SISDOC: Uma Biblioteca Digital Multimídia para Gestão de Documentos Ambientais

Hugo Feitosa de Figueirêdo[1], Ruan Victor Amorim[2], Cláudio de Souza Baptista[2], Roberta Falcão de Cerqueira Paes[3], Brígida Duarte[3], Murilo Sérgio Lucena Pinto[3],

[1]Instituto Federal de Ciência, Educação e Tecnologia da Paraíba (IFPB)
Esperança – PB – Brasil

[2]Companhia Hidro Elétrica do São Francisco (CHESF)
Recife - PE - Brasil.

[3]Universidade Federal de Campina Grande (UFCG)
Campina Grande- PB - Brasil

Abstract. *In this article, we propose a new method for managing documents generated by processes related to environmental management in the scope of a power generation company. The new method allows the optical character recognition, indexing the content, metadata extraction, annotation of attribute values created by a dynamic schema, image processing, highlighting, temporal filter and conversion to standard formats. The proposed method was implemented in the Environment Department of the Hydroelectric Company of São Francisco (Chesf) and improves the efficiency of document management, saving computer system resources (hardware, software and peopleware) adopted for management of multimedia documents, and expediting the environmental licensing procedures and supervision of reservoirs edges.*

Resumo. *Neste artigo, propõe-se um novo método para gerenciamento de documentos gerados por processos relacionados à gestão ambiental no escopo de uma empresa de geração de energia elétrica. O novo método permite o reconhecimento óptico de caracteres, a indexação pelo conteúdo, extração de metadados, anotação com valores de atributos criados por um esquema dinâmico, tratamento de imagem, destaque de termos encontrados em documentos, filtro temporal e conversão para formatos padrões. O método proposto foi implantado no Departamento de Meio Ambiente da Companhia Hidrelétrica do São Francisco (CHESF) e permite melhorar a eficiência na gestão de documentos, economizando recursos do sistema computacional (hardware, software e peopleware) adotados para gerenciamento de documentos multimídia, bem como agilizando os processos de licenciamento ambiental e fiscalização de bordas de reservatórios.*

1. Introdução

Bibliotecas digitais são um conjunto de mecanismos eletrônicos que facilitam a localização da demanda informacional, interligando recursos e usuários (Cunha, 2008). Estas bibliotecas têm sido propostas para gerenciar os recursos eletrônicos de forma distribuída. As bibliotecas digitais proveem dados e serviços e lidam com a complexidade de dados multimídias tais como imagens, textos, áudio e vídeo (de Vries,

Eberman, & Kovalcin, 1998).

Um dos grandes diferenciais das bibliotecas virtuais é o suporte a vários formatos de informações como textos, áudio, vídeo e imagens (Andrade, 2010). Este conjunto de dados heterogêneos requer mecanismos de indexação, consulta, apresentação e análise. Tais dados demandam um grande volume de repositório de dados e um dos maiores desafios enfrentados está no desenvolvimento de soluções para lidar com este volume de dados de forma eficiente, integrada, interoperável e distribuída. Por eficiência entende-se que o uso de informação multimídia envolve algoritmos paralelos, otimização de consultas e técnicas de indexação. Por integração entende-se o uso de modelos de dados que transpassam todos os tipos de objetos: simples e complexos. Por distribuição requer-se que o sistema possa prover de forma transparente para os usuários, um meio de acessar remotamente os dados, que devido aos custos de aquisição e manutenção, limita a replicação.

As bibliotecas digitais surgiram primeiro no meio acadêmico e depois foram difundidas na indústria, sendo as primeiras instituições a adotarem esse recurso no Brasil a Universidade de São Paulo (USP), Universidade de Campinas, Universidade de Brasília (UnB), entre outras (Andrade, 2010).

Uma solução que tem sido investigada é o uso de metadados para descrever a semântica, sintaxe e estrutura dos dados. Estes metadados ganharam aceitação em domínios de aplicação distintos tais como museus, bibliotecas, observatórios espaciais, dentre outros. Particularmente, algumas propostas surgiram para a gestão de metadados em bibliotecas digitais tais como: Dublin Core (DCMI, 2012), RDF (Lassila, 1998) e XML (Prescod, P. e Goldfarb, 1999).

Tais metadados descrevem a essência, os atributos e o contexto de um recurso eletrônico, além de modelar relacionamentos entre documentos, não se restringindo a dados que estejam contidos no texto, mas também podendo incluir dados da descrição física e o contexto de produção (Alvarenga, 2006).

Algumas vezes torna-se difícil separar dados de metadados (Huc, Levoir, & Nonon-Latapie, 1997). Por exemplo, alguém pode indagar se um *thumbnail* de uma fotografia pode ser considerado dado ou metadado. O mesmo se aplica ao resumo de um documento, um trailer de um filme. Neste sentido, Gunther e Voisard dizem: "Não existe uma distinção intrínseca entre dados e metadados, é uma questão de contexto indicar se um dado item representa um metadado ou não" (Gunther & Voisard, 1998).

Alvarenga (2006) destaca a importância que os metadados sejam elaborados por uma equipe técnica especializada que conheça as características do acervo, a realidade da biblioteca e de seus usuários, a fim de proporcionar uma melhor recuperação da informação. Por sua vez, documentos relacionados ao meio ambiente possuem uma grande variação nos metadados que podem ser utilizados em cada documento multimídia, por exemplo, com relação ao processo de licenciamento ambiental, cada órgão licenciador possui seu trâmite e informações distintas necessárias para a obtenção da licença. Dessa forma, uma biblioteca digital para área de meio ambiente deve permitir alteração do esquema dinamicamente por pessoas técnicas da área de meio ambiente e não da área de informática ou de ciência da informação.

Para a gestão ambiental e sociopatrimonial, uma empresa gera uma grande quantidade de documentos multimídia, sendo em sua grande maioria dos tipos textuais, áudios, vídeos e imagens. A agilidade na recuperação de um documento multimídia pode facilitar o trabalho dos funcionários da empresa na gestão ambiental e

sociopatrimonial. Outro ponto a destacar é o desperdício advindo por uma má gestão de documentos, que ocasiona aumento dos custos com material permanente (e.g., hardware), de consumo (e.g., papel, tinta) e recursos humanos (e.g., técnicos de informática, analistas). Com isso, o bom gerenciamento dos documentos multimídia torna-se essencial para o setor de gestão ambiental de uma empresa.

Neste artigo, propõe-se um novo processo para gestão de documentos na área de meio ambiente, como também a implementação de uma biblioteca digital multimídia que implementa este processo. O sistema foi testado junto ao departamento de meio ambiente da Companhia Hidrelétrica do São Francisco (CHESF).

O restante deste artigo está organizado como segue. Na seção 2, é apresentado o novo processo de gestão de documentos multimídia relacionados ao meio-ambiente. Na seção 3, é apresentada a biblioteca digital multimídia para área de meio-ambiente. Na seção 4, discute-se a avaliação do sistema proposto. Finalmente, na seção 5, são apresentadas as conclusões e trabalhos futuros.

2. Um novo Processo de Gestão de Documentos Multimídia Relacionados à Área de Meio-Ambiente

Na Figura 1, apresenta-se o modelo do processo de gestão de documentos multimídia proposto neste trabalho. Inicialmente, devem-se cadastrar os atributos e grupos de atributos para cada tipo de documento multimídia utilizado na gestão ambiental e sociopatrimonial. Em seguida, novos documentos podem ser inseridos por uma pessoa física ou por sistemas.

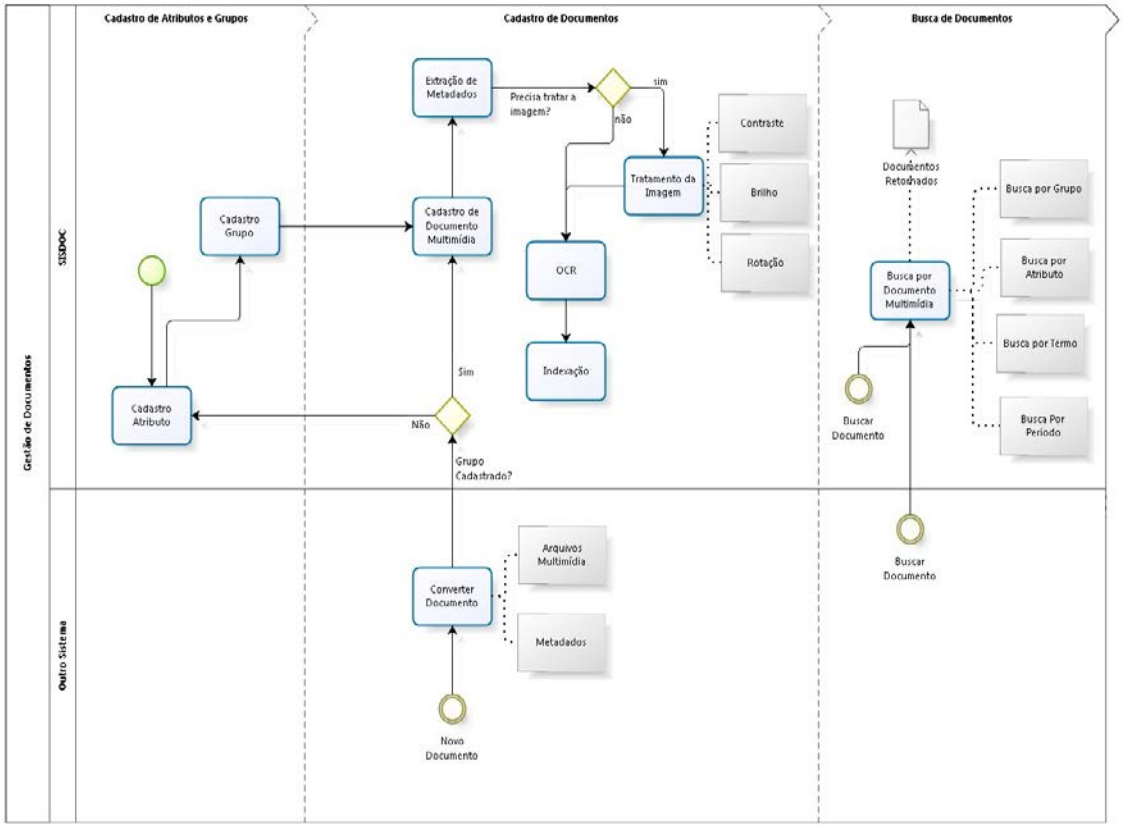

Figura 1. Modelagem do processo gestão de documentos multimídia.

A ideia central para uma boa gestão de documentos multimídia é a centralização do gerenciamento, sendo necessário que outros sistemas, que necessitem lidar com carga, geração e descarga de documentos, utilizem uma biblioteca digital multimídia central.

Quando um sistema precisar utilizar a gestão de documentos multimídia, deverá encapsular junto aos arquivos multimídia todas as informações importantes para anotação e enviar para a biblioteca digital multimídia central. Se o grupo para o qual o documento foi enviado não tiver sido cadastrado ainda, deverá ser criado na biblioteca um novo grupo com os atributos encapsulados passados. Em seguida, o documento será cadastrado na biblioteca.

O processo de cadastro pode necessitar de uma extração dos metadados dos arquivos, tratamento das imagens que podem ser melhoradas mediante alteração de contraste, brilho rotação e recorte, reconhecimento óptico dos caracteres (OCR) para extração de texto contido em imagens e, por fim, indexação. A busca de documentos poderá ocorrer por filtros de grupos, atributos, períodos de tempo e termos contidos nos documentos multimídia.

Para implantação deste processo de gestão de documentos multimídia foi implementado o SISDOC - Sistema de Gestão de Documentos Multimídia, a ser detalhado na próxima seção.

3. SISDOC – Sistema de Gestão de Documentos Multimídia

Na Figura 2, apresenta-se um diagrama com as funcionalidades e atores relacionados à biblioteca digital multimídia aqui proposta. Os atores do SISDOC podem ser classificados em três tipos: sistema externo, administrador e usuário pesquisador. Os sistemas externos representam outros aplicativos que utilizarão o SISDOC como repositório de arquivos ou enviarão os documentos criados e anexados para indexação. Como sistemas externos foram integrados aos SISDOC os seguintes sistemas: Sistema de Licenciamento Ambiental - SISLIC (Santana et al., 2015), Sistema de Fiscalização de Uso e Ocupação de APPs de Bordas de Reservatório - SISBORDAS (Paiva et al., 2015) e Sistema de Educação Sócio-Ambiental - SIPAS (ALVES, BAPTISTA, LEITE, CHAGAS, & PAIVA, 2014). Os administradores podem realizar cadastro de documentos e são os responsáveis pelo cadastro dos atributos e grupo de atributos necessários para o cadastro de documentos. O usuário pesquisador é o que realizará as buscas por documentos no SISDOC.

As principais funcionalidades providas pelo SISDOC são: cadastro de documentos, cadastro de atributos, cadastro de grupo de atributos e busca de documentos. Algumas funcionalidades secundárias são providas a partir da funcionalidade de cadastro de documentos: extração de conteúdo e extração de metadados.

Na Figura 3, apresenta-se a arquitetura do SISDOC, sendo baseada no padrão MVC e arquitetura N-Camadas. Uma implementação da especificação JavaServer Faces 2.0 (JSF) foi utilizada para a adoção do MVC e o Primefaces 3.5 (conjunto de componentes para interfaces gráficas de aplicações baseadas em JSF) foi utilizado para geração da interface gráfica com o usuário (GUI).

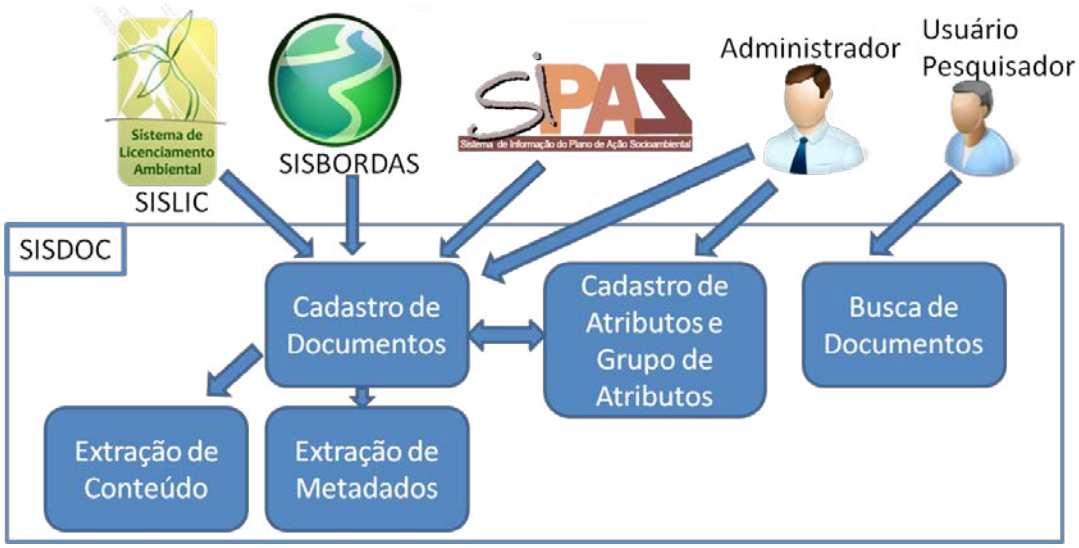

Figura 2 - Visão geral do SISDOC.

O SISDOC foi desenvolvido para ser executado em uma plataforma Java EE 5 e Java SE 6, sendo testado utilizando-se o servidor de aplicações Oracle Weblogic 10.3.5.

Na camada de dados, utiliza-se o Apache Solr 4.7 para indexação textual dos arquivos adicionados pelos usuários ou sistemas externos (SISLIC, SISBORDAS e SIPAS). O Solr é uma aplicação web que foi testada no Weblogic 10.3.5 que possui uma comunicação via REST com outros sistemas. Para comunicação entre o Solr e a camada de acesso aos dados (ver Figura 3), utiliza-se o Solrj 4.7 mediante a utilização de REST. Os arquivos adicionados e os convertidos são adicionados em uma pasta no sistema de arquivos do sistema operacional.

Para viabilizar a comunicação entre o SISDOC e as aplicações externas, foi desenvolvido um *Web Service*, permitindo a realização de intercâmbio de arquivos entre as ferramentas, seguindo o conceito de Cliente e Servidor. Para implementação do *Web Service* foi utilizado o conceito de *REST*. O framework Jersey 1.8 foi utilizado para a comunicação em REST, sendo esse uma implementação da especificação JAX-RS.

Inicialmente, um sistema externo, assumindo o papel de "cliente" na comunicação, envia um arquivo e seus metadados, por meio da Web, para o SISDOC. O SISDOC recebe o arquivo enviado e realiza todos os procedimentos necessários para indexação em seu sistema. Caso atenda todos os requisitos mínimos, o arquivo será persistido e um identificador único será gerado. Tal identificador será repassado ao cliente que realizou a requisição de envio de arquivo. Por sua vez, depois de recebida a confirmação satisfatória de envio, o sistema externo deve realizar o registro junto com o identificador recebido como resposta. Na Figura 4, é apresentado todo o processo de comunicação entre os sistemas utilizando o Sistema de Gestão de Licenciamento Ambiental (SISLIC) da Companhia Hidrelétrica do São Francisco (CHESF) como exemplo.

Alguns documentos adicionados no SISDOC possuem imagens com texto, e.g., licença digitalizada expedida por algum órgão de licenciamento ambiental. Para essas situações, a imagem passa por um processo de OCR (reconhecimento óptico de caracteres) antes da indexação, sendo esse processo realizado pelo Tesseract 3.02. Esta aplicação é externa à aplicação implantada no Weblogic. Para comunicação entre o

Tesseract e a aplicação no Weblogic foi utilizado o Tess4J 1.4.

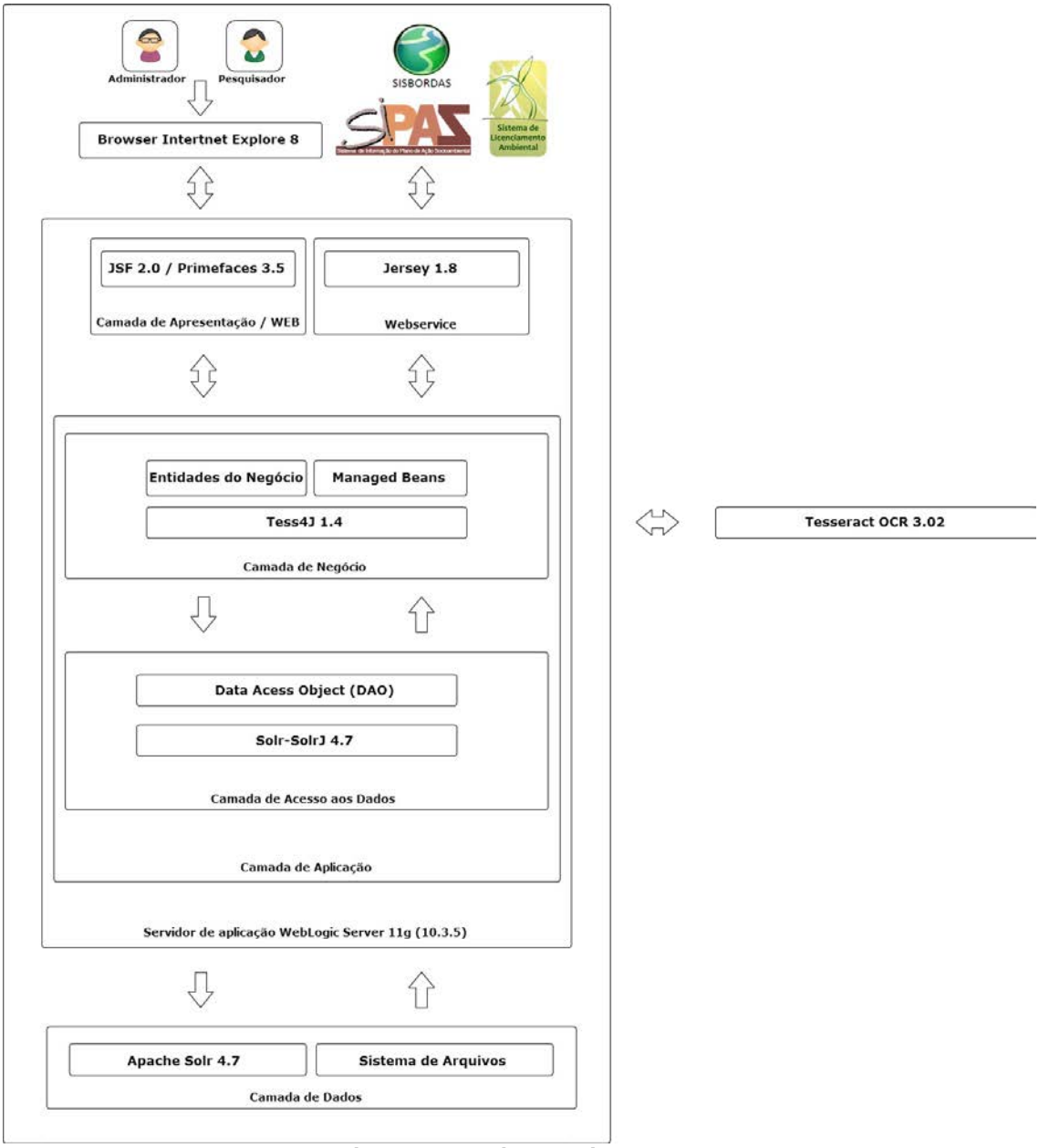

Figura 3. Arquitetura do SISDOC.

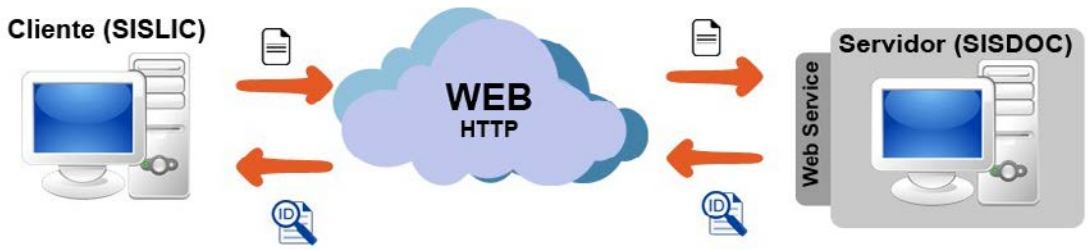

Figura 4: Comunicação entre Cliente e Servidor (SISDOC)

3.1. Mapeamento de conteúdo em documento PDF

Muitos dos documentos textuais que estarão armazenados no sistema são compostos por imagens e não possuem o reconhecimento óptico do conteúdo. Assim, foi elaborado um algoritmo que realiza o mapeamento do texto extraído do conteúdo das imagens em um arquivo PDF. Com este arquivo PDF, o conteúdo das imagens pode ser copiado para área de transferência e também pesquisado no próprio documento mediante os programas adequados, como o Adobe Reader via a tecla de atalho Ctrl+F.

O algoritmo utiliza reconhecimento óptico de caracteres para reconhecer e localizar as posições das palavras contidas nas imagens. Em seguida, um arquivo no formato PDF é gerado com a parte textual inserida de forma transparente nas posições adequadas.

Para a realização de consultas que retornem documentos que possuem um determinado texto em seu conteúdo ou nos metadados de forma eficiente, torna-se necessária a utilização de um mecanismo de indexação.

A indexação é um processamento realizado nos arquivos para extrair os metadados e conteúdo e criar uma estrutura de dados eficiente para a busca de documentos por palavras presentes no documento. Para um melhor gerenciamento dos documentos, torna-se necessário um servidor de indexação textual que possui as seguintes características: indexação e busca de documentos, destaques das informações em formato highlighting, exibição de resultados de uma busca em formato categorizado, integração com banco de dados, realização de buscas espaciais, entre outras.

4. Avaliação do SISDOC

O SISDOC foi implantado na Companhia Hidrelétrica do São Francisco no Departamento de Meio Ambiente com o objetivo de gerenciar os documentos multimídia desse setor, validando o processo proposto neste artigo. Além disso, o SISDOC foi integrado a outros sistemas da CHESF.

Outra avaliação realizada foi para determinar o quão eficiente é a técnica de busca de documentos por conteúdo. Nesta seção, apresentam-se a metodologia utilizada no processo de avaliação e os resultados obtidos a partir dos experimentos.

4.1. Metodologia

Tendo em vista a necessidade de avaliar a eficácia do método proposto, foi disponibilizada pela CHESF uma base de dados de documentos multimídia. Estes, oriundos de câmeras fotográficas e scanners, não possuem reconhecimento de caracteres e estão disponíveis em formato de imagem (JPG, PNG ou TIF) e PDF.

Para restringir o escopo da avaliação, a base de dados precisou ser particionada nas seguintes categorias: Autorização, Comprovante, Licença, Parecer e Protocolo. Os documentos são classificados de acordo com a representação do seu conteúdo e devem ser atribuídos a uma das categorias citadas.

Previamente, foi realizada uma análise textual dos arquivos cujo objetivo foi identificar padrões de texto que são comuns a uma ou mais categorias de documentos. Foi necessário construir uma tabela de resultados esperados, onde documentos e padrões são correlacionados. A utilidade desta tabela é auxiliar a verificação dos resultados da pesquisa a partir de um texto dado como entrada.

A Tabela 1 exemplifica como os resultados esperados estão dispostos. Os valores

da primeira coluna representam os padrões de texto que serão inseridos na entrada do método proposto. A segunda coluna representa os documentos multimídia que devem ser retornados na pesquisa quando o padrão relacionado for pesquisado.

Tabela 1. Exemplo dos resultados esperados para a categoria Autorização.

Padrões de Texto	Documentos
"Autorização Especial"	Doc-Aut-1.pdf
"Desenvolvimento Sustentável"	Doc-Aut-2.png, Doc-Aut-3.jpg, Doc-Aut-4.tif
"*Termo de autorização*"	Doc-Aut-1.pdf, Doc-Aut-2.png

Na Figura 5, demonstra-se a realização do experimento para um padrão selecionado. Na barra de pesquisa superior, define-se o padrão a ser pesquisado nos documentos indexados. No quadro inferior é demonstrado o resultado da pesquisa destacando o nome dos documentos e o conteúdo que foi relacionado ao padrão textual.

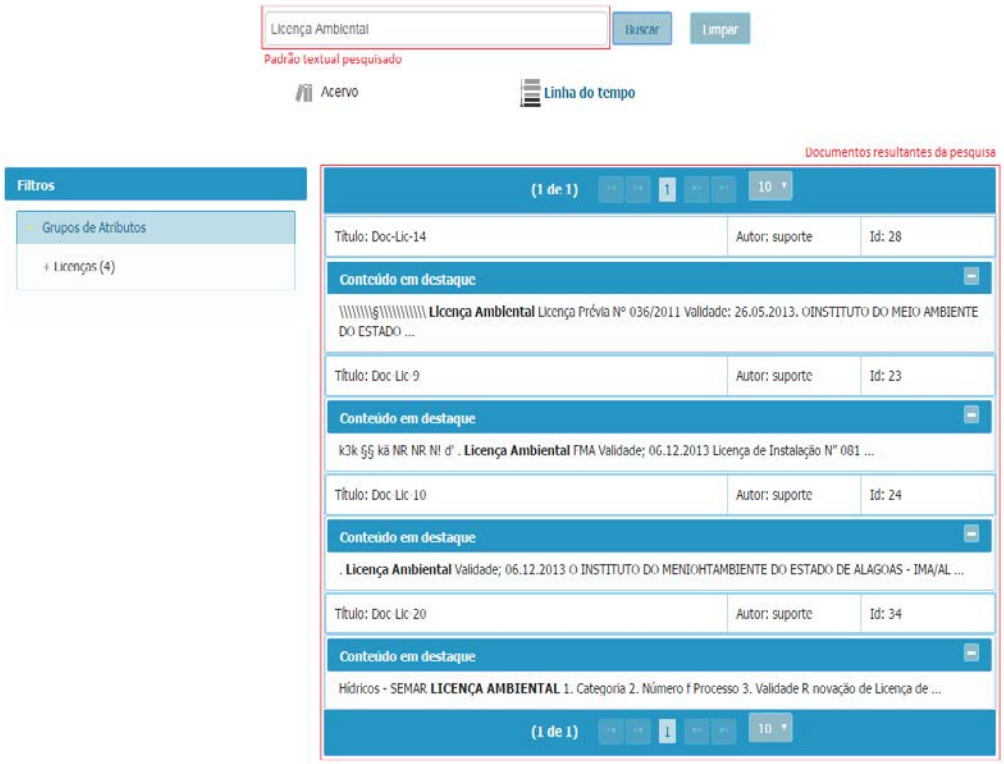

Figura 5. Exemplo de pesquisa e demonstração de resultados.

Durante a execução dos experimentos, deve-se coletar os resultados das pesquisas comparando-os com a tabela de resultados esperados. Dessa forma, é possível identificar quais documentos retornados são, de fato, relevantes. Com o número de documentos retornados e o número de documentos relevantes, as métricas de interesse: precisão, revocação e f-measure, podem ser calculadas, e assim possibilitará a avaliação da eficácia do método proposto.

4.2. Resultados

Na Tabela 2, *são* apresentados os resultados, agrupados por categorias, obtidos a partir dos experimentos. Os resultados são considerados excelentes para busca de documento por conteúdo, porém, os bons resultados do método proposto está condicionado à

qualidade visual dos documentos indexados. Documentos que possuem caracteres especiais (ex: º, ª, &), rasura, amasso, borrões ou outros defeitos, tendem a dificultar o processo de reconhecimento óptico de caracteres, prejudicando o reconhecimento de padrões textuais.

Tabela 2. Resultados da avaliação da técnica agrupado por categoria.

Categoria	Precisão	Revocação	F-Measure
Autorização	1	0,913	0,954
Comprovante	1	0,714	0,416
Licença	1	0,789	0,441

Em algumas categorias de documentos houve uma variação considerável nas métricas Revocação e F-Measure. O principal fator responsável por esta variação foi a qualidade visual dos documentos, que pode ser diferente dependendo da categoria a qual pertence. Os documentos das categorias Comprovante e Licença apresentaram defeitos como: caracteres especiais, amasso e textura irregular no background. Estes fatores contribuiram negativamente para o reconhecimento do conteúdo do documento, dessa forma, afetando as métricas supracitadas.

5. Conclusões

Neste artigo, foi proposto um processo para gerenciamento de documentos multimídia na área de meio ambiente.

O processo foi implementado por meio de um sistema intitulado SISDOC, o qual permite: (i) a evolução dinâmica do esquema, podendo ser modificado por especialistas da área de meio ambiente sem necessidade de conhecimentos avançados na área de informática; (ii) extração dos metadados contidos no documento multimídia; (iii) extração do conteúdo; (iv) busca por metadados; (v) busca por filtros temporais; (vi) mapeamento de conteúdo textual em imagens; (vii) visualização prévia de arquivos textuais, áudios, vídeos e imagens; e (viii) tratamento de imagens para melhoria de processo de reconhecimento óptico de caracteres.

O SISDOC foi implantado no departamento de meio ambiente da CHESF e foi integrado ao SISFAIXAS, SISBORDAS e SISLIC que são três soluções para a gestão do meio ambiente de uma empresa do setor elétrico que necessitam de um bom gerenciamento de arquivos.

Como pesquisas futuras, pretende-se avaliar as melhorias provindas pelo novo processo no departamento de meio ambiente da CHESF para gestão de documentos, adoção da solução para outros setores e inclusão de suporte a consultas espaciais.

Agradecimentos

Os autores agradecem o suporte financeiro da ANEEL, sob o contrato de P&D+I Nº ANEEL 0048-1119/2012.

Referências Bibliográficas

Alvarenga, L. (2006). Organização da informação nas bibliotecas digitais. *Organização Da Informação: Princípios E Tendências*, 76–98.

Alves, A. L. F., Baptista, C. D. S., Leite, D. F. B., Chagas, M. I. A., & Paiva, A. C. (2014). Aplicação de Geoprocessamento em um Sistema de Gestão Ambiental para Diagnóstico de Ações Socioambientais. In *Encontro Nacional de Geoprocessamento do Setor Elétrico - ENGEO*.

Andrade, N. S. de. (2010). Biblioteca digital: repositório de informação e conhecimento. *Fonte*, *7*(10). Retrieved from http://www.prodemge.mg.gov.br/images/com_arismartbook/download/11/revista_10.pdf#page=78

Cunha, M. B. da. (2008). Das bibliotecas convencionais às digitais: diferenças e convergências. *Perspectivas Em Ciência Da Informação*, *13*(1), 2–17. http://doi.org/10.1590/S1413-99362008000100002

DCMI. (2012). Dublin Core Metadata Initiative Metadata Terms. Retrieved April 4, 2013, from http://dublincore.org/documents/dcmi-terms/

de Vries, A., Eberman, B., & Kovalcin, D. (1998). The Design and Implementation of an Infrastructure for Multimedia Digital Libraries. In *Proceedings of the 1998 International Database Engineering and Applications Symposium* (pp. 103–120). IEEE.

Gunther, O., & Voisard, A. (1998). *Metadata in Geographic and Environmental Data Management*. (W. Klas & A. Sheth, Eds.). McGraw Hill.

Huc, C., Levoir, T., & Nonon-Latapie, M. (1997). Metadata: Models and Conceptual Limits. In *Proceedings of the Second IEEE Metadata Conference*. Maryland, USA: IEEE.

Lassila, O. (1998). Web Metadata: A Matter of Semantics. *IEEE Internet Computing*, 30–37.

Paiva, A. C. de, Campelo, C. E. C., Figueiredo, L. C. de, Rocha, J. H., Figueirêdo, H. F. de, & Baptista, C. de S. (2015). Management of Large Hydroelectric Reservoirs Surrounding Areas Using GIS and Remote Sensing. In A. Kő & E. Francesconi (Eds.), *Electronic Government and the Information Systems Perspective* (pp. 257–268). Cham: Springer. http://doi.org/10.1007/978-3-319-22389-6

Prescod, P. e Goldfarb, C. (1999). *The XML Handbook*. Prentice Hall PTR.

Santana, J. V., Figueirêdo, H. F. de, Baptista, C. de S., Paiva, A. C. de, Paes, R. F. de C., Pinto, M. S. L., & Duarte, B. (2015). SISLIC: Um metodo para gerenciamento do processo de licenciamento ambiental. In *VI Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*. Retrieved from http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Trabalho?id=23623