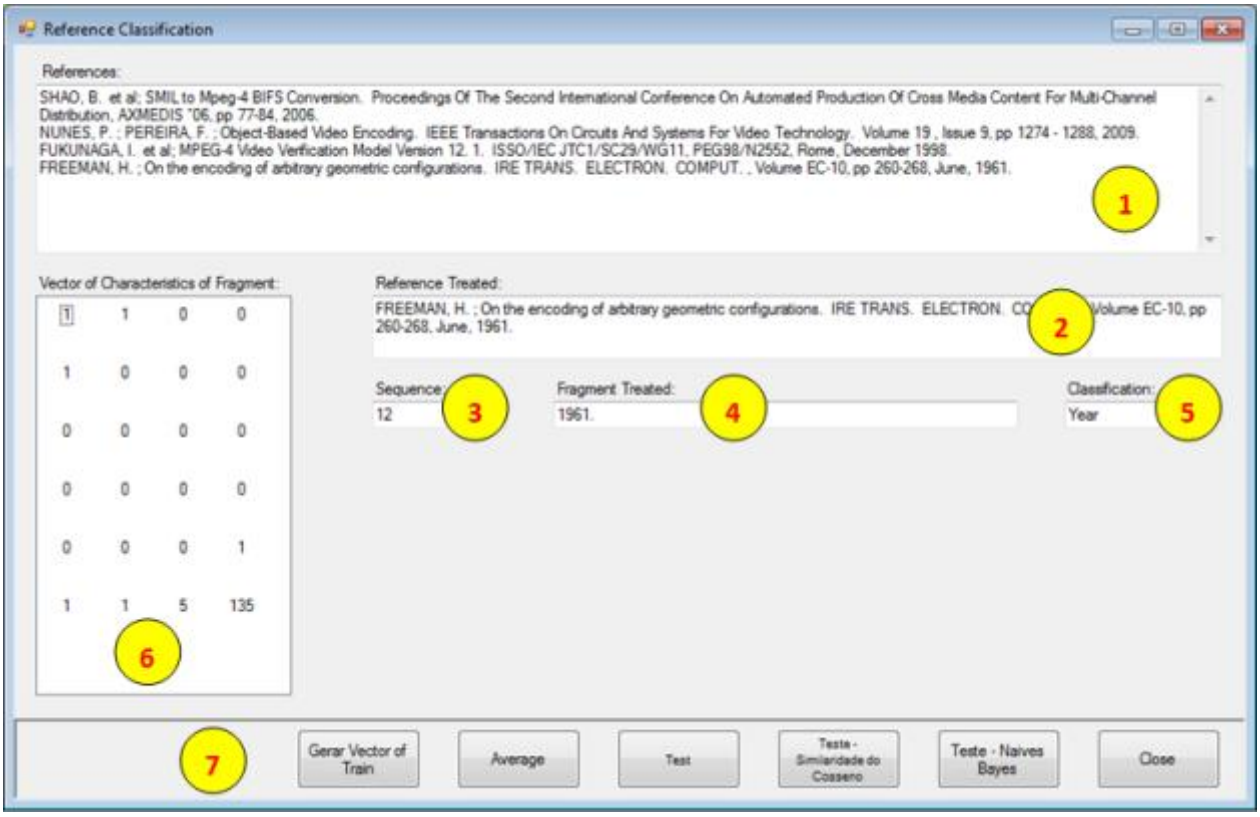


Figura 39 – *Interface de Treinamento.*

Fase de Teste

A fase de teste, inicialmente, é semelhante àfase de treinamento, mas a classificação de elementos é feita, automaticamente, pelo sistema. Para cada fragmento, o sistema compara os valores do vetor de característica gerado com os vetores de características obtidos na fase de treinamento. Para a comparação entre os vetores, três estratégias foram adotadas:

- Algoritmo K-NN com Distância Euclidiana;
- Algoritmo K-NN com Similaridade do Cosseno;
- Algoritmo*Naïve Bayes*.

Vale destacar que o algoritmo K-NN (K Nearest Neighbor), onde *k* indica a quantidade de vizinhos mais próximos, é o método de aprendizagem baseado em instâncias

mais elementar. O K-NN assume que todas as instâncias correspondem a pontos em um espaço n–dimensional.Para utilizar o K-NN é necessário:

- Um conjunto de exemplos de treinamento;
- Definir uma métrica para calcular a distância entre os exemplos de treinamento, neste trabalho as métricas adotadas foram Distância Euclidiana e Similaridade do Cosseno;
- Definir o valor de *K* (o número de vizinhos mais próximos que serão considerados pelo algoritmo);

Para classificar com o algoritmo K-NN são necessárias mais três etapas, conforme descrito a seguir:

- Calcular a distância entre o valor desconhecido e os outros valores do conjunto de treinamento;
- Identificar os *K* vizinhos mais próximos;
- Utilizar o rótulo da classe dos vizinhos mais próximos para determinar o rótulo de classe do exemplo desconhecido (votação majoritária).

A escolha do *K* não é tão simples, pois se *K* for muito pequeno, a classificação fica sensível a pontos de ruído, mas se *K* é muito grande, a vizinhança pode incluir elementos de outras classes.O cálculo da Distância Euclidiana entre dois pontos (*p* e *q*) é definido pela equação E1.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad \text{(E1)}$$

A expressãomatemática E1 é na verdade a aplicação de *n*vezes do Teorema de Pitágoras ($d^2 = p^2 + q^2$). Lembrando que neste trabalho *p* e *q* são os vetores de características de cada fragmento classificado.

Já o cálculo da Similaridade do Cosseno usa o ângulo entre dois vetores de *n* dimensões como medida de similaridade, esta é máxima quando os vetores apontam na mesma direção, ângulo de 0°, ou seja, valor 1 e mínima quando forem perpendiculares, isto é ângulo de 90°, ou melhor, valor 0. O cálculo da Similaridade do Cosseno entre dois

pontos (*x* e *y*) ou vetores, neste trabalho *x* e *y* são os vetores de características de cada fragmento,e é definido pela equação E2.

$$cos(x,y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_i^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2}\sqrt{\sum_{i=1}^n y_i^2}} \quad \text{(E2)}$$

Além do algoritmo K-NN, foi utilizado o algoritmo *Naïve Bayes*que é um classificador probabilístico, pois precisa de um conjunto de probabilidades e estas são estimadas pela contagem da frequência de cada valor decaracterística para as instâncias dos dados de treinamento. Dada uma nova instância, oclassificador estima a probabilidade dessa instância pertencer a uma classeespecífica, baseada no produto das probabilidades condicionais individuais para osvalores característicos da instância.

O cálculo exato utiliza o Teorema de *Bayes* e é poressa razão que o algoritmo também é denominado Classificador de *Bayes*, também pode ser chamado ainda de *Naïve*, uma vez que todos os atributos são condicionalmente independentes, dado o valor da variável da classe. O Teorema de *Bayes* é representado pela equação E3:

$$P(Y/X) = \frac{P(X/Y)P(Y)}{P(X)}, \quad \text{(E3)}$$

onde:

- P(*Y*/*X*) – probabilidade condicional para a variável *Y* dado *X*;
- P(*X*/*Y*) – probabilidade condicional para a variável *X* dado *Y*;
- P(*Y*) – probabilidade de ocorrer *Y*;
- P(*X*) – probabilidade de ocorrer *X*.

Inicialmente, na fase de treinamentofoi testadaa proximidade dos vizinhos utilizando Distância Euclidiana, esta é utilizada para verificar a semelhança entre o vetor de fragmentos com o vetor de médias. O mesmo foi feito usando o algoritmo K-NN com o cálculo da Similaridade do Cosseno e, finalmente, o algoritmo *Naïve Bayes* foi testado.A Figura 40exibe a entrada (conjunto de *n* referências) e a saída (classificação dos fragmentos) da fase de teste.

Depois dos testes com as três abordagens, observou-se que o algoritmo K-NN com Distância Euclidiana e Similaridade do Cosseno obteve desempenho equivalente, ou

seja,para os mesmos textos os resultados eram semelhantes. Em alguns, casos houvemelhoria na extração do título, mas parte dos autores continuou a ser classificado como “outras informações”, não obtendo desta forma um resultado satisfatório.

Testes utilizando o algoritmo *Naïve Bayes*, que considera o resultado do elemento precedente, mostrou melhores resultados, como descrito na seção 4.3.3.

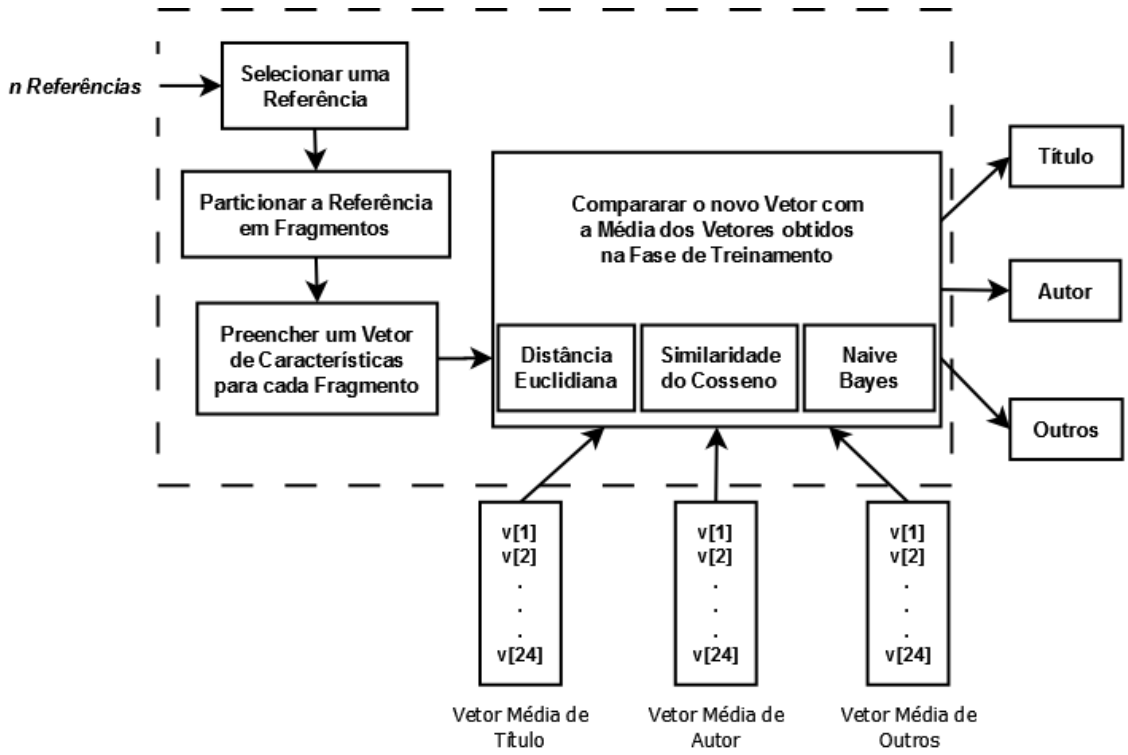


Figura 40 – *Fase de Teste das Referências Bibliográficas.*

A Figura 41 mostra a interface de teste, esta é semelhante a interface de treinamento, porém há os campos de resultados. Os itens específicos desta interface são:

- Item 1 –o texto que foi classificado como título;
- Item 2 –o texto que foi classificado como autor;
- Item 3–o texto que foi classificado como outras informações.

Quando este processo é utilizado no cadastro das referências, como já descrito anteriormente, é necessário quebrar ou extrair do campo “outros” as informações sobre ano e local, por meio de expressões regulares.

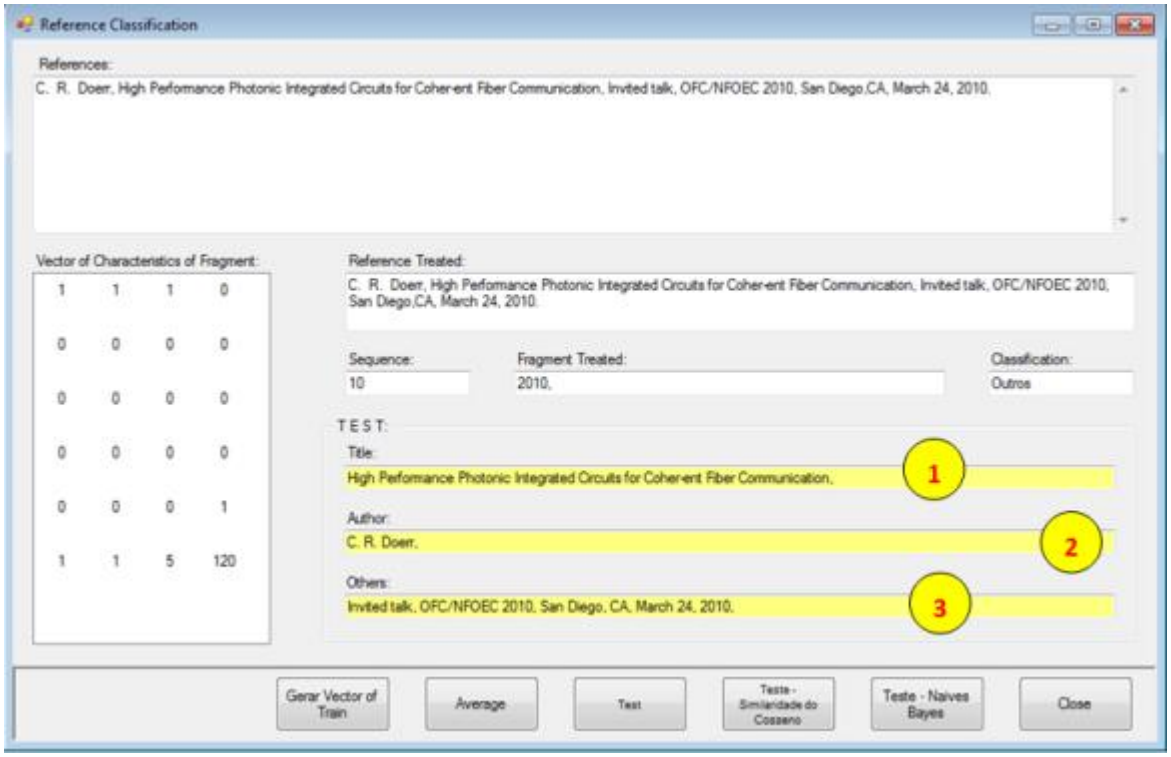


Figura 41 – *Interface de Teste.*

4.3.3 Avaliação dos Algoritmos e Resultados

Nesta seção será descrito o processo de validação do ambiente proposto para extração de referências, bem como os resultados obtidos.

Para validar o processo proposto foi utilizado os artigos da SBrT publicados em 2010 (LINS*et al.*, 2010). Os artigos estão no idioma Inglês, pois o evento é o ITS (*International Telecommunications Symposium*)uma publicação internacional da SBrT, que acontece a cada 4 anos. O formatodos arquivos estava em PDF editável.

Para a fase de treinamento, 16 artigos foram utilizados, eapartir destes 98 referências foram extraídas e particionadas em 892 fragmentos. Cada fragmento foi classificado pelo usuário, sendo que 363 fragmentos foram classificados como título, 417 como autor e 112 foram classificados como outros.A Figura 42 demonstra esquematicamente a divisão dos 892fragmentos.

A Tabela 8 mostra como cada campo, do vetor de características, influenciou na classificação, como, por exemplo, o campo 2 (ano) este foi encontrado nas classificações de título e outros, mas não foi encontrado em autor. O campo 9 (página) foi encontrado 100% em outros, da mesma forma o campo 12 (et al) foi encontrado somente nos fragmentos classificados como autor.

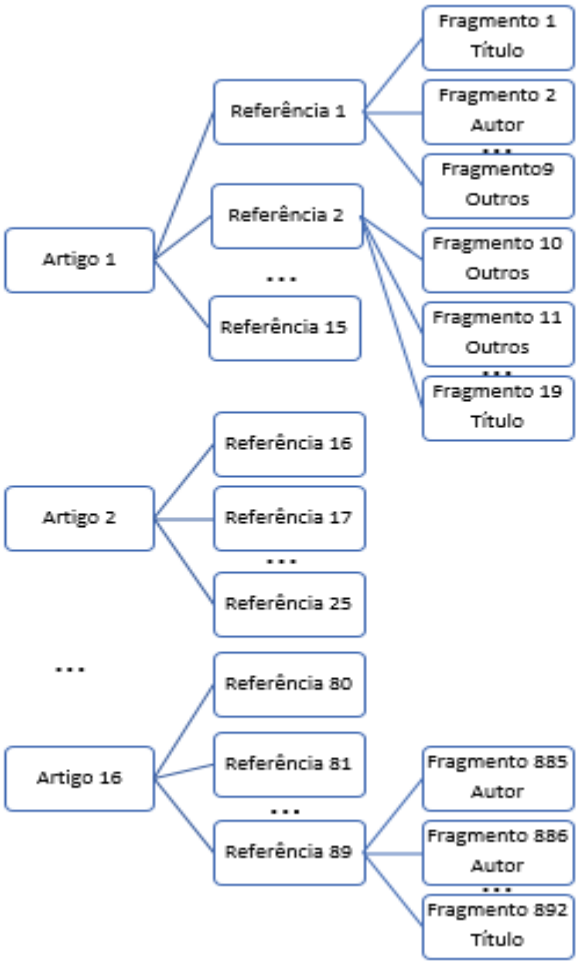


Figura 42 – *Fragmentos encontrados na Fase de Treinamento.*

Tabela 8 – *Vetor de Características.*

Seq.	Campo	Classificação (%)			Total
		Título	Autor	Outros	
1	Número	2,33	0,00	97,67	129
2	Ano	1,37	0,00	98,63	73
3	Vírgula	19,60	26,40	54,00	250
4	Ponto e Vírgula	50,00	33,33	16,67	6
5	Ponto	6,69	48,02	45,29	329
6	Traço ou Dois Pontos	35,94	3,13	60,94	64
7	1st, 2nd, 3rd, 4th, primeiro, segundo, terceiro, quarto	0,00	0,00	100,00	9
8	ed., eds., editado	0,00	0,00	100,00	9
9	p., pp., pag., pagina, página, page, pg	0,00	0,00	100,00	8
10	v., vol., volume	0,00	0,00	100,00	7
11	n., num., no., numero, número	33,33	0,00	66,67	3
12	et al	0,00	100,00	0,00	1
13	And	40,00	33,33	26,67	30
14	Mês	20,00	15,00	65,00	20

Seq.	Campo	Classificação (%)			Total
		Título	Autor	Outros	
15	País	0,00	0,00	100,00	1
16	Evento: conferência, simpósio	0,00	0,00	100,00	5
17	Palavra: available, doc, ps etc.	0,00	0,00	0,00	0
18	Preposição	81,63	0,00	18,37	49
19	Artigo	71,43	0,00	28,57	7
20	Todas Maiúsculas	0,00	56,20	43,80	242
21	Todas Minúsculas	2,19	0,00	97,81	137
22	Iniciais Maiúsculas	3,21	49,54	47,25	436
23	Tamanho do Fragmento	-	-	-	-
24	Posição do Fragmento	-	-	-	-

Após a etapa de treinamento foi iniciada a etapa de teste epara avaliar os resultados, desta última, a medida de cobertura foi utilizada. O cálculo considera a relação de cada elemento de referências (título, autor, ano e outros) corretamente retornados sobre o número total de referências, conforme equação 4.

$$Cobertura(elemento) = \frac{C_r}{C_t} \quad \text{(E4)}$$

onde:

- C_r – total de elementos identificados;
- C_t – total de elementos;

Em um primeiro momento, o foco era a identificação de referências, ea estratégia usou 10 artigos de cada ano, estes possuíam 186 referências bibliográficas.Quando as expressões regulares foram utilizadas nos 10 artigos, os resultados foram os seguintes:

- 117 títulos foram identificados do total de 186, o que representa 62,90% dos títulos existentes;
- 101 autores foram identificados, isto é, 54,30% do total;
- 170 anos foram identificados, isto é, 91,40% do total;
- 122 fragmentos foram identificados como outros, ea percentagem era de 65,59% do total.

Após os testes com as expressões regulares estas foram utilizadas juntamente com o algoritmo *Naïve Bayes*, ou melhor, as expressões são utilizadas para selecionar

primeiramente, os títulos que estão entre aspas, caso isso não ocorra, todo o texto é fragmentado e o algoritmo *Naïve Bayes* é aplicado.O resultado da classificação de cada elemento de referência foi a seguinte:

- Para os títulos, 138 foram identificados, isto é, 74,19%. Neste item houve o maior ganhoou melhora nos resultados, em relação ao primeiro experimento, de 11,29%;
- Para os autores, o sistemas identificou 114 ou 61,29%, com ganho, praticamente de 7%;
- Para os anos, identificou-se 173, que é equivalente a 93,01%, com ganho de 1,61%, como o ano possui um formato padrão, somente com as expressões regulares os resultados, já eram bastantes satisfatórios e por este motivo este foi o menor ganho;
- Para o campo outros, 128 foram identificados correspondendo a 68,82%, com ganho de 3,23%.

A Tabela 9 mostra os resultados quando expressões regulares são aplicadas em conjunto ou não com o algoritmo *Naïve Bayes*.

Tabela 9 – *Classificação: Expressão Regular x NaïveBayes.*

Classificação	**Expressão Regular (ER)**		***Naïve Bayes*e ER**		***Ganho***
	Acerto	**(%)**	**Acerto**	**(%)**	**(%)**
Título	117	62,90	138	74,19	11,29
Autor	101	54,30	114	61,29	6,99
Ano	170	91,40	173	93,01	1,61
Outros	122	65,59	128	68,82	3,23

AFigura 43 também apresenta os mesmos resultados, mas de forma gráfica. A partir do gráfico é possível observar que os pontosem azul, Expressão Regular com o algoritmo *Naïve Bayes*, representamamelhor solução.

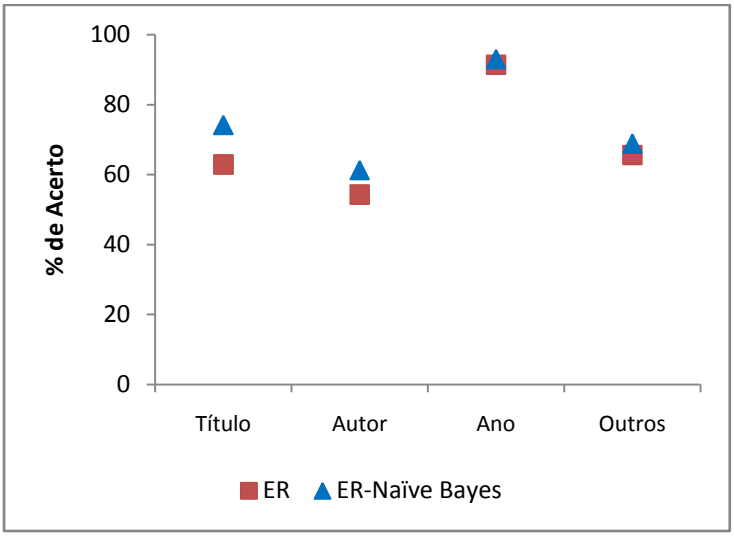


Figura 43 – *Gráfico de Resultados.*

O problema de identificação e extração de referências bibliográficas a partir de documentos digitais não é de resolução trivial, apesar da existência de muitas propostas. Dentre os algoritmosanalisados e testadoso que apresentou melhor resultado foi algoritmo *Naïve Bayes* juntamente com o uso das expressões regulares. As expressões regulares funcionam melhor se os elementos identificados nas referências possuírem um formato regular, como no caso deano. O resultado não é o mesmo quando na expressão é necessária a identificação de título ou autores, que geralmente não tem um formato regular. Mas a integração de ambos permite a obtenção de melhores resultados.

4.4 Identificação de Palavras-Chave

A busca por “palavras-chave” em artigos técnicos pode seguir abordagens diferentes, dependendo da origem do arquivo. Arquivos de imagem e no formato PDFsão totalmente diferentes quando se deseja fazer extração dos seus dados. A extração de conteúdo, em um documento de imagem, é mais difícil, e sujeito a erros, do que em um arquivo PDFeditável porque há as dificuldades do processamento da imagem, enquanto o PDF editável permite a transcrição imediata para o formato TXT, sem contar que há ferramentas, para extração de PDF, que informam o formato do texto, como o tamanho e a posição dos caracteres.

A estratégia adotada neste trabalho é a primeira a trabalhar com arquivos no formato PDF editável e criar um dicionário de palavras-chave, para minimizar os problemas de padronização de palavras-chave em dicionário. Aplataforma pLiveMemory

foi projetada para lidar com documentos em diferentes idiomas, assim, o primeiro passo é identificar a língua em que foram escritos os documentos (no caso do evento permitir multilíngues, como no caso dos artigos da SBrT). A variante do algoritmo de Lins*et al.*(2004) descrito em Cabral *et al.*(2012) como mencionado, anteriormente,foi usado para identificação do idioma.

4.4.1 Busca de Palavras-Chave em arquivos PDF Editável

Arquivos de documento portáteis são, atualmente, um padrãopara distribuição de documentos. Documentos eletrônicoseditáveisem MSWord© ou processador detexto, tais como pdf-texto ou LaTeX, são amplamente aceitos para a apresentação em conferências e também como versão final dos artigos.

Documentos no formato PDF são independentes do dispositivo e oferecem o mesmo layout independentemente de onde será impresso ou visualizado, enquanto um documento no formato MSWord, por exemplo, quando composto em uma máquina e visualizado ou impresso em outra pode ter um *layout*completamente diferente, desta forma os arquivos em PDF são os mais adequados para manipulação ou extração de informação. O processo de extraçãodas palavras-chave,adotado neste trabalho, em arquivo com formato PDFfunciona da seguinte forma:

- Receber como entrada um arquivo PDF editável e aplicar as funções do PDFBox, estas removem a formatação e o layout produzindo como saída um texto “simples”;
- Buscar a palavra “keyword” em Inglês, ou “palavra-chave” em Português ou “palabra-clave” em Espanhol, tanto no singular quanto no plural;
- Verificar se a palavra está no início de um bloco de texto após a palavra “Abstract” (ou “Sumário”, “Resumo”, ou “Resumen”) e seguido da primeira seção do artigo;
- As palavras-chave encontradas são enviadas ao “Dicionário de Palavras-Chave”do idioma correspondente.

Se nenhuma palavra-chave for encontrada no artigo há duas estratégias para tentar encontrar até 6 palavras-chave no artigo, buscandono “Título” e/ou na “Introdução”.

Antes de começar a procurar por palavras-chave no “Título” ou na primeira seção do artigo, algumas palavras que geralmente aparecem em textos, mas que não são centrais

no foco do artigo, devem ser filtradas e colocadas em um “Dicionário de palavras não relevantes”. Essas são chamadas de “palavras auxiliares” ou *stop-word*,importante para a formação da sentença e são usadas para indicar improváveis palavras-chave não candidatas, tais como: “seção”, “can”, “been” etc., frequentes nos artigos, mas não são palavras-chave.

Além dessas palavras frequentes, a estratégia usada por Lins*et al.* (2004)para identificar classes gramaticais (artigos, preposições, conjunções, numerais e advérbios) também foi utilizada não só para descobrir em que idioma o artigo foi escrito, mas também para excluir todas as possíveis palavras-chaveque pertenciam àsclasses gramaticais. A Figura 44sintetiza o processo de extração de palavras-chave.

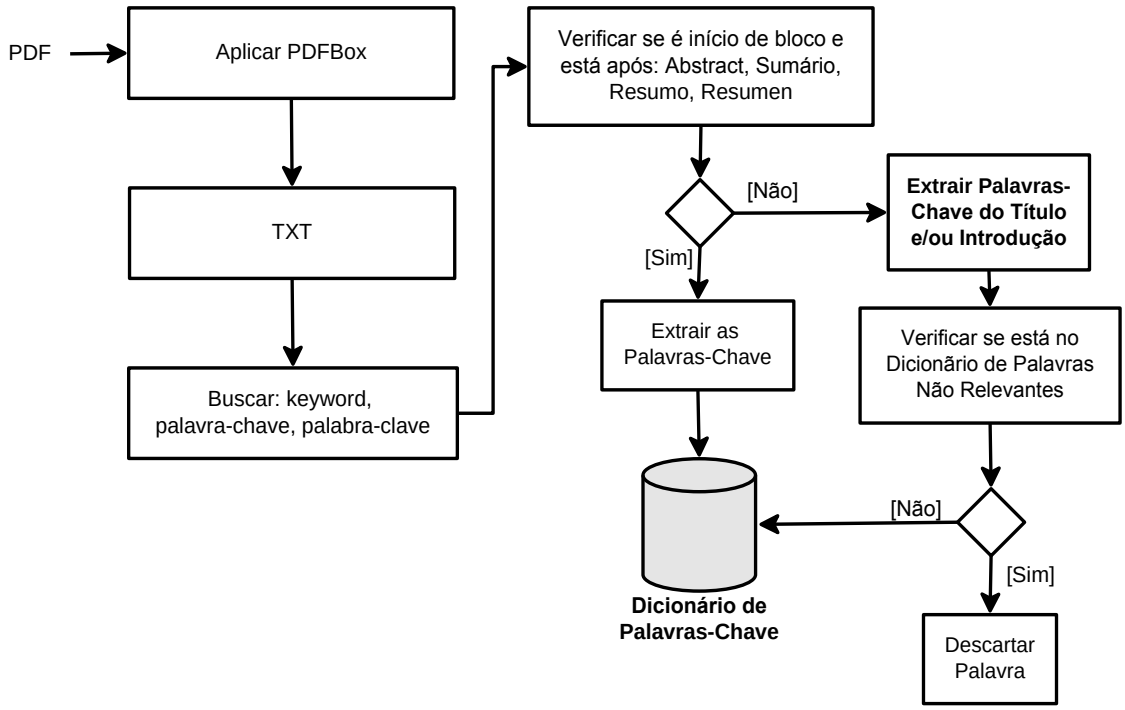


Figura 44 – *Esquema para Extração e Identificação das Palavras-Chave.*

Busca por Palavras-Chave no Título

A busca de palavras-chave no título de um artigo leva em conta os seguintes aspectos:

- Todas as palavras do “Título” são palavras-chave potenciais;
- Cada uma delas é procurada no Dicionário de palavras-chave já construído, verificando sua ocorrência relativa;

- O número de ocorrências é listado em ordem decrescente. Se uma palavra tem frequência maior que 2 no dicionário de palavras-chave, então a palavra é considerada como palavra-chave do documento em análise. O valor da frequência é configurável, e como para os primeiros experimentos não havia muitos dados na base, optou-se por trabalhar com um valor baixo.

Busca por Palavras-Chave na Introdução

Caso tenham sido extraídas menos de 3 palavras-chave do título do artigo, no passo anterior, buscam-se palavras-chave na introdução (no máximo 6,considerando as palavras extraídas do título) ou na primeira seção do artigo, poisse supõe que nessa seção, o autor apresenta a motivação básica do seu artigo e o seu contexto:

- Todas as palavras na "Introdução" são extraídas;
- A partir da lista de palavras da "Introdução"é feita uma consulta ao "Dicionário de palavras não relevantes", então as palavras restantes são pesquisadas no "Dicionário de palavras-chave". Se for encontrada a palavra torna-se uma entrada na lista de candidatas a palavra-chave. Tal lista é organizada por frequência em ordem decrescente, as palavras com frequência maior que 2 tornam-se uma palavra-chave do artigo.

A Figura 45 mostra uma visão geral do ambiente desenvolvido para extração das informações dos artigos. Neste exemplo, mesmo havendo palavras-chave, foi aplicada a estratégia para validar o processo, os campos em destaque são:

- 1 –Título;
- 2 –Introdução;
- 3 –Palavras-chave (definidas pelo sistema);
- 4 –Palavras-chave (definidas pelo autor).

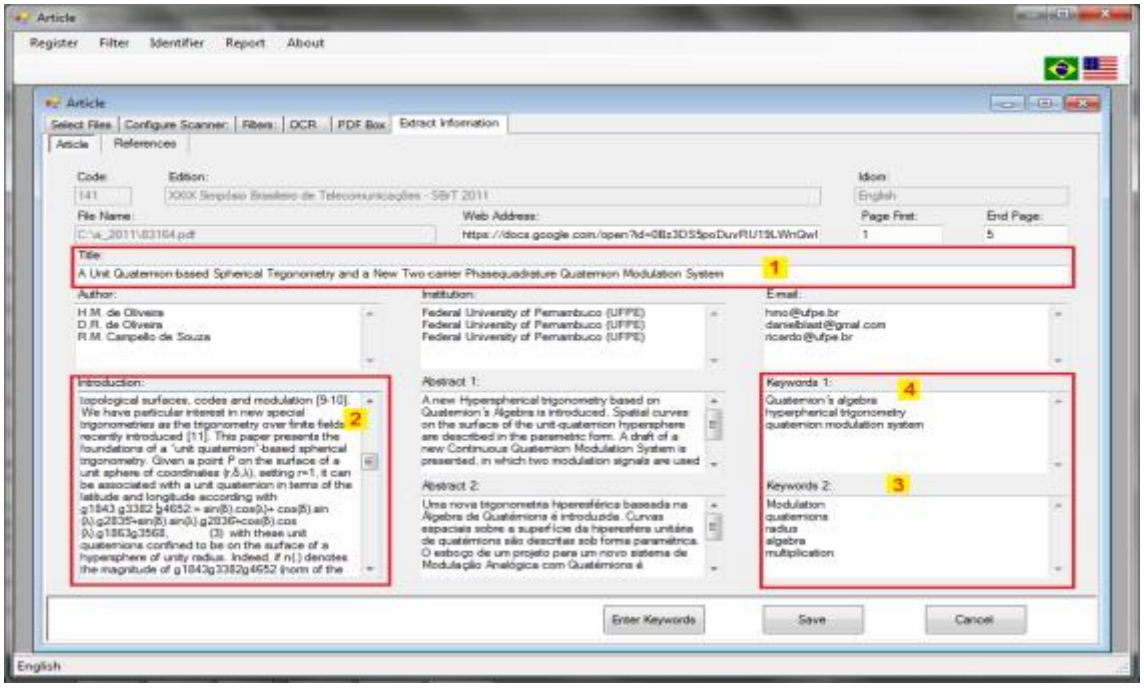


Figura 45 – *Visão Geral do Ambiente de Extração e Identificação das Palavras-Chave.*

4.4.2 Experimentos e Resultados

Para montar a base de palavras-chave foi utilizada as edições da SBrT de 2011, 2010, 2009 e 2008, as quais possuem 714 artigos e destes foram selecionados 110 artigos, sendo que nestes haviam 564 palavras-chave. Após a exclusão das palavras que se repetiram entre os artigos foram armazenadas na base de dados 531 palavras-chave distintas, a Tabela 10exibe a distribuição destes dados.

Tabela 10 – *Geração da Base de Palavras-Chave.*

Edição	Qtde de Artigos	Artigos Utilizados no Treino	Palavras-Chave Geradas
SBrT 2011	203	50	287
ITS 2010	127	50	204
SBrT 2009	196	5	38
SBrT 2008	188	5	35
Total	**714**	**110**	**564**
Total de Palavras-Chave sem Repetição			**531**

Para validar a estratégia proposta dois testes distintos foram realizados e estão divididos e descritos a seguir como caso 1 e caso 2.

Caso 1: Artigos que não continham a seção de Palavras-Chaves

Oito edições da SBrT foram usadas para testar a estratégia proposta. O processo foi semiautomático, pois houve interferência do operador para ajudar na decisão da escolha da palavra-chave, o operador podia excluir uma determinada palavra que não fazia sentido, mas não acrescentava novas palavras.

É possível observar naTabela 11, que exibe os resultados da extração de palavras-chave, que em média 15% dos artigos não possuíam palavras-chave, ou seja, 225 dos 1490 do total de artigos.

Em 2001 o número de artigos sem palavras-chave era bem maior, ou seja, atingia 81% ou 172 dos 212 artigos da referida edição. Acredita-se que neste ano não era comum o uso de palavras-chave.

Na maioria dos casos, das 6 palavras-chave identificadas 2 eram extraídas do título e 4 da introdução. Após o processo 931 palavras-chave foram inseridas, sendo que 414 palavras-chave foram inseridas por meio das palavras que compunham o título dos artigos e 517 das palavras das suas referidas introduções. Também vale ressaltar que, em alguns casos, do título foram extraídas 3 ou mais palavras-chave e, neste caso,não há necessidade de extração de mais palavras-chave da introdução.

Tabela 11 – *Resultados da Extração de Palavras-Chave.*

Edição	Qtde. de Artigos	Qtde. de Artigos semPalavras-Chave	Qtde. de Palavras-Chave extraídas do Título	Qtde. de Palavras-Chave extraídas da Introdução
SBrT 2011	203	4	3	17
ITS 2010	127	21	33	70
SBrT 2009	196	1	1	4
SBrT 2008	188	2	5	4
SBrT 2007	185	1	1	2
SBrT 2006	192	16	24	59
SBrT 2004	187	8	12	23
SBrT 2001	212	172	335	338
Total	**1490**	**225**	**414**	**517**

Na edição de 2001 havia 15 artigos que foram protegidos e, desta forma, o software PDFBox não foi capaz de ler os dados, conforme exemplificado na Figura 46.

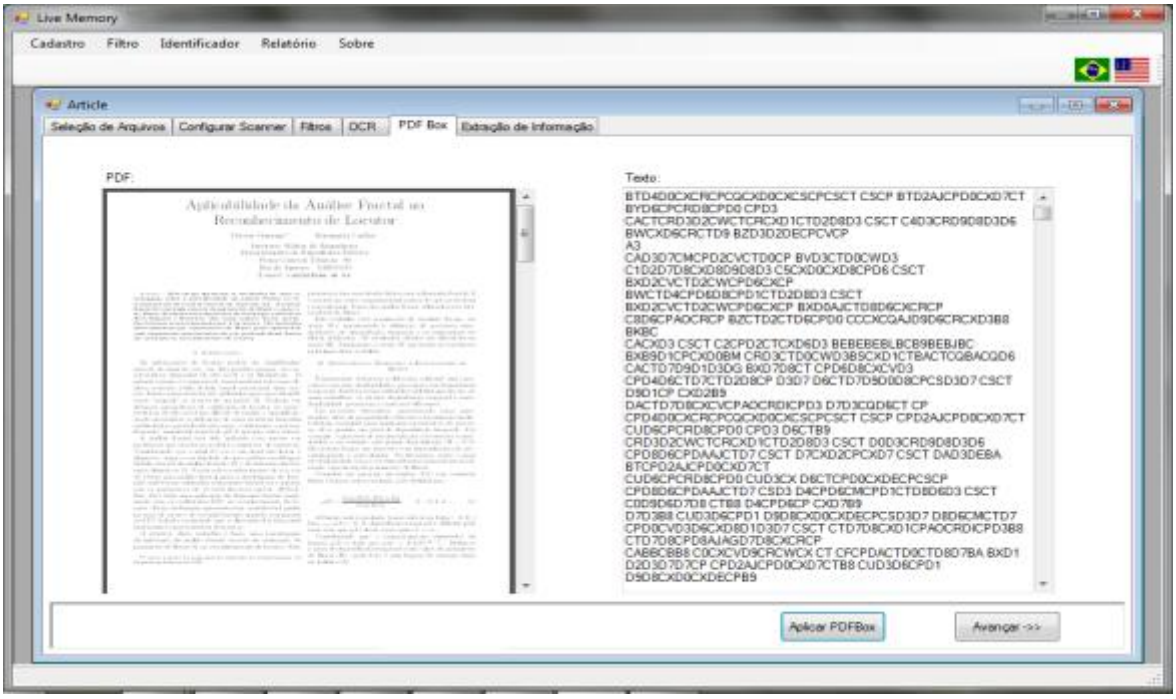


Figura 46 – *Texto produzido pelo PDFBox ao ler PDF protegido.*

Caso 2: Artigos que continham a seção de Palavras-Chaves em apenas um idioma

Nesse caso foi aplicado o método proposto, em 10 artigos (desconsiderando-se a existência da seção de palavras-chave). Posteriormente, foi feita uma análise comparativa visando avaliar o percentual de acerto confrontando o resultado obtido (palavras-chave identificadas) com o explicitado pelo autor (palavras definidas pelo mesmo, em um dos idiomas).

Os resultados obtidos, como exemplificado na Figura 47, é possível verificar, que no campo 3, das 5 palavras identificadas pelo sistema, 1 veio do título e 4 vieram da introdução. Destas, as palavras “radius” e “multiplication” não constavam das palavras-chave definidas pelo autor, como exibido no campo 4. Sendo assim, o sistema identificou 3 palavras-chave das 5 palavras-chave definidas, portanto houve um acerto de 60% para este exemplo.

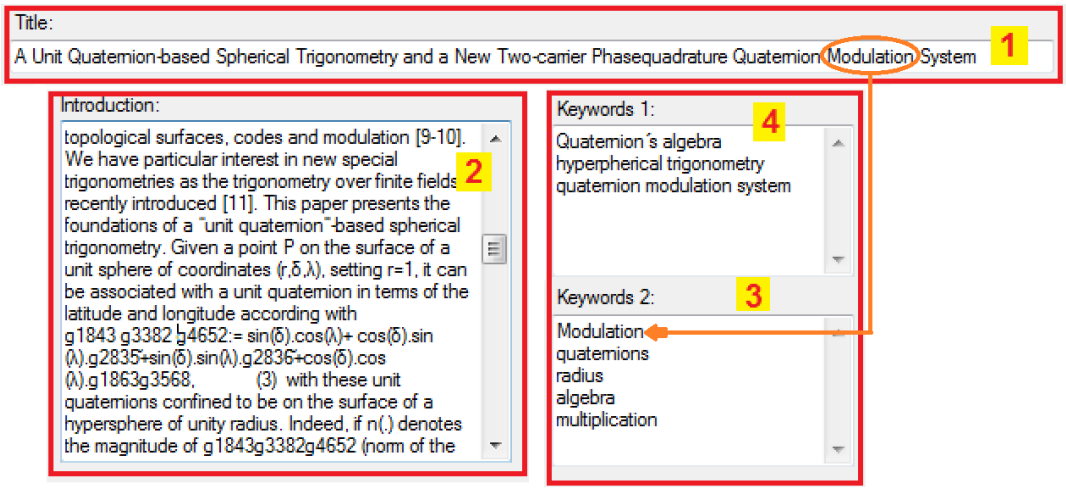


Figura 47– *Extração de Palavras-Chave (Exemplo1).*

AFigura 48 exemplifica um caso em que o usuário identificou 4 palavras distintas e o sistema identificou 6 palavras. Apenas, 2palavras coincidiram com a escolha do autor, ou seja, 33,33%, mas vale ressaltar que todas as 6 palavras-chave extraídas, após uma análise do contexto do artigo, fazem bastante sentido e poderiam ter sidos utilizadas pelo autor. Desta forma, acredita-se que a ferramenta também pode ser utilizada para gera palavras-chave candidatas e assim auxiliar um determinado autor, ao finalizar a escrita de um artigo ou mesmo trabalhos de conclusão de curso.

Em ambos os casos, pode-se observar que a escolha automática de palavras-chave, realizadas pelo sistema sLiveMemory, foi bastante razoável e consequentemente são úteis para a indexação de artigos.

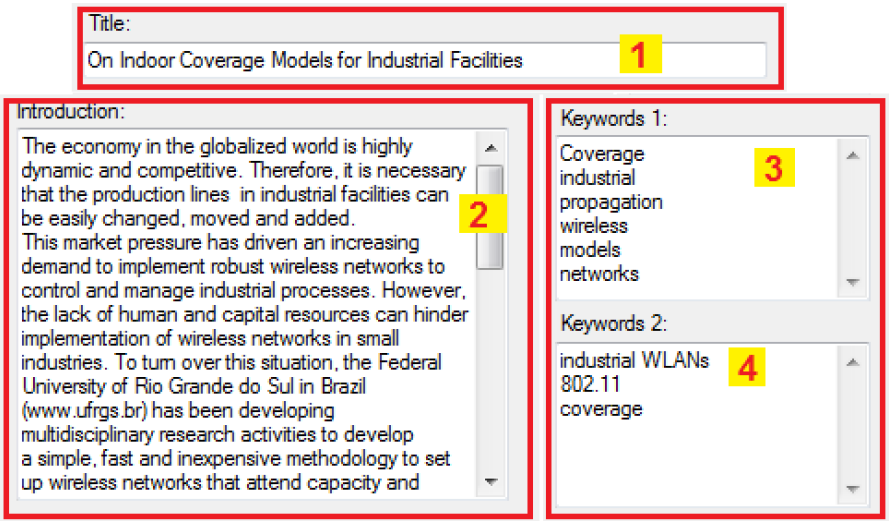


Figura 48– *Extração de Palavras-Chave (Exemplo 2).*

5 pLIVEMEMORY NA NUVEM DO GOOGLE©

Tecnologia de computação em nuvem representa um novo paradigma para o provimento de infraestrutura de computação. Esse paradigma desloca esta infraestrutura para a rede, visando reduzir os custos associados com a gestão de recursos de *hardware* e de *software*. Ela representa o sonho de longa data de idealização da computação como um utilitário, onde a economia de princípios de escala ajuda a efetivamente reduzir o custo da infraestrutura de computação (ARMBRUST *et al*., 2009).

A computação em nuvem simplifica os processos que consomem tempo de provisionamento e aquisição de *hardware*, bem como, de *software*. Portanto, ela promete uma série de vantagens para a implantação de aplicações intensivas de dados, como a elasticidade de recursos, modelo de custo *pay-per-use*, baixo tempo de mercado, e a percepção de, praticamente, recursos ilimitados e escalabilidade infinita. Assim, torna-se possível, pelo menos teoricamente, alcançar um *throughput* ilimitado por meio da contínua adição de recursos de computação, por exemplo, servidores de banco de dados, se ocorrer aumento de carga de trabalho.

Na prática, as vantagens do paradigma de computação em nuvem abrem novos caminhos para a implantação de novas aplicações, que não eram economicamente viáveis em um ambiente tradicional das infraestruturas das empresas. Portanto, a nuvem vem se tornando uma plataforma cada vez mais popular para hospedagem de aplicativos de software em uma variedade de domínios, tais como redes de varejo eletrônico, finanças, notícias e social. Assim, hoje se assiste a uma proliferação no número de aplicativos com um enorme aumento da escala dos dados gerados, bem como sendo consumida por tais aplicações. Sistemas de banco de dados hospedados na nuvem alimentando esses aplicativos formam um componente crítico na pilha de *software* destas aplicações.

Considerando as vantagens apontadas pela computação em nuvem, nesta tese foi desenvolvida uma solução, específica para a plataforma pLiveMemory, voltada para a nuvem. Também optou-se por trabalhar com o ambiente do Google©, pois as bibliotecas de hospedagem de processo na nuvem do Google têm a vantagem de um serviço disponível 24 horas por dia, 365 dias do ano com servidores grátis e eficientes, ligados a *backbones* de alta velocidade. Sendo que o Google indexa de forma independente em seus motores de busca suas páginas oferecendo o máximo de visibilidade. Qualquer pesquisa envolvendo as palavras-chave dos artigos em biblioteca, tem uma maior chance de ser apontada como

uma resposta a uma consulta, se feita usando o motor de busca Google, atualmente o melhor mecanismo e mais amplamente utilizado para pesquisa.

A proposta desta tese contempla um *site* na *web*, que usufrui do serviço gratuito de hospedagem do Google *Sites*©, usado para a publicação da biblioteca digitaldapLiveMemory. O Google *Sites* trabalha em conjunto com o Google *Drive*©, este é um ambiente que armazena e publica documentos na *web*, no caso dapLiveMemory são armazenados os dados referentes aos artigos científicos.

Além da nuvem do Google, a plataforma utiliza os servidores da UFPE. Ao consultar autores e/ou artigos, via o *site* disponível na *web*, as consultas são direcionadas para o banco de dados de artigos científicos, atualmente hospedado na UFPE. É neste banco de dados que estão todos os dados que foram extraídos por meio do ambiente para *desktop*. Os artigos em formato PDF foram enviados para os servidores do Google *Drive*, desta forma há *links* para os arquivos permitindo que o usuário consulte todos os artigos. Neste caso pode-se afirmar que a solução está em nuvem, uma vez que os dados estão espalhados por vários servidores (Google *Drive*, Google *Sites* e UFPE), mas para o usuário é indiferente, conforme Figura 49, que mostraa interação da Nuvem do Google com a base na UFPE.

Figura 49 – *Esquema dapLiveMemory na Nuvem do Google*©.

5.1 Esquema para envio de Biblioteca Digital para a nuvem do Google©

O problema básico enfrentado na publicação de uma biblioteca digital na nuvem do Google© é manter a estrutura original da biblioteca local na nuvem(SANCHATI *et al.*, 2011), (YANG, 2010). A metodologia apresentada nesta seção mostra como fazer isto de forma simples e sistemática:

- (i) Nivelar todos os diretórios, renomeando arquivos com o seu caminho. Por exemplo, o arquivo "~\sbrt\2007\p036" se torna "sbrt_2007_p036";
- (ii) Enviar todos os arquivos para o Google *Drive*© seguindo as operações padrão do mesmo;
- (iii) O Google *Drive*© gera uma lista com os endereços *web*, conforme Figura 50, onde os arquivos foram armazenados. Embora não haja nenhuma maneira sistemática de deduzir os endereços, o Google© mantém a mesma ordem que os arquivos foram enviados;
- (iv) Copiar a lista dos endereços fornecidos pelo Google *Drive*©;
- (v) Criar uma tabela de endereços.Em uma coluna tem-se o endereço do arquivo no sistema sLiveMemory e em outra coluna o endereço da *web*,isto para cada arquivo.

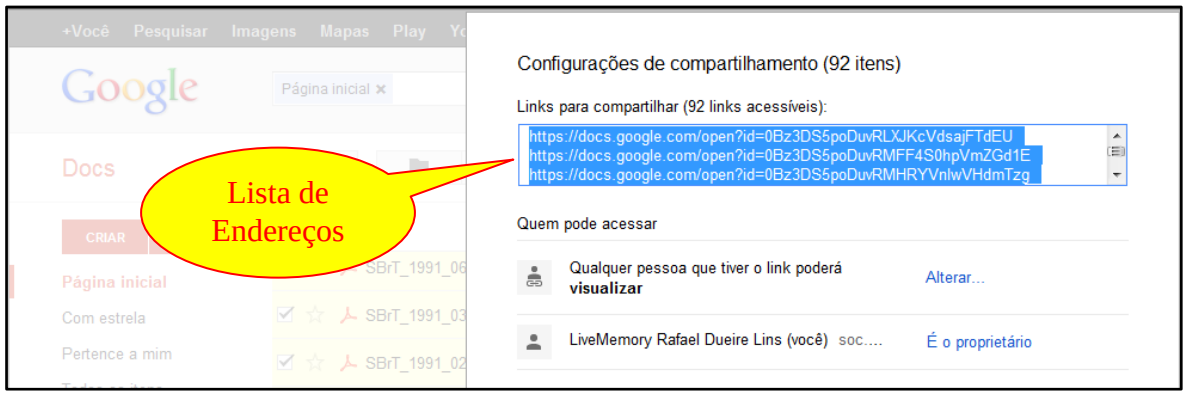


Figura 50 – *Lista com Endereços do Google*©.

A geração dessatabela é o fator chave que permite publicar a biblioteca digital na nuvem do Google©. A partir deste ponto, toda a informação da biblioteca digital está disponível na nuvem do Google©, sendo necessário apenas encontrar maneiras de traduzir a estrutura para a nuvem. A observação de que o processo de carregamento dos arquivos

mantém a ordem original do arquivo foi fundamental para ter um modo simples e sistemático. Então,deste2’ ponto em diante, o usuário deverá:

- (vi) Alterar a permissão de todos os arquivos para torná-los “público”. Muitas vezes, essa mudança do estado dos arquivos precisa do aceite do Google©, o que leva alguns dias, pois, possivelmente, os arquivos passam por uma verificação de conteúdo pelo Google© para averiguar se há material inadequado. Tal processo, embora não permita a visibilidade pública imediata para os arquivos, não impede que o desenvolvedor da biblioteca continue o seu trabalho na geração da biblioteca.

Após o envio dos documentos, o usuário deve criar o *site* e as páginas *web* são hospedadas no Google *Sites*©. A biblioteca criada, com os arquivos da SBrT, foi denominada de “LiveMemory.SBrT”, sua página principal é mostrada naFigura 51. A criação e edição de tal página seguem os procedimentos padrão do Google*Sites*©.

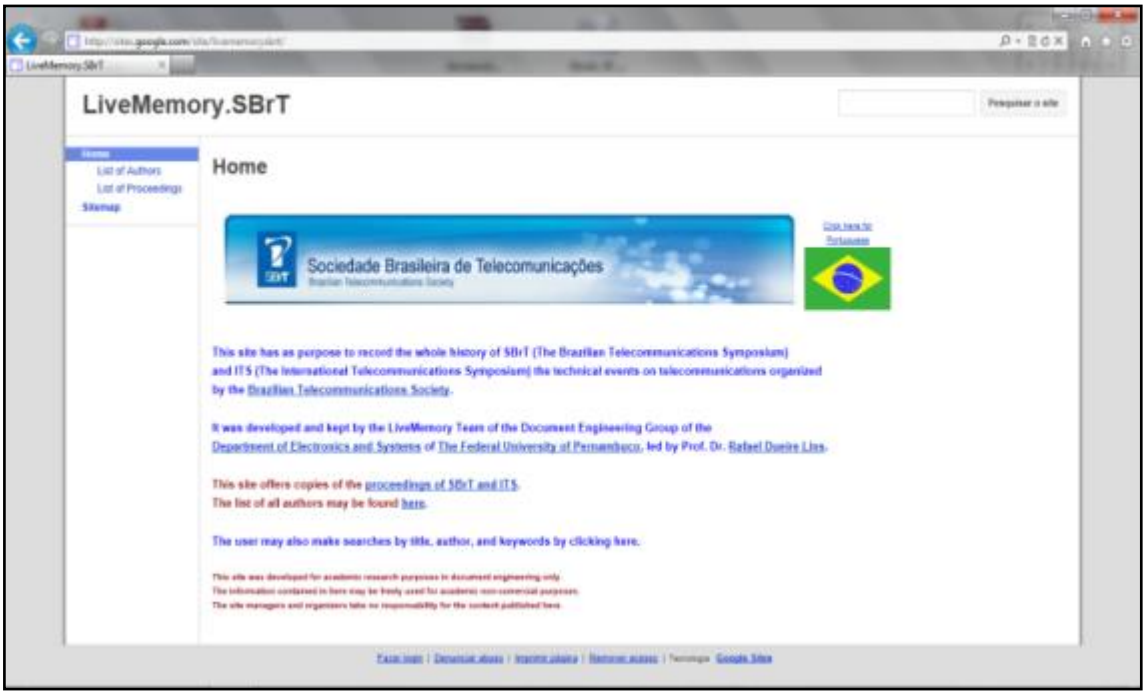


Figura 51 – *Página Principal dapLiveMemory no Google Sites*©.

Para ligar as informações da página criada no Google *Sites*© com as informações armazenadas na base de dados da UFPE e os documentos, em PDF, que foram enviados ao Google *Drive*©, os endereços dos arquivos devem ser mudadospara o seu equivalente na

nuvem do Google©, buscando cada um deles na tabela de endereços, criada no Passo *v*, assim será necessário:

- (i) Chamar o "módulo de conversão de endereços" na plataforma pLiveMemory para verificar a relação entre os endereços locais e os da *web*, os dois endereços ficam armazenados na base de dados;
- (ii) Gerar páginas *web* com o uso da linguagem PHP, estas fazem consultas a base de dados com o uso da linguagem SQL;
- (iii) Armazenar páginas em um servidor da UFPE;
- (iv) Gerar páginas *web*no Google *Sites*©;
- (v) Fazer a ligação entre as páginas do Google Sites© com as do PHP e com os arquivos que estão no Google *Drive*©.

A Figura 52 apresenta um exemplo depágina *web*com a lista de eventos ou edições da SBrT, no Google *Sites*©.

Figura 52 – *Lista de Eventos do pLiveMemory - SBrT.*

Para cada ano/edição há um *link* para uma página em PHP, na nova página o usuário pode acessar o artigo que foi previamente enviado à nuvem do Google©. AFigura

53 apresenta um exemplo de uma tela com uma visão de um artigo, em PDF, que foi selecionado.

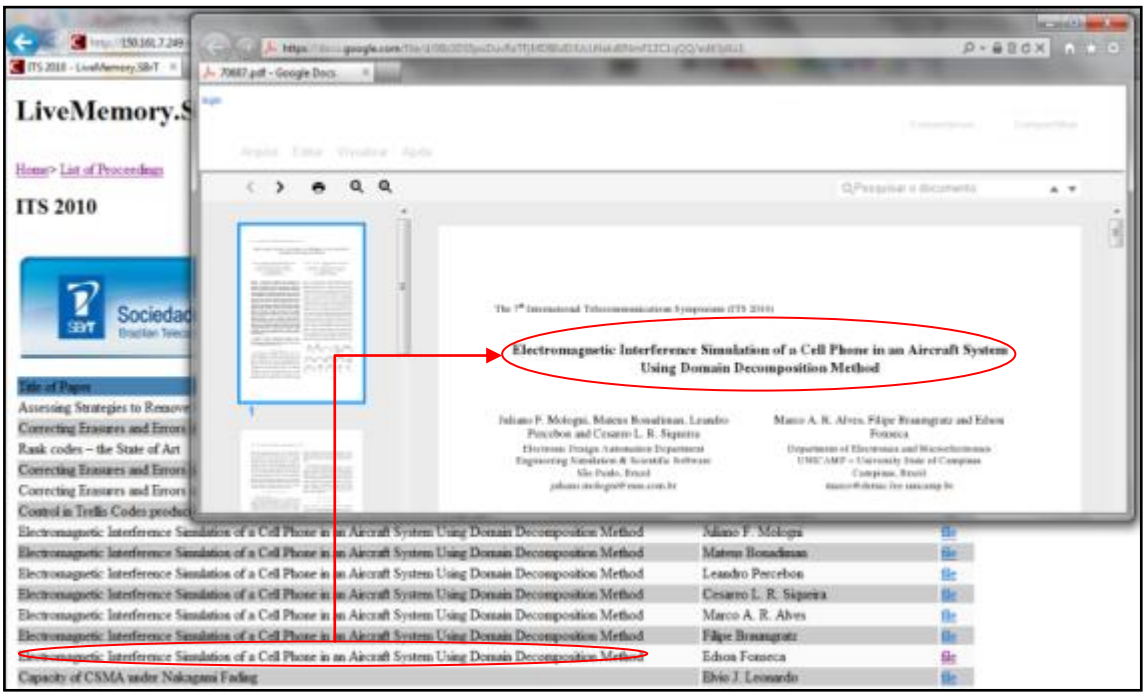


Figura 53 – *Lista de Artigos da ediçãode2010 com a visão de um documento em PDF.*

5.2 Mecanismo de Busca na Nuvem do Google©

Ser capaz de "navegar" em uma biblioteca digital é fundamental para o usuário final. A plataforma pLiveMemory, na versão da nuvem Google©, permite três possibilidades para que o usuário faça suas pesquisas:

- Tabelas estáticas;
- Máquinas de busca externas;
- Máquinas de busca instaladas localmente;

Essas possibilidades de buscas são detalhadas nas próximas subseções.

5.2.1 Tabelas Estáticas de Busca

O universo de consultas feitas por usuários de uma biblioteca digital, em geral, tende a ser muito limitado. Pode-se fazer busca por título de um artigo, pelo nome do autor ou mesmo por palavras-chave. A plataforma pLiveMemory gera tabelas estáticas com tais informações, o desenvolvedor da biblioteca pode fazer o envio para o Google *Sites*©

permitindo que os usuários naveguem facilmente na biblioteca. AFigura 54 apresenta um exemplo com a página de entrada para a “Lista de Autores”.

Figura 54 – *Página Web com o Índice da “Lista de Autores”.*

Sempre que o usuário escolhe a primeira letra do nome do autor, a lista de todos os autores com esse nome é apresentado. A Figura 55apresenta uma tela com a lista de autores que começam com a letra *E*, bem como um artigo, em PDF, que foi selecionado.

Figura 55 – *Página web com a “Lista de Autores” de uma determinada letra.*

5.2.2 Máquinas de Busca Externas

Outra possibilidade de mecanismo de busca oferecido na biblioteca gerada com a plataforma pLiveMemory na nuvem do Google© é a de manter o motor de busca na máquina local do desenvolvedor de biblioteca. O usuário da biblioteca que deseja fazer uma consulta mais complexa tem um *link* para uma página da *web*, do mecanismo de busca no hospedeiro do desenvolvedor, que pode ser acessado via PHP e produz como resposta, à consulta, uma lista de ponteiros para os artigos no Google *Drive*©.

Esta solução tem a vantagem de que as consultas,às bibliotecas e suas respostas, tendem a exigir pouca largura de banda e a máquina servidora pode ter pouco poder para processamento/armazenamento, deste modo, podeser conectada a um *link* lento. Neste caso a carga de tráfego pesado está no *download*, que é feito entre os servidores do Google© e a biblioteca cliente.

5.2.3 Máquinas de Busca Instaladas Localmente

A terceira alternativa é mais complexa, para oferecer um mecanismo de busca para a biblioteca é necessário fornecer ao usuário um código para este armazenar na sua própria máquina todo o conteúdo da biblioteca. A vantagem estaria no acesso, sem a necessidade de conexão de rede, por outro lado se os dados fossem modificados/atualizados o usuário não atualizaria os seus dados automaticamente. Para manter a base consistente, o usuário deveria, novamente, descarregar a base para sua máquina local.

Poucos usuários iriam achar esta alternativa atrativa, além disso, o desenvolvedor da biblioteca tem que dá permissão para diferentes usuários da plataformaaumentando a complexidade do sistema como um todo.

6 CONCLUSÃO

As bibliotecas digitais são grandes nichos para promover e facilitar a disseminação da informação, através da preservação de obras digitais para uso em diferentes comunidades. Permitem assim um maior progresso das sociedades, impulsionado pela geração e aquisição de conhecimento.

Esta tese aborda o universo das bibliotecas digitaiscom foco nos documentos científicos. O desenvolvimento de bibliotecas digitais,neste contexto, busca resolver o problema da preservação de conteúdo científico (muito ainda disponível apenas em formato de papel) para as gerações futuras, bem como acesso à informação por meio de redes de computadores, tornando a recuperação de informação amplamente acessível e barataaos cientistas e interessados, garantindo a manutenção da história da ciência - a ligação entre o antigo e o moderno.

Esta tese inicialmente identificaos problemas relacionados àcomplexidade de extração de informação, manutenção da integridade e integração da informação, armazenamento e distribuição da informação, propondosoluções para estes.

Na tese considera-se que uma boa solução de extração pode auxiliar no tratamento de questões relacionadas à análise da corretude da informação, integração de informações, assim como o seu armazenamento, viabilizando a busca e indexações de conteúdo. Além disso,a tese deixa claro que antes de sua propostanão existiaum ambiente integrado para acesso a artigos científicos que tratasse desde a sua fase de digitalização até à extração/disponibilização de dados. Assim a originalidade da propostaé ser uma plataforma, apoiada por um sistema, para criar e gerenciar bibliotecas digitais a partir de documentos científicos não digitalizados, com foco nos problemas de extração de informaçãoe acesso a documentos científicos de forma integrada, seguindo padrões existentes para definição de bibliotecas digitais.

Ao longo da tese, diversos artefatos para a definição de requisitos no domínio de documentos científicos foram identificados e modelados de modo a permitir o seu reuso no desenvolvimento de outras bibliotecas digitais, quer a partir de documentos impressos ou eletrônicoscomo exemplificado nas bibliotecas: *Academus* e *Thanatos*. Os artefatos apresentados incluem: requisitos, restrições, características (*features*), arquitetura e um modelo de dados que atende às necessidades exigidas para a persistência dos dados envolvidos.Com base nestes artefatos foi proposta e apresentada a plataforma

pLiveMemory,que segue o Modelo de Referências da DELOS, a partir dos quais foram elaborados mapas conceituais relacionadosà plataforma pLiveMemory.

Porém, aidentificação e extração de informações de documentos científicos éde fato o principal foco da tese. Nesse contexto, os arquivos em formato PDF foram desestruturados,com o uso da ferramenta PDFBox, para gerar um arquivo de texto simples, o qual é analisado, tentando combinar padrões descritos por expressões regulares. Nos arquivos em formato de imagens, a transcrição é feita via *software* OCR.Os resultados experimentais apresentados utilizaram os artigos publicados na SBrT.Três itens distintos foram extraídos: dados da primeira página do artigo, referências e palavras-chave. Para extração dos dados da primeira página do artigo (título, autor, *email*, resumo, palavras-chave etc.) foram utilizadas expressões regulares juntamente com as informações cadastradas no ambiente de configuração dos *templates* (número de página por artigo e início da linha de título).

Apesar da existência de muitas propostas na literatura, para a extração e classificação de fragmentos de referências, foi apresentado um processo específico, que foca na extração dos principais dados de uma referência (título, autor e ano); por considerar a sua relevância para as pesquisas.Desta forma, o ambiente tornou-se mais flexível para extração de qualquer referência, permitindo, com um número mínimo de dados de treino, a implantação deste módulo em qualquer biblioteca digital.

Vários algoritmos e técnicas (K-NN, Distância Euclidiana e Similaridade do Cosseno) foram analisados e testados, mas o que apresentou os melhores resultados foram as expressões regulares juntamente com o classificador*Naïve Bayes*. Quando o título está entre aspas, usa-se expressão regular, quando o título estáem outro formato, como em negrito, usa-se o classificador automático.As expressões regulares apresentaram um bom desempenho, quando os elementos a serem identificados tinham um formato regular, tal como o ano, entretanto o resultado não foi tão bom quando da identificação do título ou de autores, pois estes, geralmente,não possuem um formato regular. Já o classificador probabilístico *Naïve Bayes*, juntamente com as Expressões Regulares, após a construção da base de treino e a distribuição de probabilidades,obteve índice de acertos em torno de 70%, desta forma a integração das duas estratégias produziuos melhores resultados.

Outra questão relacionada aos documentos científicos, tratada nesta tese, foi a extração de palavras-chave. A estratégia descrita é capaz de trabalhar tanto com arquivos em formato PDF quanto com arquivos no formato de imagens.Se o documento original não

tem palavras-chave, estas sãoextraídas do título e/ou da seção introdutória. As palavras-chave encontradas são vinculadas ao artigo e enviadas ao “dicionário de palavras-chave”. Um dicionário de palavras auxiliares ou *stop-words*também foi criado.Esta iniciativa é única na literatura, mas importante para indicar o contexto de artigos mais antigos que não possuem palavras-chave.

Por fim, foi apresentado a metodologia para publicar uma Biblioteca Digital na *World Wide Web* usando a nuvem do Google©, mais especificamente usando o Google *Sites*© e o Google *Drive*©. Embora o processo não seja totalmente automático, as tarefas críticas e mais difíceis são realizadas pela plataforma pLiveMemory e os passos não automatizados são explicados. A hospedagem de bibliotecas no Google© tem a vantagem de um serviço disponível 24 horas por dia, 365 dias do ano para livre consulta e com servidores eficientes, ligados à*backbones* de alta velocidade. Além disso, o Google© indexa de forma independente, em seus motores de busca, suas páginas oferecendo o máximo de visibilidade, assim qualquer pesquisa envolvendo as palavras-chave dos artigos da biblioteca, possui uma maior chance de ser apontada como uma resposta de uma consulta, se feita usando o motor de busca Google©.

6.1 Contribuições da Tese

Dentre as contribuições desta tese está a proposta e apresentação de estratégias para construção de uma plataforma, apoiada por um sistema, para criar e gerenciar bibliotecas digitais integradas a partir de documentos científicos, considerando formas de reuso de informação. Para alcançar o objetivo principal do trabalho os seguintes itens foram elaborados:

- Proposta de uma estratégia para extraçãode informações dos arquivos em formato PDF ou de imagem (ALVES *et al*., 2011);
- Desenvolvimento de uma estratégia e módulo para tratar referências bibliográficas (ALVES *et al.*, 2012a);
- Desenvolvimento de uma estratégia e módulo para identificar palavras-chave;
- Criação de uma base de dados para armazenamentodas informações relacionadas aos itens anteriores (LINS *et al.*, 2010);
- Desenvolvimento de um ambiente para *web*;
- Proposta de um padrão para reuso de informação na montagem de novas bibliotecas digitais (ALVES *et al*., 2012b) a partir de conhecimento prévio

adquirido com a pLiveMemory e plataformas semelhantes no domínio das bibliotecas digitais *Academus* (LINS *et al*., 2011) e *Thanatos* (ALMEIDA *et al*., 2011).

Como consequência foi desenvolvida uma interface de um sistema para cadastro, armazenamento e extração de informações de artigos científicos (LINS *et al.*, 2010), esta passou por melhorias na extração por meio da automatização do processo com o uso de expressões regulares (ALVES *et al.*, 2011). Também foi criado um ambiente específico para extração de referências bibliográficas, o qual utilizava expressões regulares juntamente com o classificador *Naïve Bayes* entre outros classificadores (ALVES *et al.*, 2012a).Além disso, foi desenvolvida ainda uma interface para extração de palavras-chave dos títulos e/ou da introdução dos artigos científicos, por fim foi construído um *site* na nuvem do Google© para divulgar e possibilitar a consulta dos artigos da SBrT, em formato PDF.

6.2 Limitações da Proposta

Ao longo das fases do trabalho limitações foram identificadas, conforme descrição a seguir:

- Limitação no tamanho da base de treino para extração das referências;
- Necessidade de intervenção do usuário para extração de Palavras-Chave;
- As ferramentas para desktop e web trabalham independentemente, com acesso à mesma base de dados;
- Para cada nova biblioteca é necessário uma nova configuração no ambiente;
- A dificuldade para integração da informação foi reduzida, mas ainda precisa ser melhorada, ou seja, os problemas ainda não foram totalmente resolvidos.

6.3 Trabalhos Futuros

O desenvolvimento da plataforma pLiveMemory não se esgota nesta tese, há várias possibilidades de expansão, principalmente para o ambiente na *web*. Como o acesso àInternet está disponível em vários locais é possível a migração de módulos que estão em*desktop* para a *web*.

Com relação ao módulo de detecção de palavras-chave, existe a possibilidade de se incluir uma ontologia, em associação com o dicionário de palavras-chave,podendo maximizar os resultados alcançados.

Da mesma forma, a inserção de técnicas de Sumarização Automática para os artigos, bem como para os demais documentos científicos, traz a possibilidade de se trabalhar com uma pilha de artigos, devolvendo para o usuário, além de um conjunto de artigos relacionados, um histórico cronológico do assunto pesquisado.

Com relação às referências bibliográficas, seria interessante a busca destas nas demais bases de publicações (ACM, IEEE,*Springer Verlag* etc.) com o intuito de indicar onde determinada referência foi citada e mesmo confirmar se sua origem e/ou descrição é verdadeira. Também seria relevante verificar se um determinado artigo foi citado nas referidas bases, criando assim um índice de citações externas, bem como indicar quantas vezes uma referência é citada no mesmo artigo. Outra possibilidade de continuidade é a inclusão do cálculo do fator H,aplicado para estimar a produtividade dos autores baseado nos seus registros de publicações e citações.

O compartilhamento de dados pode auxiliar no desenvolvimento do conceito de *Web* Semântica, pois este se caracteriza principalmente pela organização do conteúdo e interação do usuário com o material disponibilizado.

REFERÊNCIAS

ABBYY FineReader; 2011. Disponível em: http://finereader.abbyy.com. Acesso em Junho/2011.

ADOBE; 2012. Disponível em: http://www.adobe.com/br/. Acesso em Outubro/2012.

ALEXANDER, Christopher; ISHIKAWA, Sara; SILVERSTEIN, Murray; JACOBSON, Max; FOKSDAHL-KING, Ingrid; ANGEL, Shlomo; 1977. **A Pattern Language: Towns, Buildings, Construction**. Oxford University Press: New York, 1977.

ALJABER, Bader; STOKES, Nicola; BAILEY, James; PEI, Jian; 2010. **Document clustering of scientific texts using citation contexts**. Information Retrieval. April 2010, Volume 13, Issue 2, pp 101-131. DOI 10.1007/s10791-009-9108-x.

ALMEIDA, Alessandra; LINS, Rafael Dueire; SILVA, Gabriel Pereira e; 2011. ***Thanatos*: Automatically Retrieving Information from Death Certificates in Brazil**. In HIP'2011, Sep. 16-17, 2011, Beijing, China, ACM Press.

ÁLVAREZ, Alberto Cáceres; 2007. **Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem**. USP 2007. Disponível em: http://www.teses.usp.br/teses/disponiveis /55/55134/tde-21062007-144352/.

ALVES, Neide Ferreira; LENCASTRE, Maria; LINS, Rafael Dueire; 2011. **Building Digital Libraries of Proceedings with the LiveMemory Platform**. In: Ninth International Workshop on Graphics Recognition (GREC 2011), Seoul. 9th IAPR International Workshop on Graphics Recognition, 2011.

ALVES, Neide Ferreira; LENCASTRE, Maria; LINS, Rafael Dueire; 2012a. **Improving Requirements Quality in Digital Libraries: The case of Scientific Proceedings**. In Proceedings of the Quatic 2012-8th International Conference on the Quality of Information and Communications Technology. Lisbon, Portugal, September 3-6, 2012.

ALVES, Neide Ferreira; LINS, Rafael Dueire; LENCASTRE, Maria; 2012b. **A Strategy for Automatically Extracting References from PDF Documents**. In Proceedings of the DAS 2012-10th IAPR Int. Workshop on Document Analysis Systems. (Gold Coast, Queensland, Australia, March 27-29, 2012). IEEE Press, 435-439. DOI 10.1109/DAS.2012.12.

ARMBRUST, M. et al.; 2009. **Above the Clouds: A Berkeley View of Cloud Computing**. Technical Report", EECS Department, University of California, Berkeley, 2009.

AVANCINI, Henri; CANDELA, Leonardo; STRACCIA, Umberto; 2007. **Recommenders in a personalized, collaborative digital library environment**. vol. 28, pp.253-283, Journal of Intelligent Information Systems, june 2007, doi:10.1007/s10844-006-0010-3.

BARNES, B.; BOLLINGER, T.; 1991. **Making Reuse Cost Effective**. IEEE Software, Vol. 8, No. 1, 13-24, Janeiro 1991.

BATORY, Don S.; 2005. **Feature Models, Grammars, and Propositional Formulas**. In Proceedings of SPLC. 2005, 7-20.

CABRAL, Luciano de Souza; LINS, Rafael Dueire; LIMA, R. J.; SIMSKE, Steven J.; 2012. **A Comparative Assessment of Language Identification Approaches in Textual Documents**. In: IADIS International Conference Applied Computing 2012, Madrid: IADIS, 2012. p. 67-74.

CANDELA, L.; CASTELLI, D..; FERRO, N.; IOANNIDIS, Y.; KOUTRIKA, G.; MEGHINI, C.; PAGANO, P.; ROSS, S.; SOERGEL, D.; AGOSTI, M.; DOBREVA, M.; KATIFORI, V.; SCHULDT, H.; 2008. **The DELOS Digital Library Reference Model - Foundations for Digital Libraries**. Version 0.98 (February 2008).

CLEMENTS, P.; NORTHROP, L.; 2002. **Software Product Lines: Practices and Patterns**. Boston: Addison Wesley, 2002.

CONSTANS, Pere; 2009. **A Simple Extraction Procedure for Bibliographical Author Field**. Banyoles, 2009.

DELOS - Network of Excellence; 2012. Disponível em:http://www.delos.info/index.php?option=com_frontpage&Itemid=1. Acesso em Outubro/2012.

DLF - Digital Library Federation; 1998. Disponível em: http://old.diglib.org/about/dldefinition.htm. Acesso em Outubro/2012.

FAYAD, M.; JOHNSON, R. ; 1999. **Building Application Frameworks: Object-Oriented Foundations of Framework Design**. John Wiley&Sons, 1999.

FRANK, Eibe; WITTEN, Ian H.; 1998. **Generating accurate rule sets without global optimization**. In: ICML 98: Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. p. 144-151. ISBN: 1-55860-556-8.

GAMMA, Erich; HELM, Richard; JOHNSON, Ralph; VLISSIDES, John; 1995. **Design Patterns: Elements of Reusable Object-Oriented Software**. Addison Wesley, 1995.

GONZALEZ, R. C.; WOODS, R. E.; 2000. **Processamento de Imagens Digitais**. São Paulo: Editora Edgard Blucher Ltda.

GUIZZARDI; 2005. **Ontological Foundations for Structural Conceptual Models**. Phd Thesis, University of Twente, The Netherlands, 2005.

GURRIN, Cathal; AARFLOT, Tjalve; JOHANSEN, Dag; 2009; **GARDI: A Self-Regulating Framework for Digital Libraries**. Computer and Information Technology, 2009. CIT '09. Ninth IEEE International Conference on, vol.1, pp.305-310, 11-14 Oct. 2009, ISBN: 978-0-7695-3836-5, doi: 10.1109/CIT.2009.137.

ImageJ; 2013. Disponível em:http://rsb.info.nih.gov/ij/index.html. Acesso em junho/2013.

JARGAS, Aurelio Marinho; 2009. **Expressões Regulares - Uma abordagem divertida**. 3ª. Edição. Novatec, 2009.

KRUEGER, C.; 1992. **Software Reuse**. ACM Computer Survey 24, 1992.

LAFFERTY, J. ; MCCALLUM, A.; PERREIRA, F. ; 2001. **Conditional Random Fields: Probabilistic Modelsfor Segmenting and Labeling Sequence Data**. In Proceedings of the International Conference onMachine Learning (ICML), 282-289, 2001.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade; 2003. **Fundamentos de Metodologia Científica**. 5. ed. São Paulo: Atlas, 2003.

LE BOURGEOIS, F.; EMPTOZ, H.; 2007. **DEBORA: Digital AccEss to BOoks of the RenAissance**. In: International Journal on Document Analysis and Recognition. Volume 9, Numbers 2-4, 193-221. Springer-Verlag 2007, Berlin, Heidelberg. DOI: 10.1007/s10032-006-0030-0.

LINS, Rafael Dueire; ÁVILA, Bruno Tenório; FORMIGA, Andrei de Araújo; 2006. **BigBatch: An Environment for Processing Monochromatic Documents**. In Proceedings of the ICIAR 2006, LNCS 4142, pp. 886-896. Springer Verlag. DOI: 10.1007/11867661_80.

LINS, Rafael Dueire; GONÇALVES, Paulo; 2004. **Automatic language identification of written texts**. ACM Symposium on Applied Computing SAC 04, 1128. ACM Press. DOI: 10.1145/967900.968129.

LINS, Rafael Dueire; LIMA, Paulo Hugo; SILVA, Gabriel Pereira e; 2011. ***Academus* - Generating Digital Libraries of M.Sc. and Ph.D. Thesis**.In: International Workshop on Graphics Recognition, 2011, Seoul. GREC 2011. Seoul: IAPR Press, 2011. p. 86-95.

LINS, Rafael Dueire; SILVA, Gabriel Pereira e; TORREÃO, Gabriel; ALVES, Neide Ferreira; 2010. **Efficiently Generating Digital Libraries of Proceedings with the LiveMemory Platform**. Proceedings of ITS 2010 - IEEE-SBrT International Telecommunications Symposium.

LINS, Rafael Dueire; TORREÃO, Gabriel; SILVA, Gabriel Pereira e; 2009. **Content Recognition and Indexing in the LiveMemory Platform**. In Proceedings of the GREC 2009. Springer Verlag. LNCS 6020. pp. 224 - 230, 2010. DOI: 10.1007/978-3-642-13728-0_20.

MARCHIORI, Patricia Zeni; 1997. **Ciberteca ou biblioteca virtual: uma perspectiva de gerenciamento de recursos de informação**. Ciência da Informação, Brasília, v. 26, n. 2, Maio 1997. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19651997000200002&lng=en&nrm=iso. Acesso em Outubro/2012.

MAZUREK, Cezary; WERLA, Marcin; 2005. **Digital Object Lifecycle in dLibra Digital Library Framework**. 8th DELOS Workshop on Future Digital Library Management Systems, 2005.

MySQL; 2013. Disponível em:http://www.mysql.com/. Acesso em Junho/2013.

NOVAK, Joseph Donald; 1998. **Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations**. 2a. Edition, Lawrence Erlbaum Associates, 1998.

OAI; 2012 Disponível em:http://www.openarchives.org/.Acesso em Maio/2012.

OHTA, Manabu; YAKUSHI, Takayuki; TAKASU, Atsuhiro; 2008. **Bibliographic Element Extraction from Scanned Documents Using Conditional Random Fields**. Digital Information Management, 2008. ICDIM 2008. Third International Conference on, pp. 99-104, 13-16 Nov. 2008. doi: 10.1109/ICDIM.2008.4746745

PDFBox; 2011. Disponível em: http://www.pdfbox.org. Acesso em Março/2011.

PDFTOTEXT; 2013. Disponível em:http://linux.die.net/man/1/pdftotext. Acesso em Junho/2013.

SANCHATI, R.; KULKARNI, G.; 2011. **Cloud Computing in Digital and University Libraries**. Global Journal of Computer Science and Technology, Vol. XI (XII), Version 1, July 2011.

SILVA, Eduardo Fraga do Amaral e; 2004. **Um sistema para extração de informação em referências bibliográficas baseado em aprendizagem de máquina**. 2004. Dissertação de Mestrado (Mestrado em Ciência da Computação) - Centro de Informática da Universidade Federal de Pernambuco, Pernambuco.

SOJKA, Per; HATLAPATKA, Radim;2010. **Document engineering for a digital library: PDF recompression using JBIG2 and other optimization of PDF documents**. In: Proceeding of the 10th ACM symposium on Document Engineering; 2010 Sep 21-24; Manchester-United Kingdom. ACM, New York, NY, USA 2010. p. 3-12. Doi: 10.1145/1860559.1860563.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin; 2009. **Introdução ao DATAMINING Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna Ltda.

Tesseract; 2011. Disponível em:http://code.google.com/p/tesseract-ocr. Acesso em Junho/2011. Acesso em Abril/2011.

YANG, J.; 2010. **Cloud Computing in the Application of Digital Library**. In Proceedings Intelligent Computation Technology and Automation (ICICTA), 2010, China.

APÊNDICE A – Descrição dos Casos de Uso

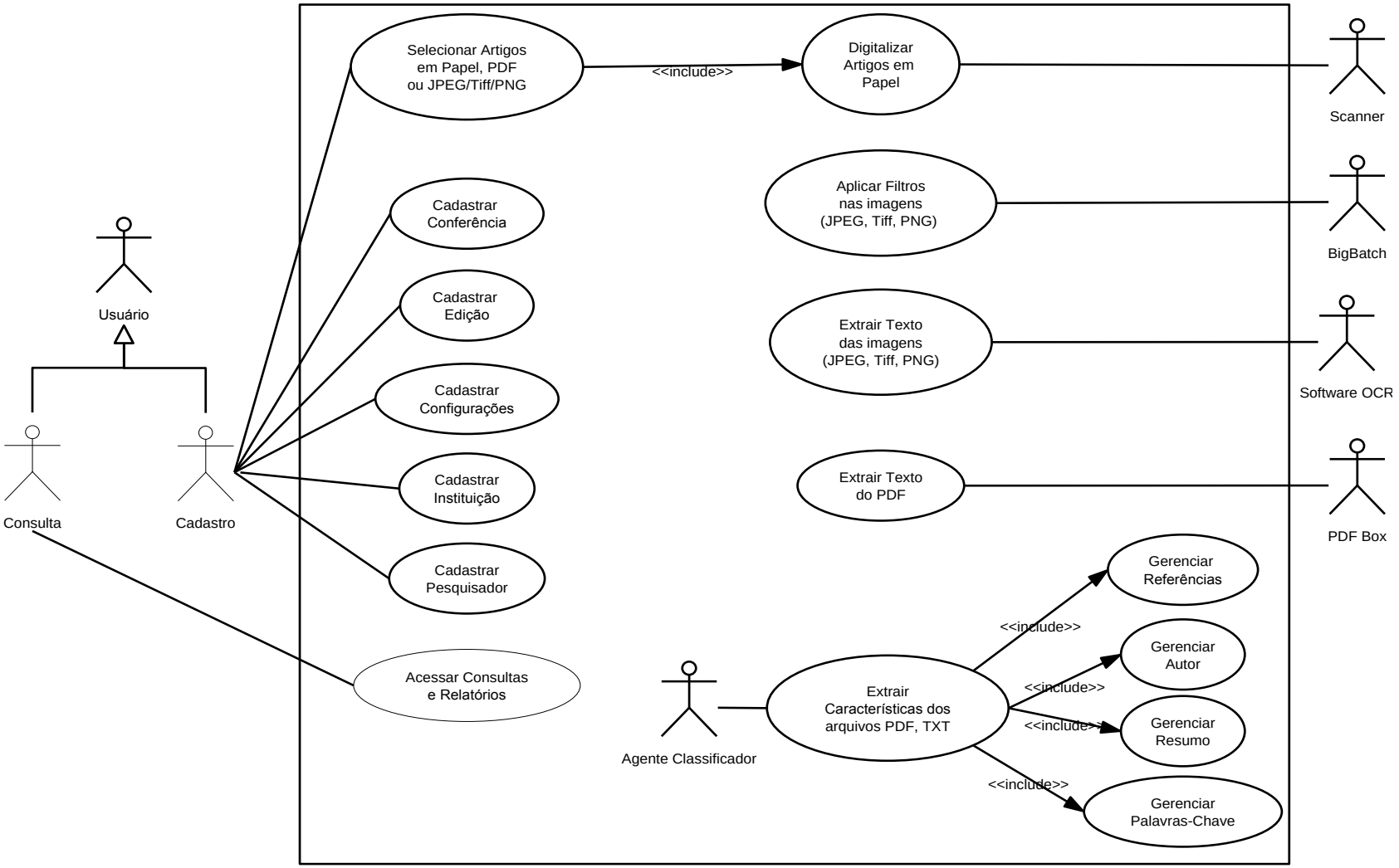


Figura 56 – *Diagrama de Caso de Uso daPlataforma pLiveMemory.*

Caso de Uso 01: Cadastrar Conferência

Objetivo

Permitir que o usuário possa cadastrar informações sobre conferências, tais como: nome e área de atuação.

Atores

Usuário – usuário do sistema, aquele que irá cadastrar as informações das conferências.

Pré-Condições

Não existem.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão "Cadastrar Conferência";

P.2 – O sistema gera automaticamente um código ou chave-primária para identificar a conferência;

P.3 - O usuário preenche um formulário onde deve informar:nome da conferência, área de atuação e quantidade de volumes;

P.4 – O sistema efetua a validação do formulário, checando se ele foi preenchido corretamente;

P.5 – O sistema redireciona o ator para o formulário "Cadastrar Edição".

Fluxos Alternativos

Nenhum.

Fluxos de Exceção

E.1 – O usuário não preenche os dados corretamente;

E.2 – O sistema indica os eventuais erros e solicita a correção dos mesmos;

E.3 – Quando o usuário informar os dados corretamente, o sistema continua da sessão P.4 do fluxo principal.

Pós-condições

Uma conferência foi criada e o usuário poderá cadastrar informações sobre as edições desta.

Requisitos Não Funcionais

Nenhum.

Caso de Uso 02: Cadastrar Edição

Objetivo

Permitir que o usuário possa cadastrar informações sobre edições de conferências já cadastradas.

Atores

Usuário – usuário do sistema, aquele que irá cadastrar as informações das edições.

Pré-Condições

A conferência a qual a edição faz parte já deverá ter sido cadastrada.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão "Cadastrar Edição" ou quando ao final do cadastro de conferência, o sistema o direciona para este formulário;

P.2 – O sistema gera automaticamente um código ou chave-primária para identificar a edição;

P.3 –O usuário preenche um formulário onde deve informar:nome da edição, local da realização, ano, idioma, nome do(s) *Chair*(s) e configurações dos artigos;

P.4 – O sistema efetua a validação do formulário, checando se ele foi preenchido corretamente;

P.5 – O sistema redireciona o ator para o formulário "Selecionar Artigos".

Fluxos Alternativos

Nenhum.

Fluxos de Exceção

E.1 – O usuário não preenche os dados corretamente;

E.2 – O sistema indica os eventuais erros e solicita a correção dos mesmos;

E.3 – Quando o usuário informar os dados corretamente, o sistema continua da sessão P.4 do fluxo principal.

Pós-condições

Uma edição será criada e o usuário poderá cadastrar informações sobre as configurações ou *templates* da edição.

Requisitos Não Funcionais

Nenhum.

Caso de Uso 03: Cadastrar Configurações

Objetivo

Permitir ao usuário cadastrar informações, relacionadas ao posicionamento dos dados,em uma edição de uma determinada conferência ou utilizar uma configuração pré-existente.

Atores

Usuário – usuário do sistema, aquele que irá cadastrar as configurações de uma edição.

Pré-Condições

A edição a qual a configuração faz parte já deverá ter sido cadastrada.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão "Cadastrar Configuração" ou quando está lançando informações sobre uma edição e o usuário seleciona a opção para cadastrar uma nova configuração e o sistema o direciona para este formulário;

P.2 – O usuário pode optar por inserir uma nova configuração ou aproveitar uma pré-existente;

P.2.1 – O sistema gera automaticamente um código ou chave-primária para identificar a edição, o usuário preenche um formulário onde deve informar os dados sobre o layout ou

posição dos itens dos artigos como: título da conferência, número de página, título do artigo, *abstract*, resumo, palavras-chave, autor e referências;

P.2.2 - O usuário pode consultar e selecionar as configurações já cadastradas de outras edições;

P.3 – O sistema efetua a validação do formulário, checando se ele foi preenchido corretamente;

P.4 – O sistema redireciona o ator para o formulário "Edição" ou para "Menu Principal".

Fluxos Alternativos

Nenhum.

Fluxos de Exceção

E.1 – O usuário não preenche os dados corretamente;

E.2 – O sistema indica os eventuais erros e solicita a correção dos mesmos;

E.3 – Quando o usuário informar os dados corretamente, o sistema continua da sessão P.3 do fluxo principal.

Pós-condições

Uma configuração foi selecionada ou cadastrada para uma referida edição.

Requisitos Não Funcionais

Nenhum.

Caso de Uso 04: Cadastrar Pesquisador

Objetivo

Permitir que o usuário possa cadastrar ou selecionar os pesquisadores ou os responsáveis (*chair*) pela conferência.

Atores

Usuário – usuário do sistema, aquele que irá cadastrar os pesquisadores de uma edição.

Pré-Condições

A instituição do pesquisador já deverá ter sido cadastrada.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão "Cadastrar Pesquisador" ou quando está lançando informações sobre uma edição e o usuário seleciona a opção para cadastrar um novo pesquisador (*chair*) e o sistema o direciona para este formulário.

P.2 – O sistema gera automaticamente um código ou chave-primária para identificar a edição, o usuário preenche um formulário onde deve informar os dados sobre o pesquisador, tais como: nome, *email* e instituição.

P.3 – O sistema efetua a validação do formulário, checando se ele foi preenchido corretamente;

P.4 – O sistema redireciona o ator para o formulário "Edição" ou para "Menu Principal".

Fluxos Alternativos

Nenhum.

Fluxos de Exceção

E.1 – O usuário não preenche os dados corretamente;

E.2 – O sistema indica os eventuais erros e solicita a correção dos mesmos;

E.3 – Quando o usuário informar os dados corretamente, o sistema continua da sessão P.3 do fluxo principal.

Pós-condições

Umpesquisador foi cadastrado e possivelmente selecionado para uma referida edição.

Requisitos Não Funcionais

Nenhum.

Caso de Uso 05: Cadastrar Instituição

Objetivo

Permitir que o usuário possa cadastrar os dados das instituições dos pesquisadores.

Atores

Usuário – usuário do sistema, aquele que irá cadastrar as instituições dos pesquisadores.

Pré-Condições

Nenhuma.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão "Cadastrar Instituição" ou quando está lançando informações sobre pesquisador ou *chair* e o usuário seleciona a opção para cadastrar uma nova instituição e o sistema o direciona para este formulário.

P.2 – O sistema gera automaticamente um código ou chave-primária para identificar a instituição, o usuário preenche um formulário onde deve informar dados sobre a instituição: nome, sigla, endereço, país.

P.3 – O sistema efetua a validação do formulário, checando se ele foi preenchido corretamente;

P.4 – O sistema redireciona o ator para o formulário "Pesquisador" ou para o "Menu Principal".

Fluxos Alternativos

Nenhum.

Fluxos de Exceção

E.1 – O usuário não preenche os dados corretamente;

E.2 – O sistema indica os eventuais erros e solicita a correção dos mesmos;

E.3 – Quando o usuário informar os dados corretamente, o sistema continua da sessão P.3 do fluxo principal.

Pós-condições

Uma instituição foi cadastrada.

Requisitos Não Funcionais

Nenhum.

Caso de Uso 06: Selecionar Artigos

Objetivo

Permitir que o usuário possa selecionar os arquivos dos artigos, quer seja no formato PDF ou de imagens (JPEG, PNG, TIFF), além dos artigos em formato em papel.

Atores

Usuário – usuário do sistema, aquele que irá selecionar os artigos.

Pré-Condições

A edição dos artigos já deverá ter sido cadastrada.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário ao acessa o sistema clica no botão "Selecionar Artigos";

P.2 - O sistema gera automaticamente um código ou chave-primária para identificar o artigo,o sistema também traz a edição;

P.3 – O usuário indica se os dados estão em papel, imagem ou PDF;

P.3.1 – Dados em papel o sistema chamará o módulo para "Digitalizar artigos em papel";

P.3.2 – Dados em formato de imagem ou PDF. O usuário seleciona os arquivos que compõem a edição, o sistema automaticamente altera o nome do arquivo selecionado para *NomedaEdição+SequênciaNumérica*e copia o arquivo para o diretório da referida edição;

P.3.2.1 – Se formato PDF, o sistema redireciona para "Extrair Texto do PDF";

P.3.2.1 – Se formato de imagem, o sistema redireciona para "Extrair Texto das Imagens";

P.4 – O sistema efetua a validação do formulário, checando se ele foi preenchido corretamente;

P.5 – O sistema redireciona o ator para o "Menu Principal".

Fluxos Alternativos

Nenhum.

Fluxos de Exceção

E.1 – O usuário não preenche os dados corretamente.

E.2 – O sistema indica os eventuais erros e solicita a correção dos mesmos.

E.3 – Quando o usuário informar os dados corretamente, o sistema continua da sessão P.4 do fluxo principal.

Pós-condições

Os artigos são associados a uma edição.

Requisitos Não Funcionais

Nenhum.

Caso de Uso 07: Extrair Texto (Imagens ou PDF)

Objetivo

Permitir que o usuário possa extrair os textos dos artigos, quer estes estejam no formato PDF ou de imagens (JPEG, PNG, TIFF).

Atores

Usuário – usuário do sistema, aquele que irá selecionar os artigos;

Software OCR – aquele que irá extrair os dados dos arquivos em formato de imagem;

PDFBox –aquele que irá extrair os dados dos arquivos em formato PDF.

Pré-Condições

Os arquivos deverão ter os formatos PDF, JPEG, PNG ou TIFF.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão “Extrair Características”;

P.2 – O usuário seleciona um arquivo de imagem e o seu conteúdo é exibido na tela;

P.3 – O usuário pode aplicar filtros (cinza, preto e branco, extrair borda, extrair ruído sal e pimenta) e o sistema exibe e o resultado, permitindo que o usuário decida qual a melhor opção;

P.4 – Ao finalizar o sistema salvar as modificações no arquivo;

P.5 – O usuário selecionar a opção “Aplicar OCR” e o sistema exibe o resultado, permitindo que o usuário visualize os dados em formato de texto;

P.6 – O usuário pode selecionar a opção “Extrair Informações” e ao final o sistema irá identificar: título do artigo, autores, palavras-chave, referências;
P.7 – O sistema armazena no banco de dados as informações extraídas do arquivo.

Fluxos Alternativos

A.1 – O usuário seleciona um arquivo PDF e o seu conteúdo é exibido na tela;
A.2 – O usuário selecionar a opção “Aplicar PDFBox” e o sistema exibe o resultado, permitindo que o usuário visualize os dados em formato de texto;
A.3–O sistema continua da sessão P.6 do fluxo principal.

Fluxos de Exceção

Nenhum.

Pós-condições

Informações dos artigos são extraídas e armazenadas na base de dados.

Requisitos Não Funcionais

Nenhum.

Caso de Uso 08: Extrair Características dos Arquivos PDF ou TXT

Objetivo

Permitir que o usuário possa extrair informações dos arquivos, quer estes estejam no formato PDF ou TXT(JPEG, PNG, TIFF).

Atores

Usuário – usuário do sistema, aquele que irá selecionar os artigos;
Software OCR – aquele que irá extrair os dados dos arquivos em formato de imagem;
PDFBox –aquele que irá extrair os dados dos arquivos em formato PDF.

Pré-Condições

Os arquivos deverão ter os formatos PDF, JPEG, PNG ou TIFF.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão “Extrair Características”;
P.2 – O usuário seleciona um arquivo de imagem e o seu conteúdo é exibido na tela;
P.3 – O usuário pode aplicar filtros (cinza, preto e branco, extrair borda, extrair ruído sal e pimenta) e o sistema exibe e o resultado, permitindo que o usuário decida qual a melhor opção;
P.4 – Ao finalizar o sistema salvar as modificações no arquivo;
P.5 – O usuário selecionar a opção “Aplicar OCR” e o sistema exibe o resultado, permitindo que o usuário visualize os dados em formato de texto;
P.6 – O usuário pode selecionar a opção “Extrair Informações” e ao final o sistema irá identificar: título do artigo, autores, palavras-chave, referências;
P.7 – O sistema armazena no banco de dados as informações extraídas do arquivo.

Fluxos Alternativos
A.1 – O usuário seleciona um arquivo PDF e o seu conteúdo é exibido na tela;
A.2 – O usuário selecionar a opção “Aplicar PDFBox” e o sistema exibe o resultado, permitindo que o usuário visualize os dados em formato de texto;
A.3 – O sistema continua da sessão P.6 do fluxo principal.

Fluxos de Exceção
Nenhum.

Pós-condições
Informações dos artigos são extraídas e armazenadas na base de dados.

Requisitos Não Funcionais
Nenhum.

Caso de Uso 09: Gerenciar Dados (Referências, Autor, Resumo e Palavras-Chave)

Objetivo

Permitir que o sistema possa relacionar as referências citadas nos artigos com os artigos cadastrados na base. Da mesma forma permite o gerenciamento dos dados sobre: autor, resumo e palavras-chave.

Atores

Usuário – usuário do sistema, aquele que irá selecionar a opção "Gerenciar Referências/Autor/Resumo/Palavras-Chave";

Classificador – este é um papel assumido pelo próprio sistema.

Pré-Condições

É necessário que existam várias edições com artigos cadastradas na base de dados.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão "Gerenciar Referências" ou "Gerenciar Autor" ou "Gerenciar Resumo" ou "Gerenciar Palavras-Chave";

P.2 – O classificador seleciona todas as referências/autores/resumos/palavras-chave cadastradas e seleciona as que foram mais citadas criando um *rank* com os artigos mais citados, assim como os autores mais referenciados ou autores que mais publicaram ou palavra-chave mais citadas ou a lista dos resumos;

P.3 – O sistema armazena os dados e os exibe ao usuário.

Fluxos Alternativos

Nenhum.

Fluxos de Exceção

E.1 – Não há referências/autores/resumo/palavras-chavescadastradas;

E.2 – O sistema indica os eventuais erros ao usuário.

Pós-condições

O *rank* de citações é gerado.

Requisitos Não Funcionais

Nenhum.

Caso de Uso10: Digitalizar Artigos em Papel

Objetivo

Permitir que o usuário possa digitalizar artigos em formato de papel.

Atores

Scanner – aquele que irá digitalizar os arquivos, gerando arquivos de imagens;

Pré-Condições

Os arquivos deverão estar no formato de papel.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão "Aplicar Scanner";

P.2 – O sistema chamará a interface do scanner e o arquivo gerado será exibido no sistema;

P.3 – O usuário pode selecionar a opção "Extrair Informações" e ao final o sistema irá identificar: título do artigo, autores, palavras-chave, referências;

P.4– O sistema armazena no banco de dados as informações extraídas do arquivo.

Fluxos Alternativos

Nenhum.

Fluxos de Exceção

E.1 – Scanner desligado ou incompatível com o sistema;

E.2 – O sistema indica os eventuais erros ao usuário.

Pós-condições

Artigos digitalizados.

Requisitos Não Funcionais

Nenhum.

Caso de Uso 11: Aplicar Filtros

Objetivo

Permitir que o usuário possa aplicar filtros nas imagens (JPEG, TIFF ou PNG), com o intuito de melhorá-las.

Atores

BigBatch –*software* para tratamento de imagens.

Pré-Condições

Os arquivos deverão ter os formatos JPEG, PNG ou TIFF.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão "Aplicar Filtros" e o sistema chama os módulos do BigBatch;

P.2 – O usuário seleciona um arquivo de imagem e o seu conteúdo é exibido na tela;

P.3 – O usuário pode selecionar os filtros (cinza, preto e branco, extrair borda, extrair ruído sal e pimenta) e o sistema exibe o resultado, permitindo que o usuário decida qual a melhor opção;

P.4 – Ao finalizar o sistema salvar as modificações no arquivo;

Fluxos Alternativos

A.1 – O usuário pode aplicar os filtros que foram implementados no próprio sistema;

A.3 – O sistema continua da sessão P.2 do fluxo principal.

Fluxos de Exceção

Nenhum.

Pós-condições

Imagens dos artigos modificadas.

Requisitos Não Funcionais

Nenhum.

Caso de Uso 12: Acessar Consultas e Relatórios

Objetivo

Permitir que o usuário possaconsultar e/ou imprimir as informações contidas na base de dados.

Atores

Usuário – usuário do sistema, aquele que irá consultar a base de dados.

Pré-Condições

Dados armazenados na base de dados.

Fluxo Principal

P.1 – Este caso de uso se inicia quando o usuário acessa o sistema e clica no botão “Relatório”;

P.2 – O usuário seleciona o tipo de relatório e o sistema exibe os dados na tela e o usuário pode optar por imprimir os dados;

P.3 – O sistema redireciona o ator para o “Menu Principal”.

Fluxos Alternativos

Nenhum.

Fluxos de Exceção

Nenhum.

Pós-condições

Nenhuma.

Requisitos Não Funcionais

Nenhum.

APÊNDICE B–Mapas Conceituais do Modelo de Referência da DELOS

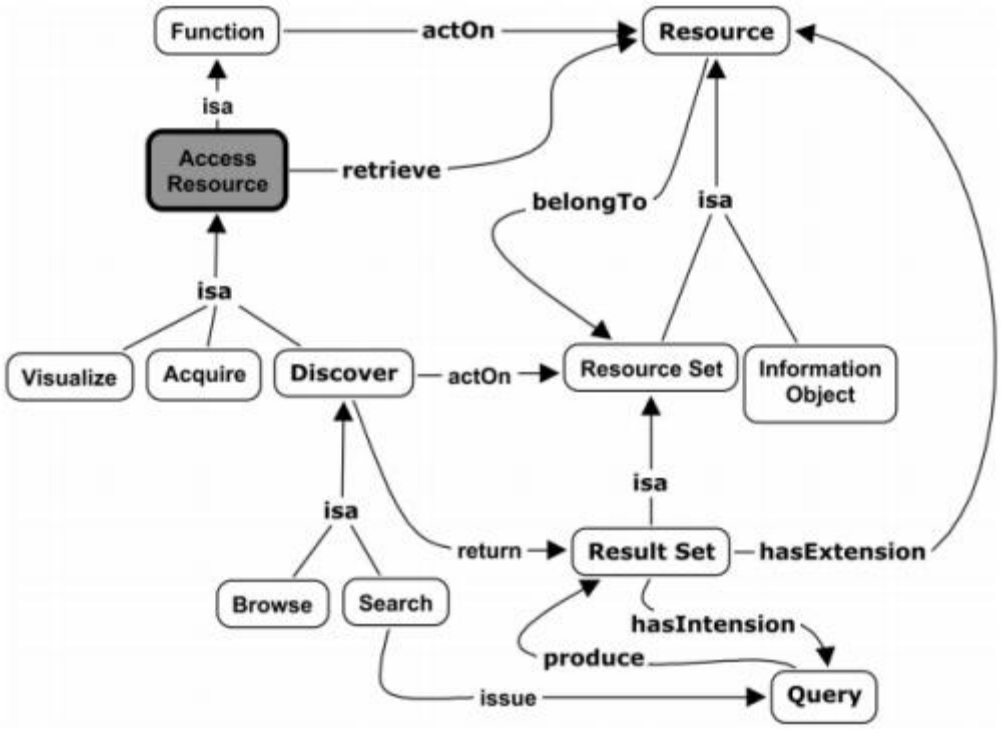


Figura 57 – *Mapa Conceitual: Recursos (CANDELA et al., 2008).*

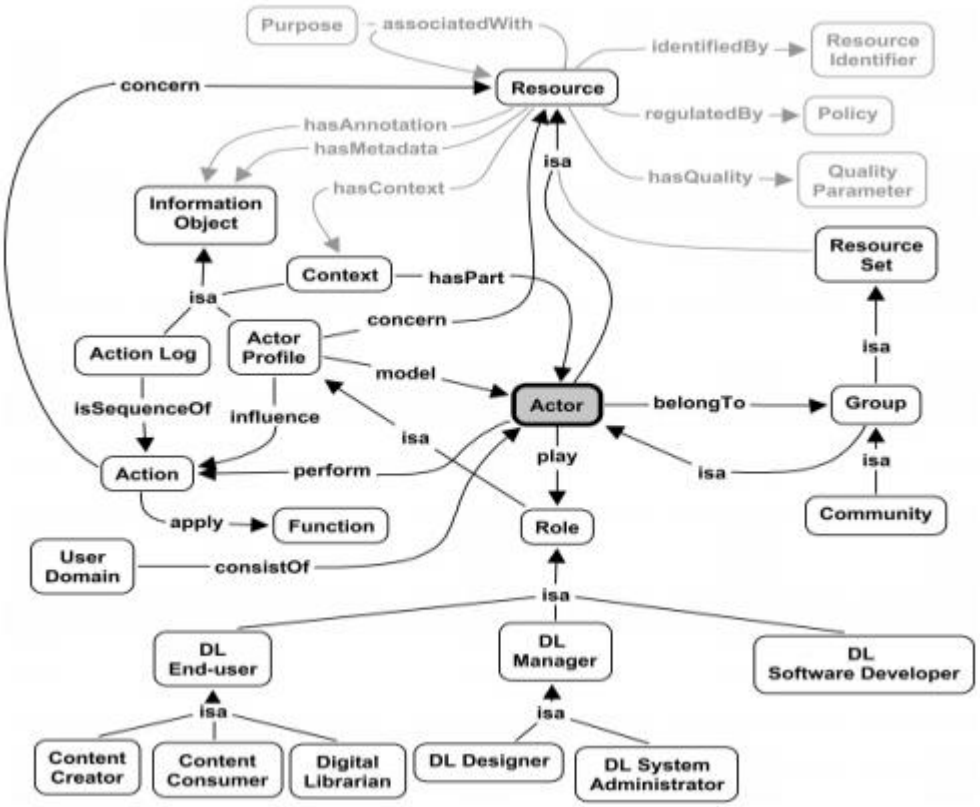


Figura 58 – *Mapa Conceitual: Domínio Ator (CANDELA et al., 2008).*

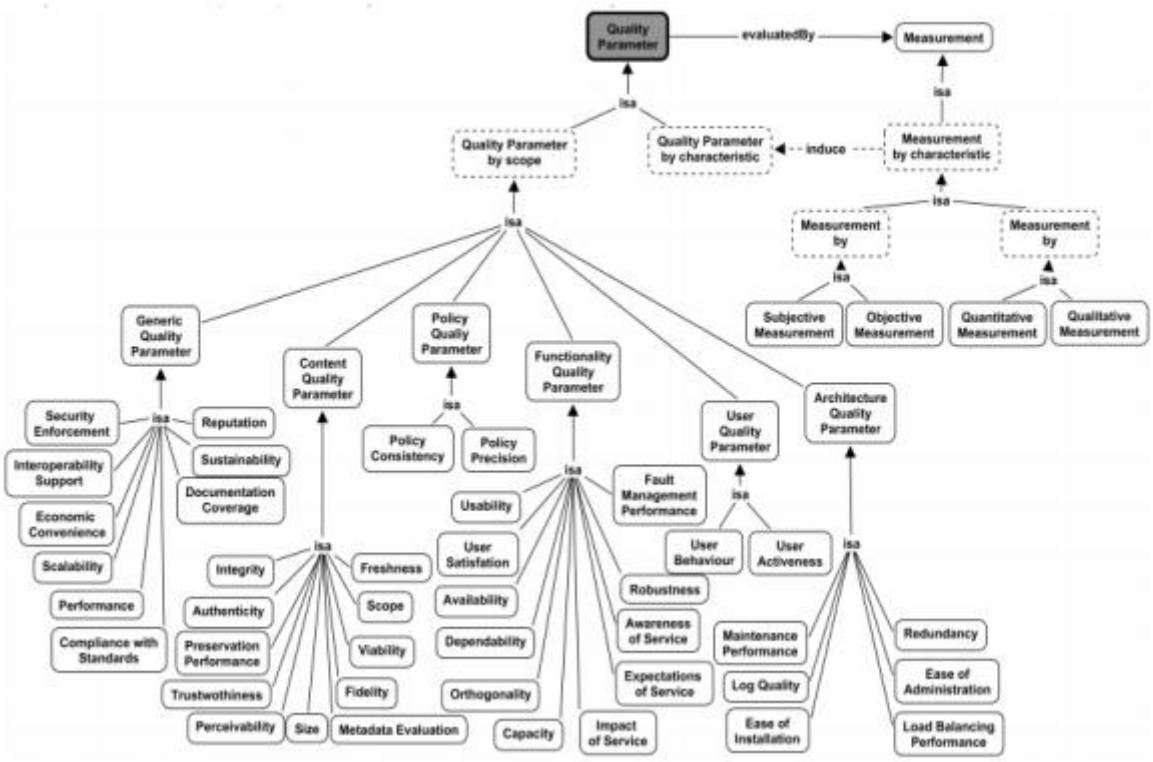


Figura 59 – *Mapa Conceitual: Domínio dos Parâmetros de Qualidade (CANDELA et al., 2008).*

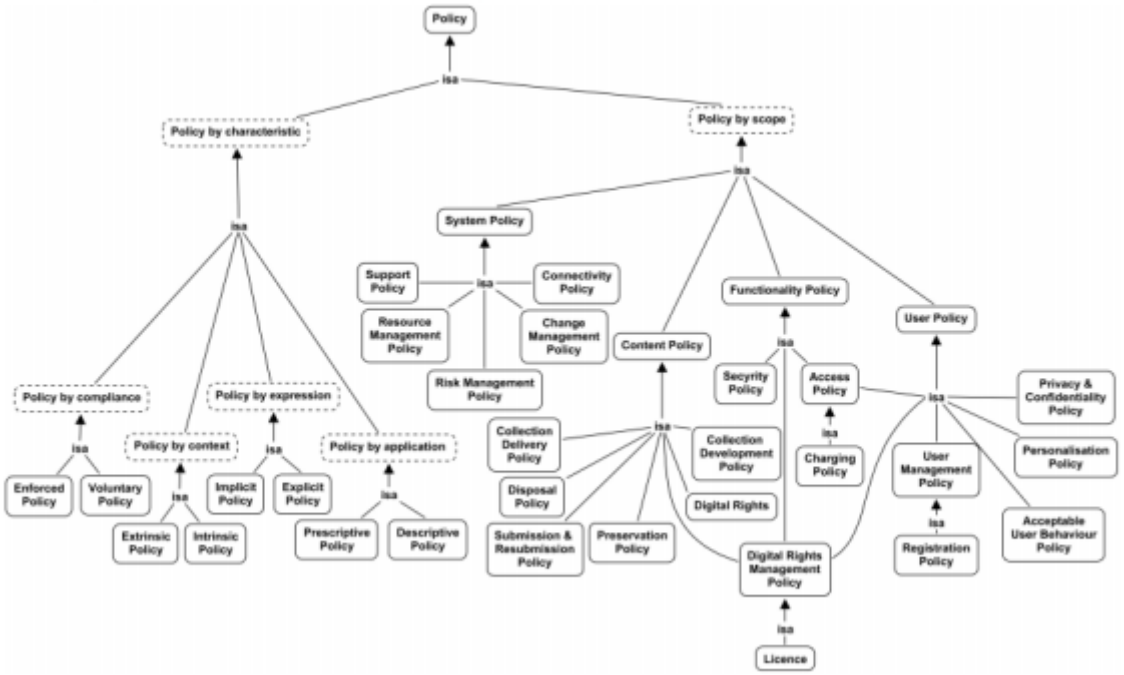


Figura 60 – *Mapa Conceitual: Domínio de Política (CANDELA et al., 2008).*

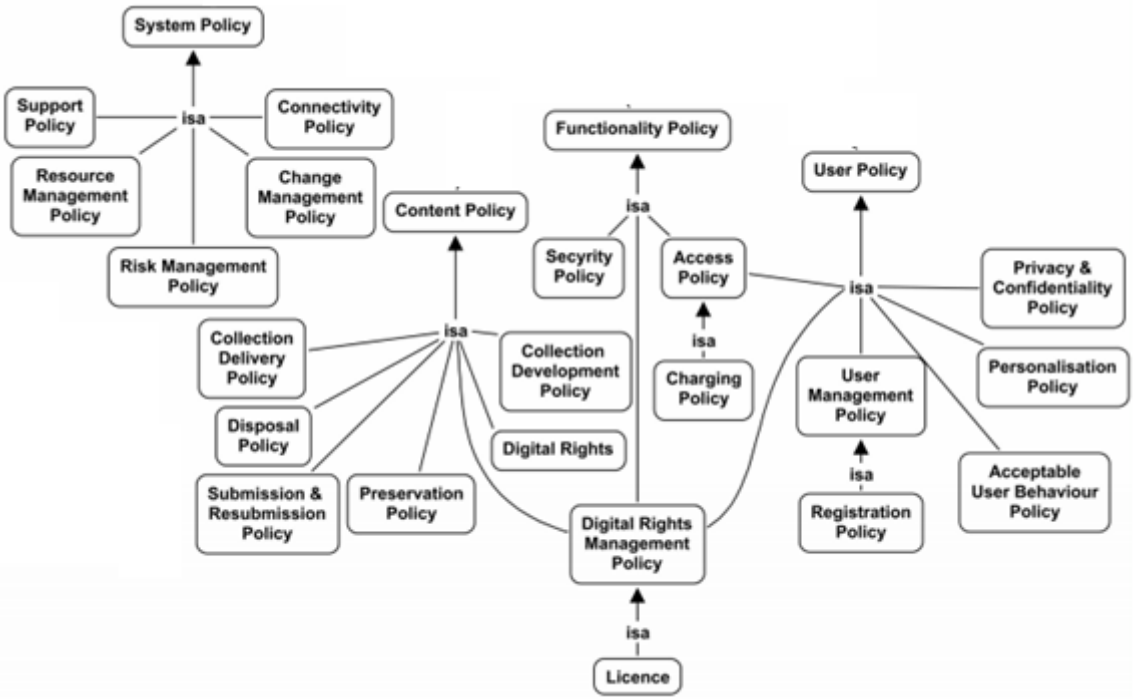


Figura 61 – *Mapa Conceitual: Domínio do Sistema de Política (CANDELA et al., 2008).*

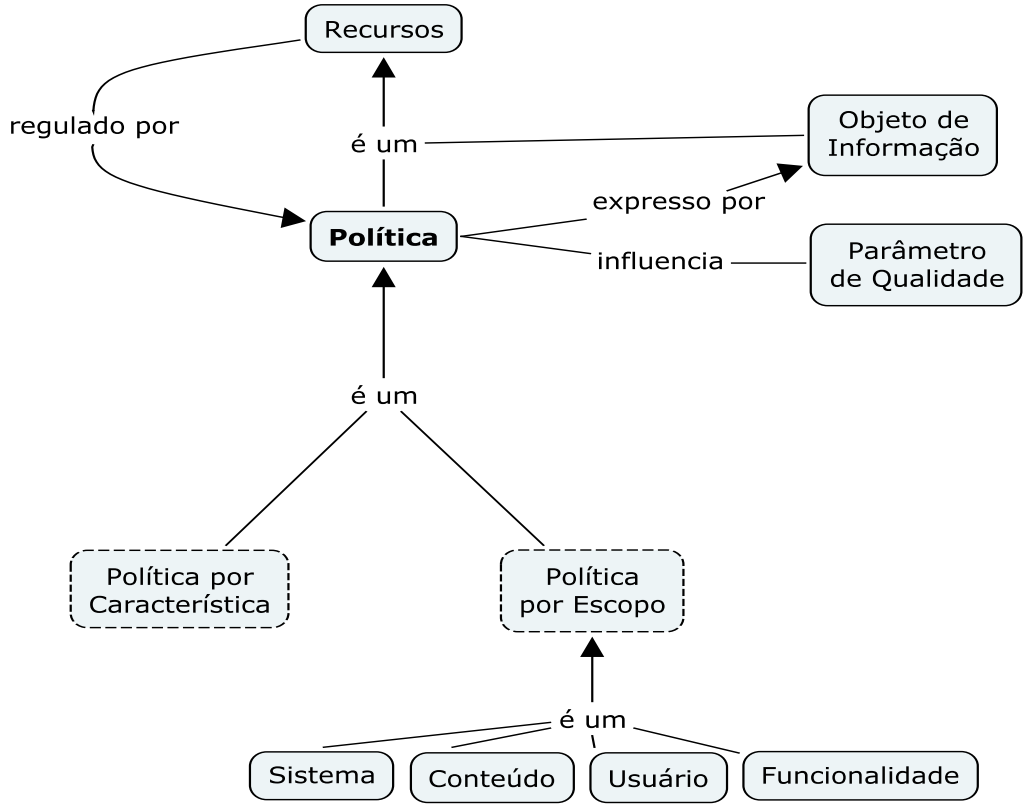


Figura 62 – *Mapa Conceitual das Políticas do LiveMemory.Adaptado de Candela et al. (2008).*

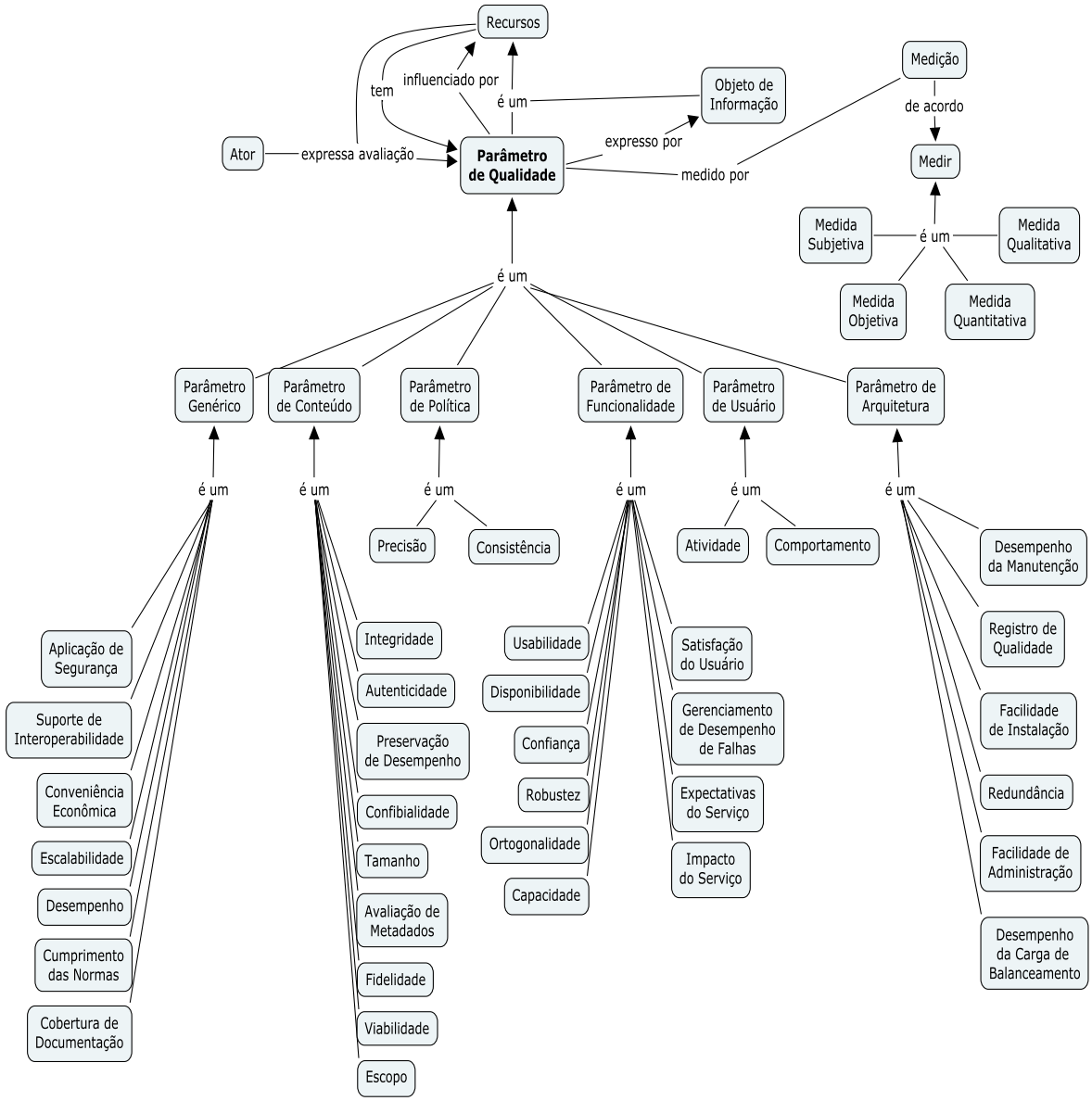


Figura 63 – *Mapa Conceitual dos Parâmetros de Qualidade do LiveMemory.Adaptado de Candela et al. (2008).*

APÊNDICE C–Modelo Lógico

Após análise dos principais requisitos do sistema, bem como dodiagrama de classe, que representou o modelo conceitual do sLiveMemory, foi elaborado o modelo lógico com 21 tabelas, conforme Figura 64. Vale ressaltar, que a tabela Artigo é a principal, pois armazena as informações sobre os artigos. Cada artigo pode ter várias referências, bem como vários autores, resumos e palavras-chave, também está relacionado a uma edição e possui um idioma. As demais tabelas auxiliam no armazenamento dos dados.

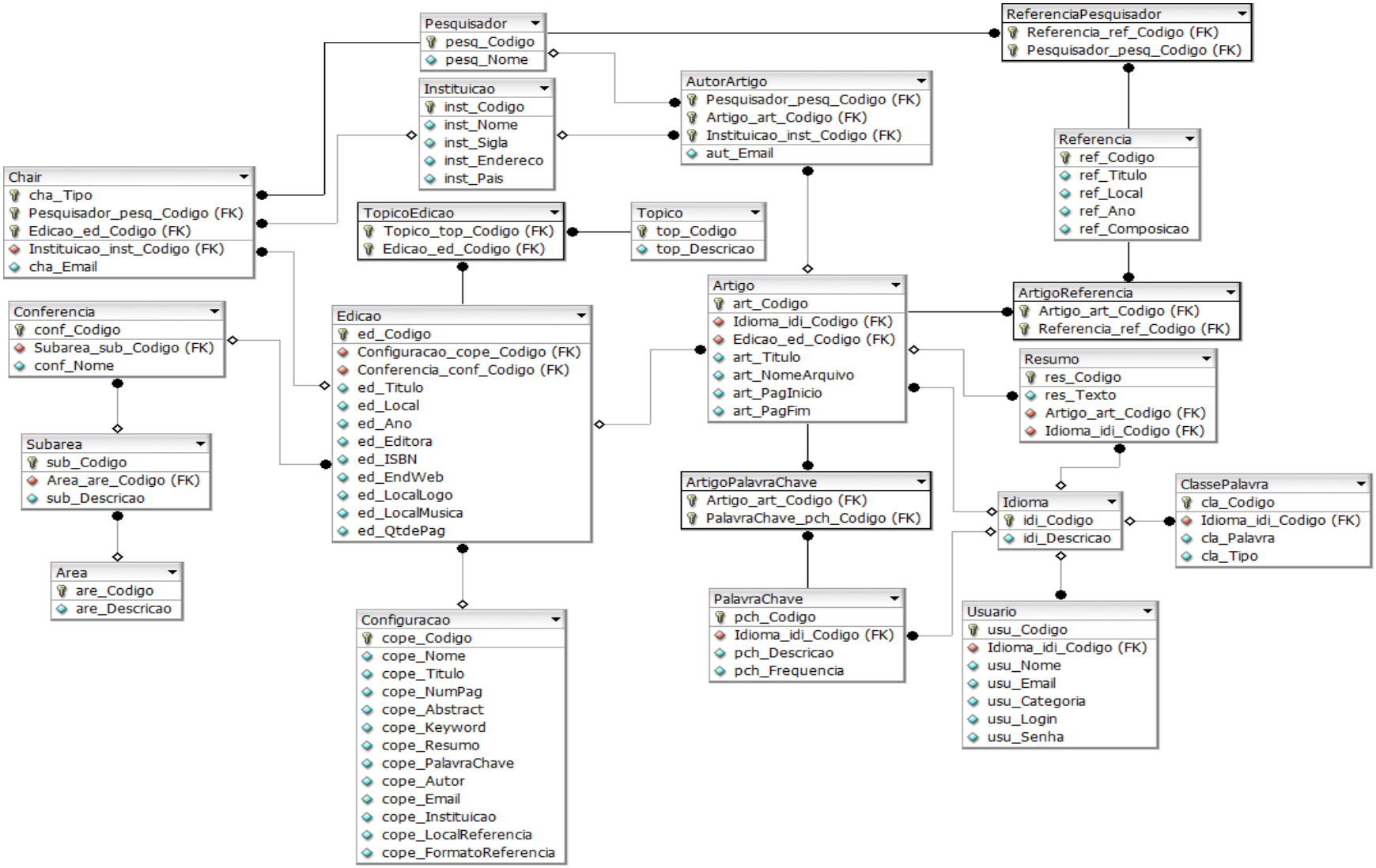


Figura 64 – *Modelo Lógico do LiveMemory.*