XIV Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB 2013)
GT 8: Informação e Tecnologia

Comunicação Oral

UMA PROPOSTA INOVADORA PARA GESTÃO DE OBJETOS DIGITAIS

Marcos Galindo – UFPE
Sandra de Albuquerque Siebra – UFPE
Vildeane da Rocha Borba – UFPE
Carlos Alexandre Barros de Mello – UFPE
Clauirton de Albuquerque Siebra – UFPB

Resumo
O conceito de gestão eletrônica de documentos (GED) tem se tornado cada vez mais importante na medida em que a quantidade de informação gerada pela sociedade tem aumentado em proporção geométrica, criando dificuldades para o seu gerenciamento e recuperação. Neste cenário, este artigo descreve uma proposta que permite a gestão, com maior eficiência, no processo de GED, utilizando dispositivos robóticos na etapa de digitalização, algoritmos de processamento de imagem para melhoria da qualidade da informação capturada, técnicas de organização da informação, armazenamento que considera a necessidade de interoperabilidade entre sistemas e de preservação digital, além de um sistema flexível e amigável para a disponibilização da informação armazenada via servidores Web. Este trabalho quanto aos objetivos tem o caráter qualitativo, e quanto aos meios se pauta em uma pesquisa prática ou experimental, utilizando-se como método o experimental e observacional no que diz respeito a construção e desenvolvimento de uma nova arquitetura para gerenciamento eletrônico de documentos mais eficiente. Com o desenvolvimento deste projeto espera-se ter uma arquitetura de baixo custo e alta eficiência que possa ser replicada entre membros redes memoriais, pra que a disponibilização da informação possa ser implementada de maneira satisfatória nas instituições envolvidas.
Palavras-chave: Gestão Eletrônica de Documentos. Robótica. Processamento de Imagem.

Abstract
The concept of electronic document management (EDM) has become increasingly important since the amount of generated information society has increased in geometric ratio, creating difficulties for its management and retrieval. In this scenario, this paper describes a proposal to permit management, with greater efficiency in the process of EDM, using robotic devices in step scanning, image processing algorithms to improve the quality of the captured information, technical information organization, storage considering the need for interoperability between systems and digital preservation, as well as a flexible and friendly to the availability of the stored information via web servers on the objectives of this work is qualitative, and the means is guided in a practical research or experimental, using as the experimental and observational method regarding the construction and development of a new architecture for electronic document management more efficient. With the development of this project is expected to have an architecture of low cost and high efficiency that can be replicated between members of networks memorials, to the provision of information that can be implemented satisfactorily in the institutions involved.
Keywords: Eletronic Document Management. Robotics. Image Processing.

1 INTRODUÇÃO

As políticas de acesso e disponibilização de informações são recentes no contexto mundial, especialmente no que se refere a instituições memoriais ou instituições de memória, nas quais as atividades de conservar, guardar e custodiar constituíam os principais, senão os únicos, exercícios, imperando, assim, um paradigma predominantemente custodialista. A partir da segunda metade do século XX, com o avanço das tecnologias, a explosão informacional e as novas demandas sociais, esse paradigma começa a apresentar os primeiros sinais de crise. Emerge, então, um novo paradigma, denominado "pós-custodial", a partir do qual as instituições memoriais, para além da guarda de documentos com o intuito de preservar a memória da sociedade, precisam promover o acesso às informações sob sua guarda, passando a necessitar de novas interfaces com o usuário, através dos novos meios de comunicação. Neste contexto, surgem novos desafios para os profissionais da informação, como, por exemplo, a "prioridade máxima dada ao acesso à informação, por todos [e] em condições bem definidas e transparentes" (MALHEIRO; RIBEIRO, 2011, p. 59), já que a custódia e a preservação somente se justificam para a promoção do acesso público.

Porém, neste novo cenário, os métodos tradicionais para o armazenamento de documentos, tais como: textos, fotografias, mapas, plantas, imagens entre outros, requerem grande quantidade de esforço para gerenciá-los. E, conforme aumenta o número de documentos, o tempo e o esforço despendidos para gerenciá-los, preservá-los e prover fácil acesso aos mesmos também aumenta. Esse é um problema que vem crescendo não apenas em instituições memoriais, mas também em organizações e empresas públicas e privadas.

Neste cenário, o conceito de gerenciamento eletrônico de documentos (GED) ganha importância na medida em que pode reduzir o tempo gasto em atividades diárias de produção, localização e recuperação documental; assegurar as informações registradas preservando os documentos; dinamizar e democratizar o acesso e racionalizar a ocupação espacial de grandes massas documentais suportadas em papel. Desta forma o acesso à informação, que muitas vezes se torna restrito, devido a fatores ligados à preservação do bem ou a dificuldade de acesso a ele, pode ser democratizado.

Neste contexto, este artigo apresenta uma arquitetura para GED baseada em automação para uma maior eficiência no processo de digitalização, a partir do uso de dispositivos robóticos. Essa arquitetura faz uso de algoritmos de processamento de imagem para melhoria da qualidade da informação digitalizada; utiliza técnicas de organização da informação reconhecidas; armazena a informação em um repositório digital, que leva em conta a necessidade de interoperabilidade entre repositórios e a necessidade de preservação

digital, além de possuir um sistema flexível e amigável para a disponibilização da informação armazenada para os usuários via Web. Esta arquitetura está em fase de implementação e já começa a apresentar resultados promissores, podendo ser uma solução eficaz e de baixo custo para instituições memoriais, uma vez que faz uso de software livre e dispositivos que podem ser compartilhados entre instituições que façam parte de uma rede.

2 GERENCIAMENTO ELETRÔNICO DE DOCUMENTOS (GED)

Para Koch (1998), o GED é a somatória de todas as tecnologias que visam gerenciar informações de forma eletrônica, desde a sua criação até o seu arquivamento, reforçando que não é necessário que os documentos estejam em meio eletrônico, mas sim, que o tratamento dispensado a estes sejam concretizados com o uso destas tecnologias. Assim, as informações podem estar, originalmente, em mídias analógicas ou digitais e em qualquer fase do seu ciclo de vida. A possibilidade de recuperação rápida de várias categorias de documentos e por várias pessoas ao mesmo tempo é um dos grande atrativos da GED. Além da economia do tempo gasto com arquivos mal armazenados e/ou mal classificados e a maior segurança das informações armazenadas. O GED

> evita a duplicação de documentos; permite classificar segundo diversos critérios cruzados; autoriza o acesso a informações e conhecimentos pertinentes; conter dados não vinculados por papel, como vídeo-som; acabar com o problema de tempo e lugar; implementa novos modos de navegação não-linear; permite e melhora a segurança e a perenidade dos arquivos. (DUARTE, 2006, p.137).

Outra vantagem, especificamente para instituições memoriais tais como bibliotecas, arquivos e museus é evitar o manuseio de documentos raros e históricos cuja integridade possa ser comprometida pelo manuseio por parte dos usuários ou pela exposição à luz ou a condições ambientais diversas.

Na literatura há várias propostas de sistemas de GED (BAX e BAX, 2006; AVEDON, 2002). A maioria das soluções tem em comum a digitalização com o uso de scanners convencionais, o uso raro de algoritmos para melhoria da qualidade das imagens digitalizadas e a falta de preocupação com a interoperabilidade com outros sistemas. Neste sentido, é proposta neste artigo, uma nova arquitetura para GED, descrita no decorrer do mesmo.

3 METODOLOGIA

Este trabalho quanto aos objetivos tem o caráter qualitativo, e quanto aos meios se pauta em uma pesquisa prática ou experimental, utilizando-se como método o experimental e observacional no que diz respeito a construção e desenvolvimento de uma nova arquitetura para gerenciamento eletrônico de documentos mais eficiente. Segundo Michel (2009) "este tipo de pesquisa se fundamenta na discussão da ligação e correlação de dados interpessoais, na coparticipação das situações dos informantes, analisados a partir da significação que estes dão aos seus atos". Neste sentido, o pesquisador deve participar, compreender e interpretar o seu objeto.

A pesquisa experimental ou prática na área das Ciências Sociais pode ser aplicada com a simulação de condições laboratoriais, simulando ambientes específicos, pode-se reproduzir e verificar na prática problemas e como estes se comportam aplicando variáveis específicas (MICHEL, 2009). O método experimental e observacional ajudará no teste do objeto e comprovação de sua validade, assim como na possibilidade de captar dados da realidade da investigação respectivamente, contribuindo para apresentar resultados que sejam compartilhados entre instituições que façam parte de uma rede. Por conseguinte, esta experimentação terá como foco a aplicação de testes em cada um dos módulos/sistemas da arquitetura proposta (módulos de digitalização, processamento de imagem, sistema de apoio à organização da informação e sistema de armazenamento e disponibilização das informações), após desenvolvidos. Também serão realizados testes gerais após a integração dos módulos propostos.

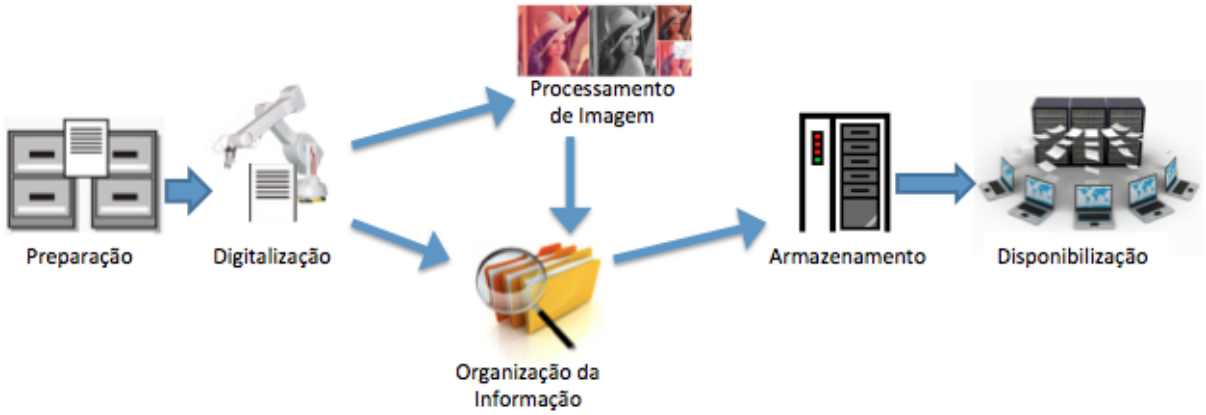
A arquitetura para GED está subdivida em módulos e sistemas independentes, de forma que o desenvolvimento deles possa ser realizado em paralelo. As equipes de desenvolvimento são multidisciplinares e contam com professores e estudantes das universidades federais de Pernambuco e da Paraíba. A comunicação entre a equipe é realizada através de ferramentas colaborativas e reuniões periódicas são realizadas para avaliação dos resultados sendo obtidos e para realizar ajustes no planejamento, quando necessário.

4 ARQUITETURA PARA GERENCIAMENTO ELETRÔNICO DE DOCUMENTOS

A arquitetura de GED proposta neste artigo possui as seguintes etapas, conforme mostra a Figura 1: Preparação, Digitalização, Processamento de Imagem, Organização da Informação, Armazenamento e Disponibilização. Vale ressaltar que, quando o documento já se encontra digitalizado, ele pode entrar nesta arquitetura já na etapa de Processamento de

Imagem ou, se a imagem já se encontra como se deseja, diretamente na etapa de Organização da Informação. Cada uma das etapas desta arquitetura serão descritas nas subseções a seguir.

Figura 1 - Etapas da arquitetura de GED



Fonte: Os autores, 2013

4.1 PREPARAÇÃO

A etapa de Preparação dos documentos consiste em tornar o documento apto à captura do seu conteúdo. Por exemplo, se o processo considera um scanner como dispositivo de captura, é necessário retirar objetos (grampos, clips etc.) que possam impedir o fluxo do documento através do scanner. Outras ações que devem ser realizadas nesta etapa são: desfazer dobras, restaurar documentos danificados, identificar os documentos com legibilidade comprometida, checar direitos autorais dos documentos/obras. Esta é uma etapa **manual** e que, dependendo do valor do documento ou do estado de conservação do mesmo, deverá ser realizada por profissional especializado.

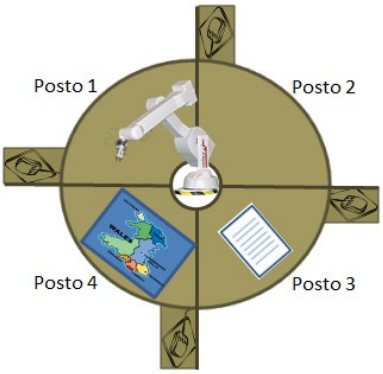
4.2 DIGITALIZAÇÃO

Em sua grande maioria, os processos de digitalização são baseados em equipamentos do tipo scanner. Porém, este processo de digitalização é custoso e não adequado para tratar o grande volume de dados das organizações. Algumas dificuldades que podem ser citadas são: a velocidade de captura do scanner; o alto custo dos scanners profissionais de alta resolução (que em sua maioria precisam ser importados) ou para grandes documentos (ex: mapas) e a adequação do tamanho do scanner ao documento que vai ser digitalizado. Além disso, os scanners mais eficientes, os quais possuem alimentação automática do material a ser digitalizado e função duplex, são direcionados à digitalização de folhas soltas, de modo que não são adequados para digitalização de livros e mapas, por exemplo.

Na arquitetura proposta a etapa de digitalização está baseada em um braço robótico, o qual controla o movimento de uma câmera de alta resolução utilizada para realizar a captura das imagens. O braço mecânico utilizado neste projeto é o modelo R17 Deucalion (Figura 2).

Figura 2 - (a) Braço mecânico modelo R17 Deucalion (b) Esquema do ambiente de digitalização

(a)



(b)

Fonte: Os autores, 2013.

Este equipamento é composto por três unidades principais: o robô, o controlador e o terminal de controle (um computador ou laptop). Uma vez programado, o controlador pode gerenciar o robô, independentemente do terminal de controle, passando a executar ações de forma autônoma. Neste modelo, o controlador é responsável por gerenciar todos os movimentos do robô, recebendo sinais de sensores instalados no próprio braço, ou dos equipamentos associados ao controlador. A função do terminal de controle é (a) programar o controlador, (b) copiar arquivos da RAM (Random Access Memory) do controlador para um dispositivo de memória permanente, e (c) realizar um processo de monitoramento via envios de comando para o robô. O terminal também apresenta informações para que o operador do robô controle o processo, quando tal não for totalmente autônomo.

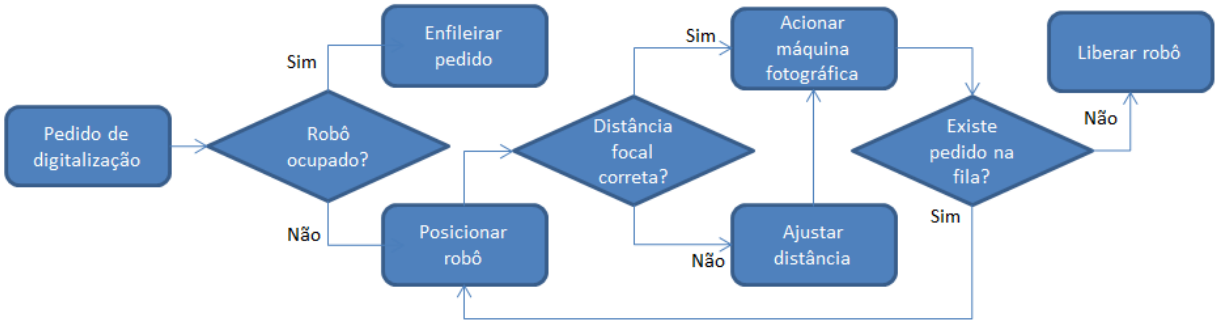
A Figura 2b ilustra como seria um possível ambiente de digitalização. A ideia é utilizar uma mesa circular, de tal forma que até quatro postos de trabalho possam ser configurados em paralelo. O braço mecânico ficaria no centro da mesa e os operadores, após prepararem o documento a ser digitalizado, acionariam um pedido para serem atendidos pelo braço robótico. Considerando tal ambiente, algumas questões estão sendo consideradas pelo grupo de pesquisa, conforme descrito a seguir.

Como qualquer outra plataforma de robótica comercial, a linguagem de programação é legada ou proprietária. Ou seja, obedece a uma sintaxe e semântica própria do seu fornecedor, não existindo uma interface de programação (API) que permita a sua fácil integração com

outros dispositivos externos ao sistema R17. A partir de uma análise inicial da plataforma R17, foi verificada a possibilidade de se trabalhar com linguagem binária, de modo que o desenvolvimento de uma interface de comunicação deverá ser a solução para abstrair os comandos de baixo nível do braço robótico. Este trabalho será realizado por especialistas da área de robótica e informática. Dessa forma, o operador do braço responsável pela captura das imagens não precisará ter conhecimento técnico de robótica ou informática, pois os comandos estarão configurados para serem de fácil compreensão.

Deverá haver um sincronismo entre os diversos componentes do sistema. Por exemplo, na configuração da Figura 2b, existem diversas entradas de pedidos de digitalização, o movimento do robô em si e o acionamento da câmera. Estes componentes deverão conversar entre si, de modo a manter o sincronismo. Além disso, como os documentos não são de um tamanho padrão único, o ajuste automático da distância focal ótima é um requisito que pode aumentar em muito a eficiência do sistema, principalmente em termos de velocidade. Dessa forma, o fluxograma das ações e decisões que o sistema deverá realizar foram definidos na Figura 3.

Figura 3 - Fluxograma das ações e decisões realizadas pelo sistema

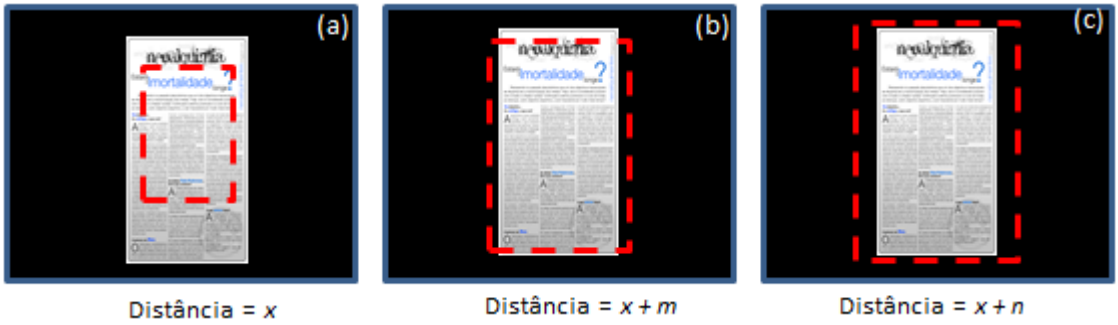


Fonte: Os autores, 2013.

De acordo com o fluxograma, quando um pedido de digitalização de um dos postos chega ao sistema, este verifica se o robô está ocupado. Caso ele esteja ocupado, o pedido é registrado em uma estrutura de dados do tipo fila (primeiro que entra é o primeiro que sai). Caso o robô não esteja ocupado, ele é posicionado em um ponto central do posto de trabalho. Neste momento o robô verifica se a distância para o documento está correta. Distância correta, neste contexto, é uma distância na qual a câmera consiga captar todo o documento com o mínimo possível de elementos externos ao mesmo (background). Esta distância permite que o ângulo de visão seja suficiente para que o campo de visão da máquina fotográfica contemple todo o documento a ser digitalizado. Este não é um problema trivial dado a variedade de tamanhos dos documentos. Uma solução que será testada é a utilização

de um fundo negro, de modo que o ajuste se dê quando uma borda negra seja detectada em todas as margens (Figura 4).

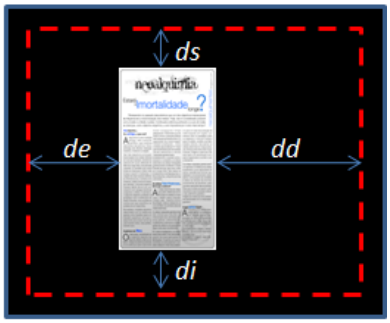
Figura 4 - Exemplos de campo de visão para diferentes distâncias focais



Fonte: Os autores, 2013.

Neste exemplo (Figura 4a) a distância da câmera ao documento (distância x) ainda é pequena, de modo que o braço robótico se movimenta m unidades para cima. A esta distância (x+m) as bordas negras laterais já são identificadas, mas não a superior e inferior. Deste modo, o braço modifica sua altura para x + n, onde n > m, de modo que todas as bordas negras são identificadas. Pode-se notar que nesta posição, a câmera vai captar imagens adicionais (bordas lalterais negras). Porém, tais partes podem ser facilmente retiradas na etapa de processamento posteriormente definida. Existem algumas limitações nesta abordagem. Primeiro, ela considera que o centro do documento é sempre posicionado no centro focal inicial da câmera, o qual pode, por exemplo, ser marcado na mesa. Segundo, o braço deveria fazer diversos ajustes na distância até encontrar a distância ótima. Em cada um desses ajustes, a câmera deveria capturar uma imagem, a qual seria processada, gerando informação para que um novo ajuste seja feito. Uma aboragem mais interessante é iniciar o ajuste na altura máxima. Neste ponto, poderíamos utilizar as medidas de, dd, ds e di (Figura 5) para calcular a 3-tupla <h,x,y>, onde h é a altura ideal da câmera e x,y o ponto onde a câmera deve ser posicionada, considerando o plano cartesiano utilizado pelo braço robôtico. Deste modo, no máximo, um ajuste seria feito. O algoritmo que faz o ajuste do braço robótico de acordo com estes parâmetros de entrada (de, dd, ds e di) será um dos produtos deste projeto.

Figura 5 - Parâmetros de entrada para o cálculo do ajuste do braço robótico.



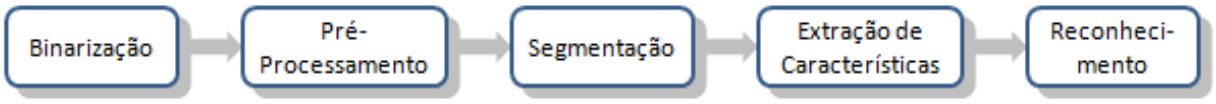
Fonte: Os autores, 2013.

4.3 PROCESSAMENTO DE IMAGENS DE DOCUMENTOS

Um sistema de processamento automático de imagens de documentos contém técnicas específicas para tratamento de imagens, ou seja, para melhorar a qualidade da imagem capturada. Em geral, podemos ter como objetivos para tais sistemas: melhoria da qualidade da imagem (ex: retirada de manchas no papel, resolução de sobreposição de textos do verso e da frente do papel, resolução de diferenças de iluminação), conversão da imagem para outro padrão de cores ou conversão da imagem do documento para texto editável.

Cada um desses possíveis objetivos pode ter diferentes aplicações como: Recuperação de Imagens de Documentos (DIR, do inglês, *Document Image Retrieval*) na qual uma imagem de referência pode ser usada na busca por outras imagens similares em uma base de dados (BOKCHOLT; CAVALCANTI; MELLO, 2011); Restauração de Imagens Antigas (como fotografias ou cartões postais) (ROE; MELLO, 2012); Reconhecimento Óptico de Caracteres (OCR) que permite a conversão da imagem do documento em texto editável, facilitando, por exemplo, a busca por palavras-chaves) (MELLO; OLIVEIRA; SANTOS, 2012). Para fins de OCR, após a digitalização da imagem (na fase de aquisição), os passos são apresentados na Figura 6.

Figura 6 - Etapas para um sistema de processamento automático de imagens visando reconhecimento óptico de caracteres



Fonte: Os autores, 2013.

Primeiro, é feita a conversão da imagem colorida para preto-e-branco, através de um processo chamado de binarização ou limiarização (GONZALEZ; WOODS, 2007,. SEZGIN; SANKUR, 2004). Na prática, há um ponto de corte que separa os tons claros, considerando que eles fazem parte do papel e os tons escuros como sendo parte da tinta. O problema principal da binarização é encontrar qual o ponto de corte ideal. Isso é particularmente um

problema em imagens de documentos antigos onde tanto o papel quanto a tinta podem sofrer degradação. A degradação do papel faz com que este assuma tons amarelados ou que surjam manchas que deixam regiões mais escuras. Nos dois casos, a presença da tinta pode gerar confusão na binarização. Caso a tinta não sofra degradação, ela pode ser confundida se aparecer em uma região escurecida do papel. Se a tinta sofrer degradação e se esvair, ela pode tornar-se tão ou até mais clara que o papel. Esse é um dos principais problemas a ser tratado no processo de binarização. Para a maioria das imagens, algoritmos simples como os apresentados em Sezgin e Sankur (2004) trazem bons resultados. No caso de documentos com manchas, degradações ou diferenças de iluminação um algoritmo eficiente e eficaz considera elementos de percepção de objetos à distância para conseguir separar os tons do papel dos tons da tinta e conseguir binarizar a imagem (MESQUITA et al, 2013). A conversão para preto-e-branco é necessária para os processos seguintes dos sistemas de processamento de imagens de documentos.

Com a imagem preto-e-branco, aplicam-se algoritmos para remoção de ruído (muitas vezes provenientes do próprio papel) que, ao se tornarem preto, podem ser confundidos com a tinta, dificultando a leitura da imagem dos documentos. Dependendo da quantidade de ruído da imagem, o processo de remoção pode ser extremamente simples, considerando apenas uma análise dos vizinhos de cada pixel (um pixel preto com oito vizinhos brancos, provavelmente, é ruído e deve ser convertido para branco, assim como um pixel branco cercado de vizinhos pretos deve ser convertido para preto). Nessa mesma fase de correção da imagem binarizada, aplica-se também métodos de estimativa e correção de inclinação do documento. Nesse caso, podemos ter uma inclinação única para toda a imagem (indicando um erro no processo de aquisição da imagem) ou inclinações diferentes em cada linha de texto (que ocorre, em geral, quando temos um documento manuscrito em um papel sem pauta). Após a estimativa de inclinação do documento (seja ela global ou local), procede-se com a rotação do documento (ou da linha de texto). Para estimativa de inclinação, nosso sistema usa o método desenvolvido em Mello, Sánchez e Cavalcanti (2011a) e para correção da inclinação detectada usa-se o método apresentado em Ávila, Lins e Oliveira (2005). Todos esses processos fazem parte do chamado pré-processamento da imagem do documento.

Em seguida, a imagem deve ser segmentada. Segmentação pode ser entendida como a divisão da imagem em seus objetos. Como o conceito de objeto pode variar para cada fase do processo, a segmentação também muda de características. No primeiro momento, temos a segmentação do documento que identifica quais as regiões onde temos elementos gráficos no documento (imagens ou figuras) e quais as regiões de texto (SHEN; LI; KWOK, 2005). As

regiões de texto são novamente segmentadas para identificar as linhas de texto, palavras e, se necessário, os caracteres (isso não é feito em documentos escritos com letras cursivas já que as letras estão conectadas, formando palavras e as conexões não são facilmente identificadas).

Tanto para segmentação de linhas quanto de palavras, usamos o método descrito em Sánchez et al (2011b). Em algumas aplicações, a segmentação de caracteres pode ser necessária e pode ser aplicada, principalmente, quando o domínio é mais restrito. A segmentação de dígitos pode servir para identificar corretamente datas em imagens de documentos. Caracteres podem aparecer conectados em um ou mais pontos ou sobrepostos (quando há interseção de um sobre o outro). No primeiro caso, sugere-se o uso do algoritmo apresentado em Lacerda e Mello (2013) e, no segundo caso, é apropriado o uso do algoritmo descrito em Roe e Mello (2009). No caso do caractere se apresentar degradado, pode ser necessário recuperar esse caractere, restaurando os pontos de degradação conforme apresentado em Lopes Filho e Mello (2013).

Feitas as segmentações, para um sistema de reconhecimento de caracteres, faz-se a representação dos caracteres em características que consigam descrever corretamente um caractere, dificultando a confusão de um com outro. Essa é uma das etapas mais complexas do reconhecimento já que temos uma grande quantidade de possíveis fontes de caracteres. Diferentes formas de representar caracteres por características podem ser encontradas em Mello, Oliveira e Santos (2012). A partir dessas características, o reconhecimento é feito através de técnicas de aprendizagem de máquina como Redes Neurais ou Máquinas de Vetor de Suporte (HAYKIN, 1988; NEVES et al, 2011). A partir daí, temos um documento em formato de texto sendo possível ser editado.

4.4 ORGANIZAÇÃO DA INFORMAÇÃO

Resumidamente, pode-se afirmar que o objetivo da organização da informação é dar suporte ao fluxo de tratamento e recuperação dos objetos informacionais estruturados, semiestruturados e não-estruturados nas organizações. Em outra palavras, a organização da Informação é um processo, atividade, técnica, operação, que remonta os primórdios da antiguidade e tem o objetivo principal de subsidiar a recuperação da informação, a partir da descrição das características físicas e do tratamento temático do conteúdo dos objetos informacionais. Considera-se Objeto Informacional, qualquer informação disponibilizada nos seus mais variados suportes. (BRASCHER; CAFÉ, 2008; FUJITA, 2003). A descrição física diz respeito à representação descritiva ou, como é mais conhecida catalogação. A descrição de conteúdo diz respeito aos processos de representação temática ou, como é mais conhecida à

classificação e indexação. Dessa forma, as atividades desenvolvidas no módulo de organização da informação serão:

Representação descritiva - que pode ser definida como uma operação de reunir os elementos que possibilitem a identificação de um documento em uma coleção. Pode-se dizer que ela contempla os dados ligados à produção editorial dos documentos, tais como o responsável pela obra, título da publicação, editor, ano de publicação, número de páginas. Para este fim, será utilizado o padrão de descrição de informação (padrão de metadados) Dublin Core (ROSETTO; NOGUEIRA, 2002), por ser um padrão internacional e que tem a flexibilidade necessária para descrever qualquer tipo de objeto informacional. A definição dos metadados, seguindo um padrão é importante para a identificação, a organização e a recuperação da informação digital. (SCHAEFER, 1998).

Representação temática – que é a ação de descrever e identificar um documento de acordo com o seu assunto. É a classificação propriamente dita de cada objeto informacional a ser armazenado. Além da representação temática da informação, existe a indexação, que é uma prática que visa extrair termos de determinado documento com o fim de promover e facilitar o acesso do usuário ao conteúdo informacional de que necessita.

Indexação – que tem a função de identificar a pasta e os arquivos capturados, com atributos de pesquisa ou campos de indexação que permitem o acesso posterior ao documento, ou partes do documento, digitalizado. Esse acesso pode ser realizado com recurso de reconhecimento automático de caracteres (OCR) para documentos digitados, ou (ICR) no caso de documentos escritos à mão;

A etapa de organização da informação necessitará de intervenção manual de especialistas da área de Ciência da Informação, que terão o apoio ferramental necessário desenvolvido de forma a facilitar a realização do seu trabalho.

4.5 ARMAZENAMENTO E DISSEMINAÇÃO

Todo documento capturado, processado e organizado necessitará ser armazenados visando sua preservação de longo prazo. Dessa forma, o sistema de armazenamento considera as alterações na tecnologia e o aumento do número e volume dos documentos, além de ações de segurança como a realização de backups periódicos e espelhamento do serviço de disponibilização (BORBA et al, 2012). Também, a fim de garantir a capacidade de leitura a longo prazo, os arquivos serão todos armazenados em formato não-proprietário e que sejam reconhecidos como formatos para preservação, tais como PDF (para textos) e TIFF (para imagens), sendo geradas imagens para disponibilização, quando necessário, em formato

otimizado tal como o JPG. Além de serem documentados também em metadados de preservação (DUBLIN..., 2013). A etapa de Armazenamento também implementa requisitos configuráveis de segurança de acesso a informação, para garantir que a informação esteja disponível apenas para quem de direito.

A etapa de disponibilização tem a função de permitir o acesso ao conteúdo armazenado via Web, quebrando, deste modo, a barreira espaço-tempo. Dessa forma, para armazenamento e disponibilização será utilizado um repositório digital. Segundo Viana, Márdero Arellano e Shintaku (2005), um repositório digital é uma forma de armazenamento de objetos digitais que tem a capacidade de manter e gerenciar material por longos períodos de tempo e prover o acesso apropriado. O uso de repositórios digitais trás como vantagem a disponibilização da informação sem limitações de espaço e tempo e a facilidade na recuperação da informação); o não desgaste do material original pela manipulação por parte de diversos usuários; a possibilidade do mesmo material ser acessado por vários usuários ao mesmo tempo.

O repositório digital escolhido para essa finalidade é o CLIO (SIEBRA et al., 2011; CARDOSO et al, 2011). O Clio é um repositório digital que possui características diferenciadas, tais como: ferramentas de busca simples e avançada (agilizando a recuperação dos documentos armazenados); um módulo de visualização das informações armazenadas que permite operações sobre os documentos eletrônicos (ex: zoom, negativação, corte, entre outros); uso do protocolo OAI-PMH para prover a Interoperabilidade com outros repositórios digitais; uso de metadados no padrão internacional Dublin Core (DUBLIN, 2013), versão 2010; funções que promovem acessibilidade; desenvolvimento com foco na facilidade de uso por parte do usuário.

5 CONSIDERAÇÕES FINAIS

Este projeto está sendo realizado no laboratório LIBER (Laboratório de Tecnologia do Conhecimento) por uma equipe multidisciplinar que envolve o Departamento de Ciência da Informação e o Centro de Informática da Universidade Federal de Pernambuco e o Departamento de Ciência da Computação da Universidade Federal da Paraíba, agregando o conhecimento de vários grupos de pesquisa.

O projeto encontra-se em fase de desenvolvimento. Porém, todas as etapas estão sendo desenvolvidas em paralelo a fim de otimizar o tempo de finalização do projeto. Espera-se com a realização desse projeto conseguir patentear uma solução replicável e de baixo custo

para uso por instituições memoriais, a fim de aumentar a eficiência do trabalho de gestão eletrônica de documentos.

Nessa solução, a digitalização será feita com auxílio do braço robótico o qual se ajustará automaticamente de acordo com o tamanho do papel a ser digitalizado. A imagem do documento será processada, visando a remoção de ruídos, correção de problemas como rotação, segmentação e, por fim, reconhecimento de caracteres. Além disso, diversas formas de visualizar a imagem e a informação que ela contém estarão disponíveis ao usuário, quando a informação for disponibilizada.

Este projeto apresenta-se como uma grande oportunidade de cooperação entre especialistas e instituições e para a realização de parcerias entre a universidade e organizações públicas e privadas.

REFERÊNCIAS

AVEDON, D. M. **GED de A a Z:** tudo sobre gerenciamento eletrônico de documentos . São Paulo: CENADEM, 2002. 200 p.

ÁVILA, B.T.; LINS, R. D.; OLIVEIRA, L. A New Rotation Algorithm for Monochromatic Images. **DocEng**, Bristol, 2005.

BALDAM, R. L.; VALLE, R. CAVALCANTI, M. **GED:** gerenciamento eletrônico de documentos . 2. ed. rev. e atual. São Paulo: Érica, 2004. 204 p.

BAX, M.P., BAX, M.L.P. Gestão da Documentação por Imagens: um Tipo Específico de GED.

Revista Perspectivas em Ciência da Informação, Belo Horizonte, Escola de Ciência da Informação da UFMG. v. 7, n.02. p.141-154. dez. 2002.

BORBA, V. da R. et al. Política de Preservação Digital: Diretrizes para o LIBER. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 13., 2012, Rio de Janeiro. **Anais**... Rio de Janeiro: FIOCRUZ, 2012.

BOKCHOLT, T. C.; CAVALCANTI, G. D. C.; MELLO, C. A. B. Document image retrieval with morphology-based segmentation and features combination. **SPIE XVIII Document Recognition and Retrieval (DRR)**, San Francisco, 2011.

BRASCHER, Marisa; CAFÉ, Lígia. Organização da Informação ou Organização do Conhecimento? In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 9., 2008, São Paulo. **Anais**... São Paulo: USP, 2008.

CARDOSO JR. et al. CLIO-I: primando pela usabilidade e acessibilidade em um sistema para gerenciamento e interoperabilidade de repositórios digitais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 12., 2011, Brasília. **Anais**... Brasília: IBICT, 2011.

DUARTE, E. N.; SILVA, A. K. A.; SANTOS, E. G.; LIMA, I. F.; RODRIGUES, M. P. F.; COSTA, S. Q. Vantagens do Uso de Tecnologias para Criação, Armazenamento e Disseminação do Conhecimento em Bibliotecas Universitárias. **Transinformação**. v. 18, n. 2, mai-ago, 2006. p 131-141. Disponível em: <http://200.18.252.94/seer/index.php/transinfo/article/view/675> . Acesso em: 25 de jul. 2013.

DUBLIN Core Metadata Initiative. Disponível em: <http://dublincore.org >. Acesso em: 13 mai. 2013.

FUJITA, M. S. L. **A leitura documentária do indexador**: aspectos cognitivos e lingüísticos influentes na formação do leitor profissional. 2003, 321f. Tese (Livre-Docência em Análise Documentária e Linguagens Documentárias Alfabéticas) – Faculdade de Filosofia e Ciências da UNESP.

GONZALEZ, R.; WOODS, R. **Digital Image Processing**. 3. ed. New Jersey: Prentice Hall, 2007.

HAYKIN, S. **Neural Networks**: A Comprehensive Foundation. 2. ed. New Jersey: Prentice Hall, 1988.

KOCH, W. **Gerenciamento eletrônico de documentos – GED.** São Paulo: Cenadem, 1998.

LACERDA, E. B.; MELLO, C. A. B. Segmentation of Connected Handwritten Digits Using Self-Organizing Maps. **Expert Systems with Applications**, v.40, n.15, p.5867-5877, 2013.

LOPES FILHO, A.N.G; MELLO, C. A. B. Degraded Digit Restoration Based on Physical Forces. **International Conference on Document Analysis and Recognition (ICDAR),** Washington, EUA, 2013.

MELLO, C. A. B.; SÁNCHEZ, A; CAVALCANTI, G. D. C. Multiple Line Skew Estimation of Handwritten Images of Documents Based on a Visual Perception Approach. **Lecture Notes in Computer Science**, v.6855, p.138-145, 2011a.

MELLO, C. A. B.; OLIVEIRA, A.L.I.; SANTOS, W. P. dos. (Ed.) **Digital Document Analysis and Processing**. New York: Nova Science Publishers, 2012.

MESQUITA, R.G. *et al.* A new thresholding algorithm for document images based on the perception of objects by distance. **Integrated Computer-Aided Engineering**, 2013.

NEVES, R.F.P. et *al.* A SVM Based Off-Line Handwritten Digit Recognizer. **IEEE International Conference on Systems, Man, and Cybernetics, Anchorage**. Alaska, p. 510-515, 2011.

ROE, E.; MELLO, C. A. B. Automatic system for restoring old color postcards. **IEEE International Conference on Systems**, Man and Cybernetics, Seul, Coreia do Sul, p.451-456, 2012.

ROE, E.; MELLO, C. A. B. Simulating Inertial and Centripetal Forces for Segmentation of Overlapped Handwritten Digits. **IEEE Systems, Man and Cybernetics (SMC)**, San Antonio, EUA, p. 149-153, 2009.

ROSETTO, M.; NOGUEIRA, A. Aplicação de elementos metadados Dublin Core para descrição de dados bibliográficos on-line da biblioteca digital de teses da USP. In: SIMPÓSIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 12. São Paulo. **Anais**.... 2002. Disponível em: < http://www.sibi.ufrj.br/snbu/snbu2002/oralpdf/82.a.pdf>. Acesso em: 10 Mar. 2013.

SÁNCHEZ, A. *et al.* Automatic line and word segmentation applied to densely line-skewed historical handwritten document images. **Integrated Computer-Aided Engineering**, v.18, p.125–142, 2011b.

SEZGIN, M.; SANKUR, B. Survey over image thresholding techniques and quantitative performance evaluation. **Journal of Electronic Imaging**, v.13, n.1, p.146-168, 2004

SHEN, Q.; LI, S.; KWOK, J. Page Segmentation Using Mathematical Morphology. **International Symposium on Intelligent Signal Processing and Communication Systems**, p.89-92, Japão, 2005.

SIEBRA, S. A.; Cardoso Jr, M.; Borba, V.; Miranda, M. K. F. O. Um Sistema para Gerenciamento e Interoperabilidade de Repositórios Digitais com foco na Simplicidade, Usabilidade e Acessibilidade. In: **Conference on Technology, Culture and Memory - CTCM, 2011**, Recife – PE. Disponível em: < http://www.liber.ufpe.br/ctcm/anais/anais_ctcm/14_sistema_clio.pdf>. Acesso em: 13 Mar. 2013.