UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

COORDENAÇÃO DE ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

RENATA ANDJARA WISNIEWSKI

DESENVOLVIMENTO DE UM *WORKFLOW* PARA PROCESSAMENTO E A PRODUÇÃO DIGITAL DE DOCUMENTOS

TRABALHO DE DIPLOMAÇÃO

PONTA GROSSA

2012

RENATA ANDJARA WISNIEWSKI

DESENVOLVIMENTO DE UM *WORKFLOW* PARA PROCESSAMENTO E A PRODUÇÃO DIGITAL DE DOCUMENTOS

Trabalho de conclusão de curso de graduação, apresentado à disciplina Trabalho de Diplomação, do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas da Coordenação de Análise e Desenvolvimento de Sistemas – COADS – da Universidade Tecnológica Federal do Paraná – UTFPR - como requisito parcial para a obtenção do título de Tecnólogo.

Orientador: Profº Rogério Ranthum

Co-Orientador: Profº Edson Armando Silva

PONTA GROSSA

2012

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS PONTA GROSSA
GERENCIA DE ENSINO E PESQUISA
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS
DISCIPLINA DE TRABALHO DE DIPLOMAÇÃO

TERMO DE APROVAÇÃO

DESENVOLVIMETNO DE UM *WORKFLOW* PARA PROCESSAMENTO E A PRODUÇÃO DIGITAL DE DOCUMENTOS

Renata Andjara Wisniewski

Este Trabalho de Diplomação foi considerado adequado como cumprimento das exigências legais do currículo do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas e aprovado em sua forma final pela Coordenação de Informática da Universidade Tecnológica Federal do Paraná – Campus de Ponta Grossa.

_____
Profº ROGÉRIO RANTHUM
Orientador

_____
Profª HELYANE BRONOSKI BORGES
Responsável pelo Trabalho de Diplomação

_____
Profª SIMONE ALMEIDA
Coordenadora do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas

Banca Examinadora:

_____
Profº Danillo Leal Belmonte

_____
Profº Geraldo Ranthum

* Termo de Aprovação disponível na coordenação do curso

Dedico este trabalho à minha família,
a base da minha vida.

AGRADECIMENTOS

Agradeço antes de tudo a Deus, que com certeza guiou meu caminho e me conduziu até aqui.

Agradeço a minha família, que mesmo estando longe, sempre me apoiaram e me incentivaram em todos os momentos da minha vida. Uma graduação não é resultado somente desses últimos anos e sim de uma vida toda e vocês são a minha vida. Mãe, pai, vó, vô e João Ricardo, amo muito vocês.

Quero agradecer em especial à minha mãezinha Rosane A. Boesing, por seu amor, seu carinho, por dedicar sua vida para me fazer feliz. Mãe, você estava comigo desde a primeira vez que entrei nesta Universidade, acompanhou toda minha trajetória, nos momentos alegres e também nos difíceis, sempre me apoiando, me aconselhando e me incentivando a nunca desistir, tenho certeza que se não fosse você hoje eu não estaria aqui. Minha maior conquista é poder te encher de orgulho! Mãezinha, muito obrigada por cada palavra, abraço, risada, esta conquista também é sua! Queria te dizer que te amo muito e que você é com certeza a melhor mãe do mundo.

Agradeço ao meu amor Thiago L. Gasparetto, foi durante o curso que eu te conheci, e esse foi o maior presente que eu poderia ganhar. Obrigada por todos os momentos ao seu lado, obrigada pelo seu apoio, incentivo e pela sua paciência. Você sabe da grande jornada que eu percorri para chegar até aqui, e, com seu jeito, com seu sorriso, com suas palavras, tornou tudo mais fácil e feliz na minha vida, você também faz parte desta conquista. Obrigada pelo seu amor e por me fazer tão feliz. Te Amo!

Agradeço ao Professor Edson Armando Silva e ao meu orientador Professor Rogério Ranthum, pela oportunidade que me deram em participar da iniciação científica, por seus ensinamentos e por todo o auxílio prestado para a construção do trabalho.

Agradeço a todos que de alguma forma fizeram parte de forma positiva em minha vida, com certeza levo comigo um pouquinho de cada um de vocês.

Obrigada!

RESUMO

WISNIEWSKI, Renata A. **DESENVOLVIMENTO DE UM *WORKFLOW* PARA PROCESSAMENTO E A PRODUÇÃO DIGITAL DE DOCUMENTOS.** 2012. 83 f. Trabalho de Conclusão de Curso – Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2012

Este trabalho apresenta um estudo detalhado acerca do processo de digitalização de documentos. É relatada a importância e os benefícios advindos da adoção dessa tecnologia, discutindo-se os conceitos que envolvem o gerenciamento eletrônico de documentos e a tecnologia *workflow*. Através de um estudo de caso com base no projeto de iniciação científica, cada uma das etapas pertencentes ao processo de digitalização de documentos (captura, tratamento da imagem, reconhecimento de caracteres óticos, meta dados e armazenamento) são devidamente especificados, contendo a relação de suas atividades e os respectivos *softwares* livres que foram utilizados. Por fim, é estabelecida uma possível integração destas ferramentas através de um *workflow*. Como resultado, o trabalho fornecerá base para que instituições possam adotar esta tecnologia de baixo custo para a digitalização de seus acervos.

Palavras-Chave: Digitalização de Documentos. Gerenciamento Eletrônico de Documentos. Tratamento de Imagem. OCR. Objeto Digital. Repositório *Online*.

ABSTRACT

WISNIEWSKI, Renata A. **DEVELOPMENT OF WORKFLOW TO PROCESSING AND DIGITAL PRODUCTION OF DOCUMENTS**. 2012. 83 f. End of Course Work - Degree in Course Technology Analysis and Systems Development, Federal Technological University of Paraná. Ponta Grossa, 2012.

This work presents a detailed study about the process of scanning documents. It is reported the importance and benefits arising from the adoption of this technology, discussing the concepts involving electronic document management and *workflow* technology. Through a case study based on scientific initiation project, each of the steps belonging to the document scanning process (capture, image processing, optical character recognition, metadata and storage) are properly specified, containing a list of their activities and its free *software* that were used. Finally, we established a possible integration of these tools through a *workflow*. As a result, the work will provide basis for institutions to adopt this low-cost technology for the digitization of its collections.

Keywords: *Document Scanning. Electronic Document Management. Image Treatment.* OCR*. Digital Object. Online Repository.*

LISTA DE FIGURAS

LISTA DE SIGLAS

CENADEM	Centro Nacional de Desenvolvimento do Gerenciamento da Inform
CONARQ	Conselho Nacional de Arquivos
GB	*Gigabyte*
GED	Gestão De Documentos Eletrônicos
GIMP	GNU *Image Manipulation Program*
ICA-ATOM	Conselho Internacional de Arquivos - Acesso à Memória
OBE	*Open Business Engine*
OCR	Reconhecimento Ótico de Caracteres
RGB	*Red, Green e Blue*
SGBD	Sistemas de Gestão de Banco de Dados
TI	Tecnologia da Informação
WfMC	*Workflow Management Coalition*

SUMÁRIO

1 INTRODUÇÃO

É encontrado em acervos, memoriais, museus e bibliotecas um grande volume de obras, livros, coleções, jornais, documentos administrativos entre outros. Alguns destes documentos são muito antigos e ficam sob proteção e conservados de forma especial, o que acaba restringindo muitas vezes o acesso ao público. Outro problema encontrado é quanto à forma de conservação e vida útil do papel impresso, “há certa dificuldade em se conservar o papel impresso por períodos muito longos e – ainda que bem conservado – o papel impresso tende, nesses casos, a se tornar muito sensível ao manuseio”. (MARQUES et al., 2010).

A conversão de documentos impressos para a forma de mídia eletrônica está sendo cada vez mais utilizada, pois é oportuna e extremamente viável. Esta conversão engloba documentos de vários tipos e épocas e os disponibiliza de forma prática e rápida ao público. A informática está em grande desenvolvimento nesta área. É importante ressaltar que a digitalização de documentos não compreende somente a captura e o armazenamento dos mesmos, os objetos digitais devem ser devidamente preparados seguindo alguns procedimentos, como será visto detalhadamente no decorrer do trabalho e, para isso, a informática nos fornece diversos recursos.

A digitalização de documentos deve passar por algumas fases, como: captura da imagem, tratamento da imagem, reconhecimento óptico de caracteres (OCR), reconhecimento de caracteres, organização de meta-dados e meios de armazenamento. No mercado já existem *software*s que auxiliam na realização das tarefas pertinentes a cada fase, porém muitos têm um custo muito alto e outros são de difícil acesso. Como cita Silva (2001):

> Para gerenciar e manter os documentos, se faz necessário o uso de sistemas operacionais e de banco de dados cliente/servidor, *software*s de *workflow*, *software*s de tratamento e recuperação de imagens, além de implementações de segurança no acesso e backups (cópias de segurança). (SILVA, 2001)

Diante disso, a proposta deste trabalho é realizar um estudo acerca dos procedimentos pertencentes a cada uma das etapas do processo de digitalização de documentos. Posteriormente aplicar a informática no desempenho das tarefas, através, primeiramente, de uma análise seguida de uma seleção de *software*s livres que auxiliem nas etapas para elaboração dos objetos digitais. Os

*software*s foram analisados conforme requisitos estudados na disciplina de *Interface* Humano-Computador, sendo eles: viabilidade, eficiência e usabilidade. Esta pesquisa servirá de base para que as instituições possam fazer uso das ferramentas aqui descritas para que possam digitalizar seus acervos. Ao final será proposto um modelo para a implantação de um *workflow* referente ao processamento e a produção digital de documentos, visando a otimização do processo. O resultado é uma ferramenta de baixo custo que pode ser usada por acervos locais para a digitalização de seus documentos.

1.1 OBJETIVOS

O objetivo geral e os específicos serão descritos nas subseções 1.1.1 e 1.1.2 respectivamente.

1.1.1 Objetivo Geral

Estudar as etapas que envolvem o processo de digitalização de documentos e posteriormente fazer uma seleção de *software*s livres que oferecem tecnologia de baixo custo para o desempenho de cada uma dessas etapas. Em seguida, desenvolver um modelo de *workflow* englobando todas as fases da digitalização de documentos, para que assim o trabalho como um todo possa ser mais bem controlado e otimizado.

1.1.2 Objetivos Específicos

- Fazer um estudo sobre as atividades pertencentes a cada fase do processo de digitalização de documentos e as respectivas atividades.
- Fazer um levantamento e posteriormente uma seleção de *software*s livres que realizem: captura de imagens, tratamento de imagem, OCR,

reconhecimento de caracteres, transcrição de manuscritos, organização de meta-dados e armazenamento de arquivos.

- Instalar cada um dos *software*s e testá-los em relação à usabilidade, tecnologia e eficiência.
- Documentar os projetos para possibilitar sua integração em um *workflow* de Gerência Eletrônica de Documentos (GED).
- Desenvolver um *workflow* para o processamento digital de imagens.

1.2 ORGANIZAÇÃO DO TRABALHO

Este trabalho está dividido em 6 capítulos. O capítulo 1 apresenta a introdução e os objetivos gerais e específicos do trabalho.

O capítulo 2 aborda os conceitos envolvidos com a tecnologia *Workflow* a qual possibilita a automatização das diversas etapas envolvidas na digitalização de documentos.

O capítulo 3 descreve sobre a Gestão Eletrônica de Documentos, seus conceitos, aplicações, benefícios, entre outros.

O capítulo 4 relata as atividades realizadas no estudo de caso que tem como base o projeto realizado na iniciação científica acerca da digitalização de documentos. Neste capítulo, cada fase do processo será detalhadamente apresentada juntamente com os *software*s livres específicos que foram utilizados e com as telas para ilustrar o que foi trabalhado e, no final, será apresentado um protótipo de *workflow* para a possível integração destes *software*s.

O capítulo 5 exibe a conclusão do trabalho e o capítulo 6 apresenta idéias para pesquisas futuras que possam ser desenvolvidas com base neste trabalho.

2 *WORKFLOW*

Este capítulo é destinado ao estudo da tecnologia *Workflow*, serão apresentados a seguir a sua conceituação, funcionalidades, desenvolver histórico e os tipos existentes.

O processo de transformação de documentos físicos para mídias digitais, como veremos de maneira mais específica no decorrer do trabalho, é envolvido por diversos processos, desde a captura até seu armazenamento. Estes processos, quando feitos isoladamente, não permitem um controle de quais atividades e de que maneira estão sendo executadas, prejudicando assim o fluxo de trabalho, deixando-o mais lento e mais sujeito a falhas.

Cruz (2004) define que processo não é apenas um conjunto de atividades visando um objetivo comum, mas também “uma entidade regida por leis muito particulares, inerentes a sua natureza”. Assim, para otimizar o fluxo de trabalho e evitar o desperdício de tempo e esforço, todos os processos devem estar bem controlados e organizados.

Este capítulo abordará aspectos relacionados à tecnologia *Workflow*, um breve histórico, definições e os tipos de *workflow*, os quais servirão de embasamento para entendermos como ocorrerá o desenvolvimento de todo o processo para digitalização de documentos.

Podemos observar algumas funcionalidades da tecnologia *workflow* no Quadro 01 – Funcionalidades do *Workflow*.

Funcionalidade	Descrição
Definição de pesos para grupos	Possibilidade de definir pesos dentro de um grupo de forma que uma atividade possa ser designada a um membro do grupo baseada nos pesos definidos.
Visão de carga de trabalho	Possibilidade de determinar a carga de trabalho de cada usuário, permitindo que o administrador do processo determine quanto e que tipo de atividades estão pendentes para um usuário. Com estas informações, o administrador pode re-designar algumas ou todas as tarefas

	para outros usuários.
Designação de funções pelo cliente	Possibilidade dos participantes do processo designar ou retirar a execução de atividades para/de outros participantes. Esta capacidade é interessante para os casos em que um participante estará ausente ou deseja delegar a tarefa a outros usuários.
Subprocessos	Possibilidade de definir subprocessos em um processo. Isto permite o projeto de implementação de processos aninhados.
Rejeição de um passo de trabalho	Possibilidade de um participante rejeitar a execução de um passo de trabalho ou tarefa, fazendo com que o fluxo retorne ao passo anterior de execução.
Notificações	Possibilidade de configuração de notificações sobre ocorrência de eventos ao longo da execução de um processo sobre atrasos em sua execução.
Auditoria automática	Possibilidade de manter em um sistema de gerência de documentos versões dos documentos/formulários em cada passo do processo.
Priorização automática de atividades	Possibilidade de priorizar automaticamente as atividades de acordo com seu prazo estimado.
Grupos dinâmicos	Possibilidade de definir grupos responsáveis por uma atividade no momento de sua execução.
Flexibilidade	Possibilidade de redesenhar os processos conforme a necessidade.

Quadro 01 – Funcionalidades do *Workflow*
Fonte: Araújo e Borges (2001) apud ULTIMUS (1998)[1]

2.1 DEFINIÇÃO

Encontramos diversos conceitos para *Workflow*, porém, antes de apresentar as definições de diferentes autores, o conceito oficial de *workflow*, estabelecido pela

[1] ULTIMUS, 1998, 40 *Essential Features of Workflow Software You Will Not Find in Lotus Notes*, Disponível em: http://www.*workflow*zone.com, acesso em junho/2001

Workflow Management Coalition (WfMC), uma organização fundada em 1993 e composta por diversas empresas com o objetivo de definir padrões para a utilização do *workflow* em sistemas de gestão. Alguns dos padrões estabelecidos pela WfMC podem ser observados na Figura 01. De acordo com a WfMC, *workflow* é "é a automação de processos envolvendo combinações de humano e máquina de atividades baseadas, principalmente aqueles envolvendo interação com aplicações de TI e ferramentas", (HOLLINGSWORTH, 1995).

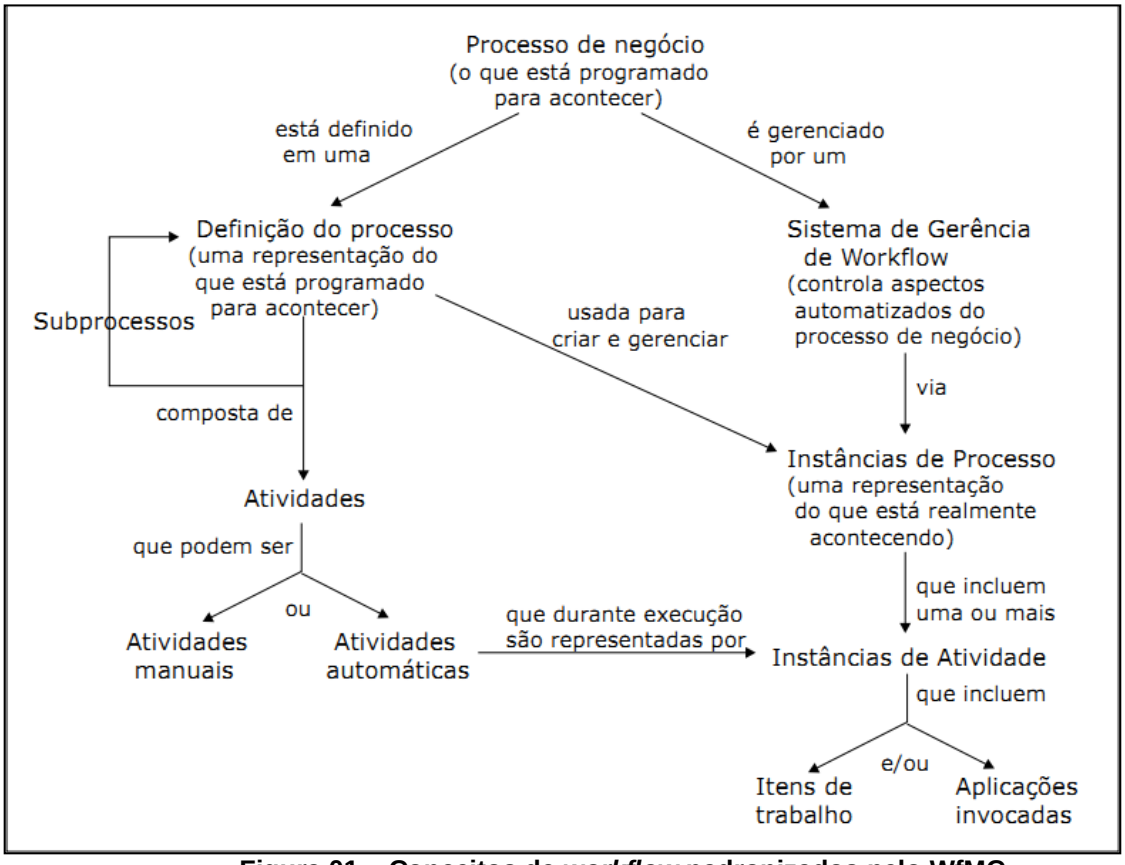

Figura 01 – Conceitos de *workflow* padronizados pelo WfMC
Fonte: Adaptado de Hollingsworth (1995)

A utilização do *workflow* consiste em regras pré-estabelecidas juntamente com designação para que cada componente do grupo saiba qual papel deve desempenhar. Estas regras podem ser estabelecidas tanto de forma manual, como também (e mais recomendado) através de utilização de um sistema de tecnologia da informação. O autor Junior (2008), conceitua *workflow* como uma "automação de processos de negócio, onde as atividades são passadas de um participante para o outro de acordo com um conjunto de regras definidas". Ainda segundo o autor,

podemos dizer que o gerenciamento do *workflow* é composto por várias ferramentas que controlam e administram os clientes integrantes do sistema, para que assim, cada um trabalhe em sua atividade e os resultados pertençam a um fluxo como um todo.

Em uma abordagem mais detalhada, Cruz (2004) explica a ligação existente entre outros importantes conceitos que compõe o *workflow*, entre eles: processo, tarefa, atividades e procedimentos. Ele define *workflow* como a automatização de processos, o termo processo, por sua vez, é vulgarmente utilizado para tudo que realizamos, perdendo seu verdadeiro sentido. O autor esclarece que o processo é formado por atividades que tendem a atingir um determinado objetivo e quando o processo torna-se muito complexo, ele é subdivido em chamados subprocessos. Cada atividade é composta por uma cadeia de procedimentos. O procedimento acaba sendo uma forma mais específica de executar a atividade, os procedimentos podem ser formais, quando indicam de que forma e em que tempo cada evento deve ser executado, ou podem ser informais quando formam práticas não escritas, mas que também são incorporadas à realização da atividade. E por último, porém não menos importantes encontram-se as tarefas, as quais são a última divisão da cadeia, as tarefas são a decomposição do procedimento. Podemos observar na Figura 02 um resumo de todos estes elementos.

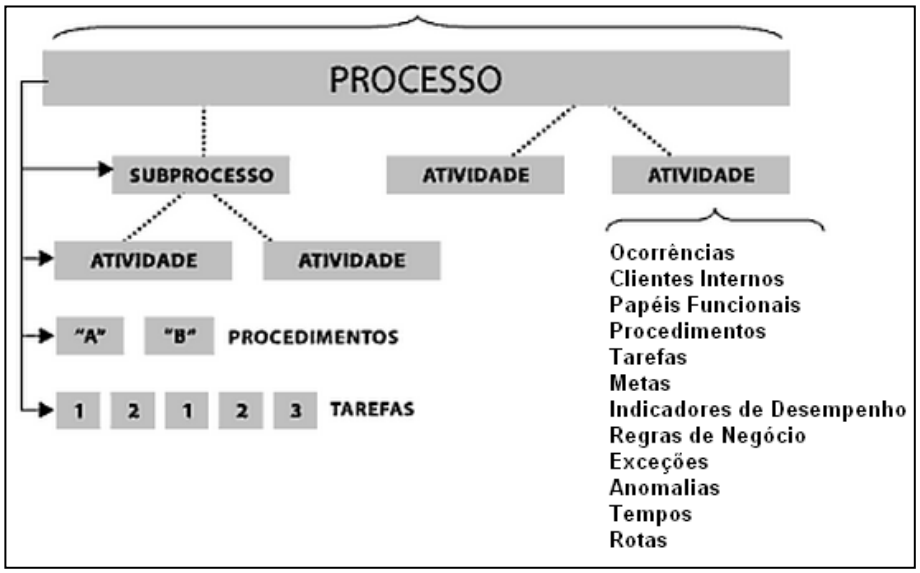

Figura 02 - Processo, Divisões e Elementos
Fonte: Cruz (2004)

Ao se tratar de sistema de gestão de *workflow*, Hollingsworth (1995) cita em sua obra que “é um sistema que completamente define, gerencia e executa fluxos de trabalho através da execução de *software* de computador que executa e representa a lógica de fluxo de trabalho”.

A utilização de *workflow* objetiva o melhoramento da coordenação do trabalho, disponibilizando e facilitando uma comunicação eletrônica no ambiente local de trabalho. O *workflow* integra processos realizados tanto por humanos como também por recursos de informática (*software*s e hardwares). (CRUZ, 2004).

O tempo de ciclo pode variar de acordo com sua natureza, objetivos, complexidade e atividades envolvidas.

2.2 HISTÓRICO

De acordo com Vieira (2005), os sistemas *Workflow* já existem há algum tempo, porém é a partir da última década que ele vem se popularizando. A autora cita que a origem do *workflow* deu-se na tentativa de automatização do trabalho em escritórios por volta da década de 70. O objetivo era diminuir o fluxo de documentos que circulavam nos escritórios, organizando e otimizando as transações. Destacam-se como pioneiros Skip Ellis e Michael Zisman, os quais trabalhavam na empresa Xerox e já utilizavam modelos de *workflow* baseados em redes Petri (denominação para as redes de controle de informação). (AALST; HEE, 2009).

Em se tratando de sistemas de informação, Aalst; Hee (2009) descrevem que do ano 1965 a 1975 as aplicações eram feitas todas de modo independente, e, cada uma continha suas próprias definições e base de dados. Não havia comunicação entre as aplicações. De 1975 até 1985 o período foi caracterizado pelo surgimento dos Sistemas de Gestão de Banco de Dados (SGBD), os quais inicialmente eram hierárquicos e em rede e mais tarde passaram a ser relacionados.

A partir dos anos 90, a tecnologia *Workflow* teve um grande avanço devido ao crescimento acelerado da informática, principalmente na área de redes de computadores, muitas organizações passaram a fazer uso de arquiteturas *workflow*, pois necessitavam de uma maior interação intra-organizacional e mais tarde até mesmo inter-organizacional. (ARAUJO; BORGES, 2001). Nesta época, os processos

passaram a serem extraídos das aplicações. “Um sistema de *workflow* organiza o encaminhamento de dados de um caso entre os recursos humanos e aplicações”. (AALST; HEE, 2009).

Atualmente, segundo Pereira e Casanova (2003), a tecnologia *workflow* vai mais além do que simplesmente reduzir o fluxo de papel dentro de uma organização, seria uma “ferramenta para a coordenação do trabalho de equipes, impulsionando seu desenvolvimento”. Os sistemas passaram de “orientado a dados” para “orientado a processos” (VIEIRA, 2005). Na citação a seguir, a autora faz uma síntese das gerações pelas quais o *workflow* desenvolveu-se:

> A primeira geração dos sistemas de workflow compreendeu aplicações monolíticas de uma área de domínio particular. A segunda geração já dividiu os sistemas em componentes diversos, mas ainda os deixou fortemente dependentes das aplicações. A terceira geração apresentou máquinas de workflow genéricas que forneciam uma infraestrutura robusta para workflows orientados à produção. Nesta geração a descrição dos workflows era construída através de ferramentas gráficas e posteriormente, era interpretada pelas máquinas de workflow, as verdadeiras responsáveis pela sua execução. Uma quarta geração seria a que se está presenciando nos últimos anos, a de sistemas de workflows que oferecem uma gama de serviços. (Vieira (2005) apud ELMAGARMID (1997)[2]).

As pesquisas relacionadas a *workflow* agora enfocam um trabalho cooperativo e descentralizado, com o uso de arquiteturas distribuídas de execução de processos. Assim, cada processo pode ser executado em locais diferentes, fazendo uso de recursos diferentes, tornando-se o mais independente possível dos demais, porém, ainda assim, todos sendo coordenados por um fluxo único. (ARAÚJO; BORGES, 2001).

Tendo uma noção de como surgiu o conceito de *Workflow*, passaremos a estudar as suas classificações na sessão seguinte.

2.3 TIPOS

A classificação de *workflow* varia conforme o autor. Esta variação é positiva ao ponto que consegue abranger as diversas necessidades particulares de cada

[2] DU, W.; ELMAGARMID, A.. *Workflow Management: State of the Art vs. State of the Products. HP Labs technical Report* HPL-97-90, HP *Labs*, 1997.

caso. Cruz (2004) divide o *workflow* em cinco tipos: *Ad hoc*, Administrativo, Produção/transação, orientado a objeto e baseado no conhecimento.

Workflow Ad hoc, é o modelo que está presente na maioria das classificações dos autores. O termo "*Ad hoc*" é uma expressão que deriva do latim e significa "para isto", "para este caso". Segundo Cruz (2004), *workflow* do tipo *Ad hoc*:

> É aquele criado para ser usado dinamicamente por grupos de trabalho cujos participantes necessitem executar procedimentos individualizados para cada documento processado dentro do processo de negócio. (CRUZ, 2004)

Uma das características deste tipo de *workflow* é que não possui regras de negócios formalizadas, são apropriados para processos de natureza simples, flexíveis e não exijam muita segurança. Ele é o mais simples e mais elementar de todos. O *Workflow Ad hoc* "requer ferramentas gráficas de desenvolvimento que possam ser utilizadas pelo usuário para criar e modificar seus procedimentos" (CRUZ, 2004). Os *software*s específicos para a implantação do *workflow Ad hoc* permitem tratar e armazenar diversos tipos de documento como dê sons, imagens, textos e realidade virtual. Porém, o ponto negativo é que muitos não são capazes de trabalhar com grande fluxo de dados.

Santana (2006) destaca que este tipo de *workflow* é utilizado em ambientes onde não existe um "padrão fixo para o fluxo de informações", por isso a característica de ser flexível e adaptável.

O *workflow* administrativo também é adequado para processos simples, porém agora estruturados de forma moderada, com processos e regras repetitivas e bem definidas, ou seja, orientado a processos administrativos. (CRUZ, 2004).

O terceiro tipo de *Workflow*, chamado de produção/transação é considerado para muitos autores dois modelos distintos, Cruz (2004) coloca-os na mesma categoria. Este tipo de *workflow* é adequado para processos que não envolvam muito a participação de pessoas e sim de *software*s integrados com sistemas que trabalham com um grande volume de dados em aplicações mais complexas. Geralmente as aplicações não podem sofrer interrupções nem pausas. Filho (2000) cita que "O *Workflow* do tipo produção ou transação, geralmente, envolve grandes quantidades de dados, muitas regras de negócio e recursos financeiros em grande escala." É aconselhável que tenha também uma auditoria do *Workflow*, para tornar o desenvolvimento mais seguro.

O quarto tipo de *Workflow* é chamado orientado a objetos. Este é o mais sofisticado modelo e incorpora a tecnologia orientada a objetos.

> Com o *Workflow* OO, permite-se a convivência de várias versões de fluxos de trabalho e regras diferentes para um mesmo objeto. O *Workflow* de produção e o orientado a objeto são semelhantes quanto ao tratamento de volumes de dados, a diferença é somente o uso da tecnologia Orientada a Objetos pelo último. (FILHO, 2000)

O modelo de *workflow* baseado no conhecimento utiliza a tecnologia de Inteligência Artificial, pois consegue “aprender com seus próprios erros”. Este modelo vai além da aplicação de regras, podendo prever e estabelecer exceções conforme as necessidades surgem. Esse tipo de *Workflow* ainda não está disponível para comercialização, existindo apenas como protótipos (CRUZ, 2000).

Com isso, pode-se ter uma idéia geral das características e dos conceitos que envolvem a tecnologia *Workflow*, os quais serão aplicados no estudo de caso deste trabalho.

3 GESTÃO DE DOCUMENTOS

Este capítulo trata da Gestão de documentos de modo indutivo, delineando, a *priori*, seus conceitos mais amplos, para em seguida tratarmos especificadamente sobre a gestão de documentos informatizados.

Antes de passarmos a falar sobre gestão de documentos, é imprescindível delinearmos um conceito para documentos de arquivo. Para o autor Belloto (2002) documentos são “unidades constituídas pela informação e seu suporte”, ele define mais especificamente documentos de arquivo como:

> Arquivos são conjuntos orgânicos de documentos produzidos / recebidos / acumulados por um órgão público, uma organização privada ou uma pessoa, no curso de suas atividades, independentemente do seu suporte, e que, passada sua utilização ligada às razões pelas quais foram criados, podem ser preservados, por seu valor informativo, para fins de pesquisa científica ou testemunho sociocultural. (BELLOTO, 2002).

Segundo Rondinelli (2002), pode-se traçar um breve histórico da trajetória dos documentos, destacando seus momentos significativos. De acordo com ele, as primeiras preocupações em armazenar documentos surgiram com a criação do Arquivo Nacional da França, em 1789. Mais tarde, através do chamado Decreto Messidor Francês, os arquivos passaram a ter acesso público. Outro momento significativo foi o pós II Guerra Mundial, onde houve um grande aumento no volume de documentos expedidos por instituições públicas, assim, veio à necessidade de tratar e organizar essa grande demanda de documentos, como diz a autora “sob pena de as organizações inviabilizarem sua capacidade gerencial e decisória”. Nesta etapa surge então o conceito de Gestão de Documentos:

> O conjunto de procedimentos e operações técnicas referentes às atividades de produção, uso, avaliação e arquivamento de documentos em fase corrente e intermediária, visando a sua eliminação ou recolhimento para guarda permanente. Rondinelli (2002 apud Indolfo, 1995)[3]

De um modo geral, pode-se dizer que o conjunto de vários procedimentos que visam à proteção, organização e armazenamento dos documentos é conceituado como gestão de documentos. Podemos observar a definição de gestão de

[3] INDOLFO, Ana Celeste. **Gestão de Documentos: conceitos e procedimentos**. Rio de Janeiro: Arquivo Nacional, 1999.

documentos na Lei 8.159 de janeiro de 1991, a qual Dispõe sobre a política nacional de arquivos públicos e privados e dá outras providências:

> Gestão de documentos é o conjunto de procedimentos e operações técnicas à sua produção, tramitação, uso, avaliação e arquivamento em fase corrente e intermediária, visando a sua eliminação ou recolhimento para guarda permanente. (BRASIL, 1991)

Juntamente com o conceito de gestão de documentos, também se sobressai a descoberta do conceito do ciclo vital, que compreende a trajetória de um documento. Rondinelli (2002) explica quais são as três fases da vida do documento. A primeira fase é caracterizada pela organização dos documentos e assim, utilizados ativamente por seus criadores. Em seguida, na segunda fase, os documentos passam a ser armazenados e o uso acaba por não se tornando mais tão freqüente. Por último, como terceira e última fase, quando seu uso operacional encerra, eles passam por uma seleção, dessa seleção é possível dividir os documentos, onde alguns são transferidos e arquivados e outros são descartados. Esta idéia de ciclo vital, fez com que cada uma das fases (também chamadas de corrente, intermediário e permanente) tivessem entendimento, cuidados e procedimentos específicos. Partindo desta premissa, podemos constatar as três fases da gestão de arquivos, que compreende a produção, utilização e destinação.

A partir da década de 80, os documentos eletrônicos passaram a ser muito utilizados, mudando os rumos da arquivologia e passando a ser um novo marco nesta trajetória. (RONDINELLI, 2002).

No âmbito do Brasil, nesta mesma época, também percebemos a preocupação do governo com o acesso dos cidadãos aos documentos, surgindo uma legislação Arquivística através da Lei nº 8.159 de 08 de janeiro de 1991 – Lei da Política Nacional de Arquivos Públicos e Privados. Conforme esta lei anteriormente citada, gestão de documentos é “o conjunto de procedimentos e operações técnicas referentes à sua produção, tramitação, uso, avaliação e arquivamento em fase corrente e intermediária, visando a sua eliminação ou recolhimento para guarda permanente”. (BRASIL, 1991).

Juntamente com a Lei nº 8.159 foi criado o Conselho Nacional de Arquivos – CONARQ que foi regulamentado pelo Decreto nº 1.173 de 29 de junho de 1994.

Os autores RONCAGLIO et al. (2004) sintetizam alguns dos benefícios da gestão de documentos:

> Administrar ou gerenciar documentos arquivísticos, a partir da aplicação de conceitos e teorias difundidas pela Arquivologia, garante às empresas públicas ou privadas obter maior controle sobre as informações que produzem e/ou recebem, racionalizar os espaços de guarda de documentos, desenvolver com mais eficiência e rapidez suas atividades, atender adequadamente clientes e cidadãos. (RONCAGLIO et al., 2004).

3.1 GESTÃO DE DOCUMENTOS INFORMATIZADOS

Um Sistema de Gestão de Documentos Informatizados, de acordo com o Conselho Nacional de Arquivos (2010) é um conjunto de ferramentas informatizadas (podendo ser composto por um ou mais *softwares*) desenvolvidos para produzir, receber, armazenar, dar acesso e destinar documentos arquivísticos.

Veremos a seguir de forma mais detalhada a gestão eletrônica de documentos, definição, tipos, exemplos e aplicações.

3.2 GESTÃO ELETRÔNICA DE DOCUMENTOS (GED)

A sigla GED significa Gerenciamento Eletrônico de Documentos ou Gestão Eletrônica de Documentos. É uma tecnologia já existente que permite a digitalização e disponibilização de acervos documentais, porém trata-se de uma tecnologia ainda cara e disponível apenas para grandes instituições em nível nacional.

A GED foi introduzida no Brasil pelo Centro Nacional de Desenvolvimento do Gerenciamento da Informação - CENADEM, que iniciou suas atividades em 1976.

3.2.1 Definição

O volume de documentos em meios físicos que circulam por empresas e organizações tem aumentado significativamente. Rondinelli (2002) coloca como as causas deste acontecimento: aumento dos negócios, aumento da burocracia, disseminação da educação, crescimento dos meios de comunicação. Assim, as

pessoas passaram a se expressar e se manifestar mais através do papel. O que antes eram apenas palavras, conversas, acordos, hoje se efetivou como documento.

Os documentos formam uma parte importantíssima dentro de uma empresa ou organização, eles são registros de acontecimentos, parte do patrimônio e principalmente fonte de pesquisa. Estas pesquisas históricas necessitam de fontes e como cita Pena e Silva (2008):

> Às vezes as fontes existem, mas não estão organizadas, catalogadas ou mesmo disponíveis. Foi tentando minimizar este problema que surgiram museus, arquivos, bibliotecas e, mais recentemente, os centros de documentação. No entanto, os mesmos, se não possuírem um sistema organizacional e de consulta rápido e eficiente, não lograrão bons resultados e seu trabalho ficará comprometido. (Pena; Silva, 2008)

Por isso, visando diminuir e organizar essa grande massa de documentos surge a nova tendência é transformá-los em mídias digitais.

O GED pode trabalhar com documentos originalmente eletrônicos e com os originalmente em papéis. Este último caso recebe a designação de digitalização que pode ser explicada como o processo que transforma a informação do papel para um formato digital, o qual pode ser manipulado em um computador. Os documentos em mídias digitais podem ser mais bem organizados e pesquisados. A transformação do formato do documento também possibilita a modificação e o melhoramento na qualidade do mesmo.

O Conselho Nacional de Arquivos – CONARQ conceitua digitalização de documentos como:

> Entendemos a digitalização como um processo de conversão dos documentos arquivísticos em formato digital, que consiste em unidades de dados binários, denominadas de bits - que são 0 (zero) e 1 (um), agrupadas em conjuntos de 8 bits (*binary digit*) formando um byte, e com os quais os computadores criam, recebem, processam, transmitem e armazenam dados. De acordo com a natureza do documento arquivístico original, diversos dispositivos tecnológicos (hardware) e programas de computadores (*software*) serão utilizados para converter em dados binários o documento original para diferentes formatos digitais. (CONARQ, 2010)

A Digitalização contribui com a proteção do acervo documental, porém, não se deve esquecer que apesar do documento digitalizado ser uma cópia fiel, ele não substitui o original.

.

> Em linhas gerais, podemos descrever GED como um conjunto de tecnologias que permite a uma empresa gerenciar seus documentos em forma digital. Esses documentos podem ser das mais diversas origens, tais como papel, microfilme, imagem, som, planilhas eletrônicas, arquivos de texto etc.(GESTÃO ELETRÔNICA DE DOCUMENTOS, 2011).

O GED pode abranger o processo de conversão para forma digital vários formatos de documentos, podendo ser imagens, documentos de áudio ou textos.

Através de uma união de *software*s e hardwares específicos pode-se realizar as etapas de digitalização as quais compreendem a: captação, tratamento, armazenamento e gerenciamento de documentos.

Este gerenciamento eletrônico de arquivos não compreende somente a função de armazenar, mas sim também as demais etapas como organização, acompanhamento do processo (desde a captura até o seu arquivamento) juntamente com o controle sobre os mesmos, ou seja, o GED atua nas três fases do ciclo de vida do documento, como visto anteriormente. (RONDINELLI, 2002).

Os sistemas de Gerenciamento Eletrônico de Documentos permitem que os usuários acessem as informações de forma ágil e mais segura.

3.2.2 Importância e Benefícios

Apesar de algumas organizações não terem acesso a esta tecnologia, é importante destacar alguns benefícios que a digitalização e controle de documentos por esta ferramenta traz. De uma forma sintetizada, podem-se citar como principais vantagens a otimização no tempo na busca dos mesmos, além de economia em espaço físico destinado ao arquivamento e acesso.

Os documentos ficam armazenados de forma que as pessoas possam ter acesso à informação de forma segura em um acervo organizado e controlado.

> Já que o acervo foi digitalizado durante o processo de implantação da gestão eletrônica de documentos, agora que os documentos estão no computador, fica bem mais fácil controlar o acesso, distribuindo senhas e definindo níveis de acesso para estas senhas. (GESTÃO ELETRÔNICA DE DOCUMENTOS, 2011).

Outro grande benefício trazido pela utilização da GED é a possibilidade de fazer *backup* automatizado e de forma simples dos dados, garantindo melhor segurança e preservando a versão original.

De acordo com o CONARQ (2010) os benefícios que a gestão eletrônica de documentos traz são:

• Facilitar o amplo acesso e disseminação dos documentos arquivísticos através das mídias eletrônicas.

• “Permitir o intercâmbio de acervos documentais e de seus instrumentos de pesquisa por meio de redes informatizadas;”

• Promover a difusão e reprodução dos acervos arquivísticos não digitais.

• Preservar os documentos originais, pois somente as cópias serão manuseadas.

Silva (2001) também destaca benefícios da utilização do GED como, por exemplo, a facilidade de executar pesquisas, a disponibilidade do mesmo documento para diversos usuários, a fácil manutenção do acervo, uma possível correção de erros para tornar o documento mais legível, a economia de papel e de espaço físico. O autor estima a quantificação dessa economia de espaço, “um arquivo metálico de quatro gavetas, com capacidade de 22.000 documentos, pode ser facilmente acomodado em 1GB de espaço de disco.”

Muitos documentos históricos ficam armazenados e preservados devido a sua importância histórica, assim, o público em geral não tem acesso a eles. A gestão eletrônica de documentos permite que os acervos tornem-se mais acessíveis ao público, disponibilizando-os através de uma consulta eletrônica, o que acaba sendo mais fácil rápido e eficaz. As pessoas podem encontrar diretamente o que estão buscando no momento em que quiserem.

O gerenciamento eletrônico de documentos - GED possui diversos benefícios, porém, como tudo, também vem acompanhado de algumas desvantagens. Os equipamentos necessários para a implantação desta tecnologia acabam tendo um custo muito alto. O planejamento e a programação deste sistema devem ser extremamente bem planejados, pois corre um grande risco do sistema não funcionar da forma esperada. A proposta do projeto que será explicado na sessão seguinte visa encontrar *software*s livres que possam ser utilizados na execução das tarefas que compões as etapas do GED. (SILVA, 2001).

Atualmente, encontramos no mercado diversas empresas que oferecem o serviço de GED. Apesar de o GED ser regido por aspectos genéricos, ele é bastante flexível e adaptável especificamente para cada situação. Por isso, cada caso deve

ser bem analisado para que se possa estabelecer uma comparação e escolher qual empresa fornece um sistema de GED mais viável para ser contratado.

De acordo com Nascimento et al., (2006), a digitalização também pode aumentar o valor intelectual de seus materiais sob a ótica educacional, exemplo: as cópias digitais podem circular por diferentes instituições e países em tempo real, o acervo digital criado pode ser acessado por novos públicos o que possibilita novas formas de explorar o material, o material pode ser pesquisado e manipulado em formato eletrônico, a digitalização pode resultar em catálogos eletrônicos, que permitem consultas rápidas e uma maior flexibilidade na utilização dos mesmos.

3.2.3 Aplicações

O GED pode ter aplicação em diversas áreas. Haddad (2000) divide a aplicação do GED em dois macro grupos: o gerenciamento de documentos e o gerenciamento de imagens. Segundo ele no primeiro grupo as “informações estão em estado dinâmico e no segundo estão em estado estático”.

. Silva (2001) listou mais detalhadamente essas áreas, chamando-as de categorias genéricas. De acordo com ele são sete categorias genéricas: “Aprimoramento do processo de publicação; Suporte a processos organizacionais; Suporte à comunicação entre pessoas e grupos na organização; Aperfeiçoamento do acesso a informações externas; criação, manutenção e distribuição da documentação; Sustentação dos registros da corporação e Promoção de Treinamentos e Educação.”

A primeira categoria chamada de aprimoramento do processo de publicação, de acordo com Silva (2001), enquadra organizações que “possuem documentos como seus produtos”, por exemplo, que produzem manuais, guias, jornais, entre outros. A tecnologia GED auxilia na organização das atividades, desde a criação até a impressão e posteriormente sua distribuição.

> Serão criados num meio eletrônico, distribuídos através de uma rede de computadores e impressos quando for necessário. Com o armazenamento da versão digital do documento, o trabalho de revisão pode ser feito de maneira mais freqüente, e assim, o tempo entre versões se torna menor. (SILVA, 2001)

Os custos da organização ou da empresa também diminuem, os depósitos e o tempo da entrega são melhores controlados.

A segunda área é chamada de processos organizacionais. O GED nesse caso é aplicado a organizações nas quais circulam formulários, relatórios, memorandos, autorizações e outros documentos internos. O GED gerencia o fluxo dos documentos, fazendo com que cheguem de forma mais rápida e certa no destino, aumentando assim o nível de confiabilidade das transações. (SILVA, 2001)

A terceira área é a de suporte à comunicação entre pessoas e grupos na organização. Tudo que precisa ser comunicado pode tornar-se documento (por exemplo, uma apresentação ou uma conversa gravada.). O GED atua como auxiliador na comunicação dentro da organização, compartilhando os documentos entre pessoas que fazem uso dessas informações. (Silva, 2001)

A quarta área é a de aperfeiçoamento do acesso a informações externas, onde o GED atua na busca de fontes externas e essenciais para a organização. Silva (2001) cita como exemplo desta área o “consórcio de universidades, agências governamentais, institutos de pesquisas, entre outros” onde disponibilizam um grande acervo de materiais para consulta.

A quinta área é a de criação, manutenção e distribuição da documentação. O GED auxilia trabalha com todas as etapas e ainda permite que usuários internos acessem a documentação através da rede e de uma estação de trabalho. Os usuários externos recorrem aos internos, os quais poderão atendê-los de forma ágil, pois terão a localização e um melhor controle dos documentos em questão. (SILVA, 2001).

A penúltima área de aplicação do GED definida por Silva (2001) é a de sustentação dos registros da corporação. O autor especifica a atuação do GED: ‘Nesta área, o GED estará lidando com documentos de história, *performance* financeira, acordos, deveres, contratos e outros tipos que dão suporte legal para a organização” (SILVA, 2001). O GED controlará os documentos oficiais corporativos, permitindo o armazenamento e as consultas ocasionais.

E, por fim, a última área de aplicação do GED é a de Promoção de Treinamentos e Educação. Nesta categoria enquadram-se documentos com fins de aprendizagem e treinamento do usuário. Os documentos são de natureza interativa e o GED auxiliará na sua distribuição, permitindo que o usuário possa interagir no processo de aprendizado. (SILVA, 2001).

Há ainda outros exemplos de aplicação da tecnologia GED, como: comércio eletrônico, bancos, automação de cartórios, bibliotecas digitais, contratos (de diversas naturezas), conversão de acervo histórico, disponibilização de documentos oficiais, tais como diário oficial, documentação e processos jurídicos, documentação em escritórios, gestão de documentos em hospitais, entre outros. Atualmente está em voga a gestão de documentos eletrônicos no âmbito jurídico, pois recentemente houve uma grande mudança no modo como os processos acontecem, agora, utilizando-se somente de métodos eletrônicos. (GESTÃO ELETRÔNICA DE DOCUMENTOS, 2011).

4. ESTUDO DE CASO

Este capítulo conterá a descrição do projeto. Aqui, cada etapa, resultados e dificuldades serão relatados.

Pode-se observar, no decorrer do trabalho, os benefícios advindos da gestão eletrônica de documentos, porém, o grande problema que as instituições ou organizações ainda encontram é no alto custo para tal implantação. O projeto focou na pesquisa e seleção de *softwares* livres que possam ser utilizados para a realização de cada uma das etapas na produção de fontes digitais. Assim, o processo como um todo acaba tornando-se de baixo custo, o que possibilita que instituições menores tenham acesso e possam digitalizar seus acervos.

O desenvolvimento das etapas que serão vistas a seguir foi realizado no ambiente operacional Ubuntu 11.10, sistema de código aberto e construído a partir do núcleo LINUX. Todos os *softwares* envolvidos foram utilizados na versão para Linux.

As etapas que foram estabelecidas para o desenvolvimento do projeto abrangeram desde a captura dos objetos até sua colocação no repositório digital. Podemos observar conforme a Figura 03 apresentada a seguir o esquema das etapas da produção dos objetos digitais, os quais são o resultado da conversão do documento físico para o formato digital.

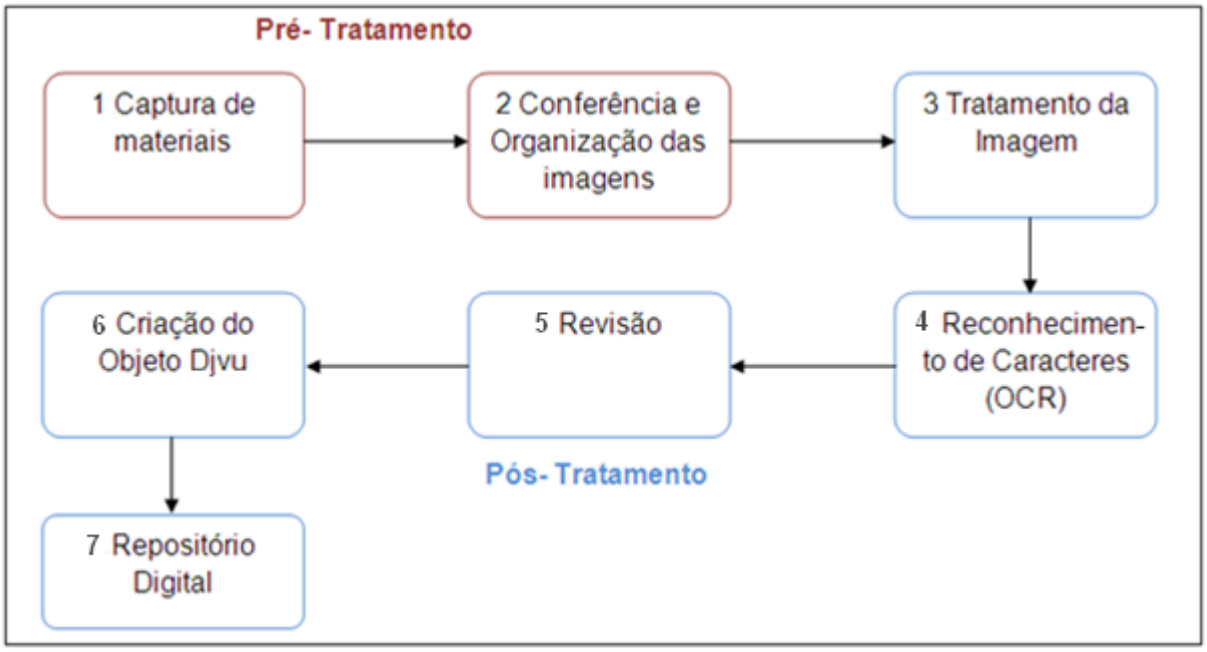

Figura 03 – Esquema de Etapas da produção do objeto digital
Fonte: Autoria Própria

4.1 Captura de materiais

Nascimento et al.(2006) afirmam que, em geral, antes de iniciar o processo de digitalização, é necessário buscar o equipamento ideal de acordo com o tipo de documento que será digitalizado. De acordo com as recomendações do Conselho Nacional de Arquivologia – CONARQ (2010) é necessário fazer a escolha dos materiais de captura que possibilitem a reprodução, no mínimo, da mesma dimensão e com as mesmas cores do documento original.

O processo de captura deve ser feito sempre com o objetivo principal de aproximar ao máximo a cópia com o documento original. Deve-se também se atentar quando as condições físicas do material original, tomando o cuidado necessário para o seu manuseio. È interessante que partes como capas, e páginas sem conteúdo, mas com a numeração, também sejam capturadas, para assim auxiliar na organização do documento como um todo. Os Meta-dados (informações e características sobre os dados já existentes) devem ser adicionados aos documentos ou armazenados em um banco de dados. Quando reproduzimos um documento, é essencial que todo o seu conteúdo acompanhe a cópia, o que facilita para os futuros pesquisadores e preserva a essência do documento original. (CONARQ, 2010).

> A qualidade da imagem digital é o resultado dos seguintes fatores: da resolução óptica adotada no escaneamento, da profundidade de bit, dos processos de interpolação (quando utilizados) e dos níveis de compressão, além das características dos próprios equipamentos e técnicas utilizadas nos procedimentos que resultam no objeto digital. (CONARQ, 2010)

O projeto baseou-se na conversão digital de livros e documentos escritos. Em razão disso, os aparelhos que fazem a captura dos documentos podem ser *scanner* ou câmera digital.

É importante escolher bem a ferramenta de captura, porque como cita Alves (2003) os problemas na digitalização “podem afetar o formato de um caractere, podendo, inclusive, perder *pixels*, deixando “buracos” que podem dificultar o reconhecimento”.

Conforme Nascimento et al. (2006) existem *scanners* de vários tipos e tamanhos, ao escolhê-los deve-se observar e analisar requisitos como: sua

resolução, profundidade de cores, área de escaneamento, tempo de digitalização e usabilidade.

Quanto à resolução, Nascimento et al. (2006) explicam que se refere ao máximo de detalhes que podem ser capturados de um documento, “a resolução óptica de um *scanner* é medida pela capacidade de leitura de seu sensor de imagem”. A profundidade de cores representa o número de cores que cada *pixel* da imagem pode ter, quanto maior a profundidade de cores, mas próximo ao original a reprodução consegue ficar. A área de escaneamento é a superfície de vidro do aparelho, esta área varia de acordo com o modelo do *scanner*, geralmente o tamanho padrão é para folhas A4. Alguns *scanners* possuem a tampa removível, o que facilita no escaneamento de documentos espessos. O tempo de digitalização varia de acordo com a resolução escolhida, quanto maior a resolução mais tempo o *scanner* levará para fazer a leitura da imagem. A usabilidade do *scanner* depende do *software* que vem integrado a ele, em muitos casos o *software* é de difícil utilização.

Para os documentos mais antigos e frágeis, é recomendado que se utilize câmeras digitais. As câmeras digitais são viáveis para o escaneamento de grandes arquivos, pois não sofrem limitações quanto a área de escaneamento. As câmeras permitem que o usuário configure a resolução que será utilizada, a iluminação, o modo de cena da imagem, entre outros quesitos. Deve-se ter cuidado quanto a iluminação utilizada para a fotografia. As imagens fotografadas são armazenadas no cartão de memória da câmera e dependendo da resolução, a imagem pode ficar com um tamanho maior, necessitando que seja continuamente descarregada em um computador. (NASCIMENTO et al., 2006)

As câmeras digitais podem ser adaptadas e acopladas a equipamentos para a captura em massa de documentos.

4.2 Conferência e organização dos Imagens

Após a fase de captura é necessário que os arquivos sejam conferidos, a fim de verificar se todas as partes do documento foram capturadas e se nenhuma apresenta falhas. Em seguida os arquivos devem ser organizados, porque devido a alguns documentos serem compostos por diversas páginas, para tornar mais fácil as próximas etapas é recomendado que seja estabelecido um prefixo ou sufixo para

arquivos do mesmo grupo. Por exemplo, um livro digitalizado pode ter genericamente o prefixo “livro01” e em seguida o respectivo número de cada página: livro01-01, livro01-02, livro01-03 e assim sucessivamente.

Esta tarefa auxiliará no desenvolvimento das próximas etapas. Dependendo da demanda de documentos, existem alguns *softwares* que renomeiam os arquivos automaticamente, bastando o usuário designar o prefixo ou o sufixo desejado.

Os *softwares* levantados e analisados para realizar esta tarefa foram os seguintes: GPRename e KRename, ambos são de distribuição gratuita e podem ser utilizados no sistema Ubuntu 11.10. Os dois *softwares* foram instalados através da Central de Programas do Ubuntu. Utilizando os dois *softwares* é possível compará-los quanto à eficiência e a usabilidade de cada um. O *software* GPRename falha no quesito de usabilidade, sua *interface* apesar de parecer simples, é de difícil utilização comparado com o outro. A Figura 04 mostra a *interface* do *software* GPRename.

O KRename é um *software* mais completo e mais prático. O KRename utiliza abas que o usuário possa controlar o seu trabalho. É possível anexar um grupo de arquivos e, para organizá-los, colocar prefixos e sufixos no nome do arquivo e números para controlar o índice dos mesmos. A *interface* do programa pode ser vista na Figura 05.

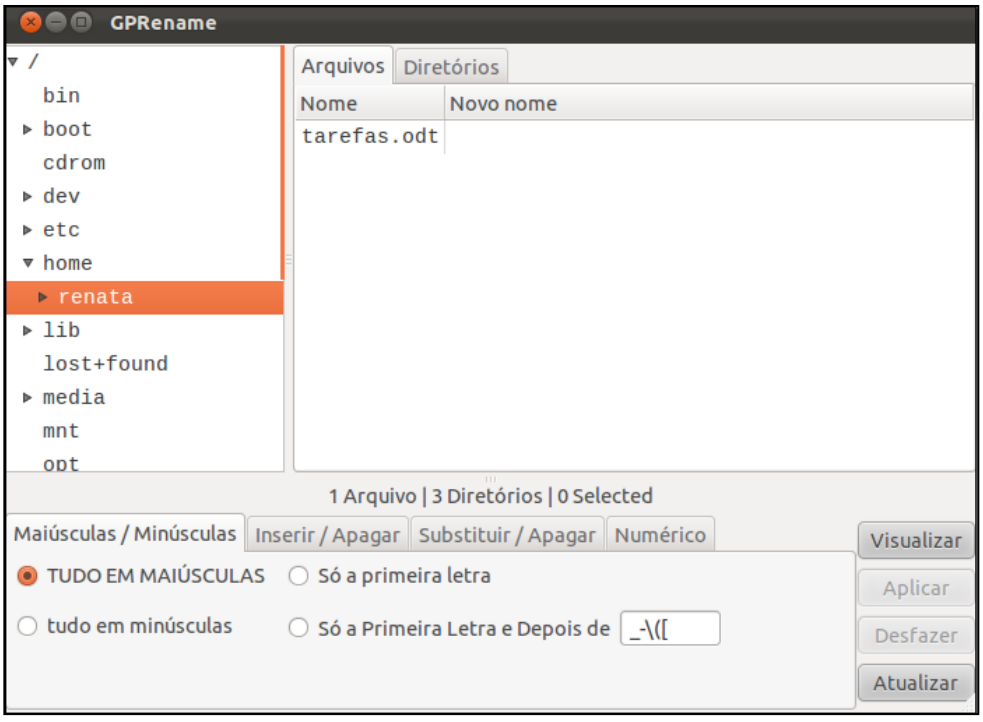

Figura 04 – *Interface GPRename*
Autor: TRISTESSE (2012)

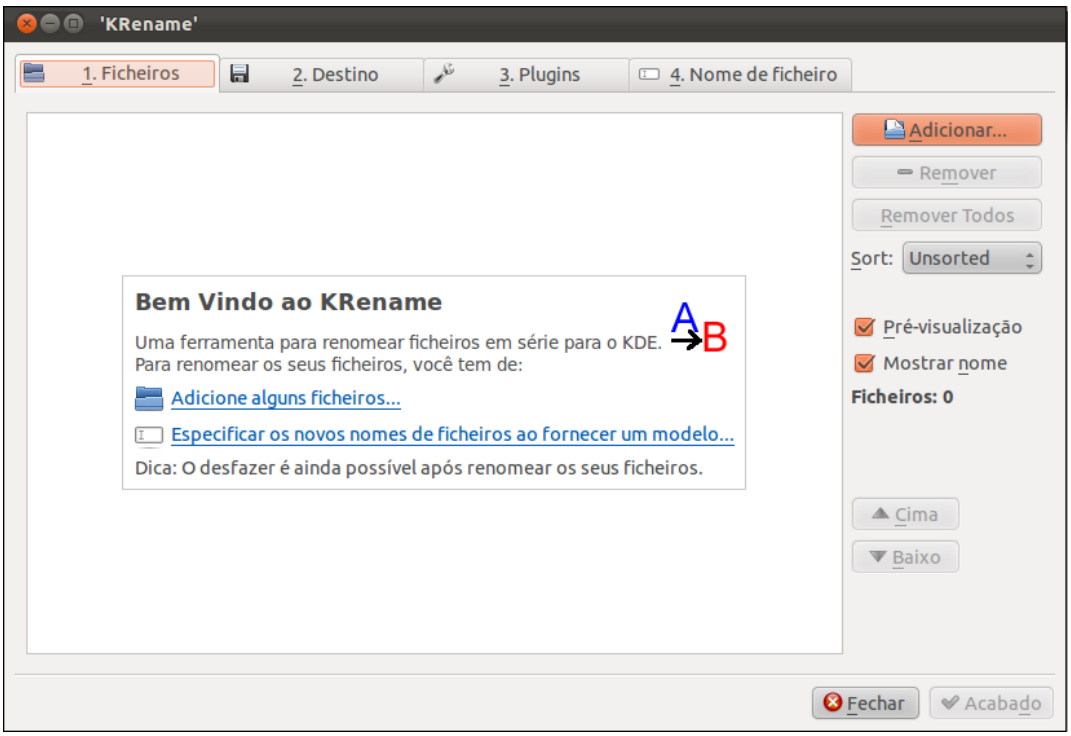

Figura 05 – *Interface KRename* 1
Autor: SEICHTER (2012)

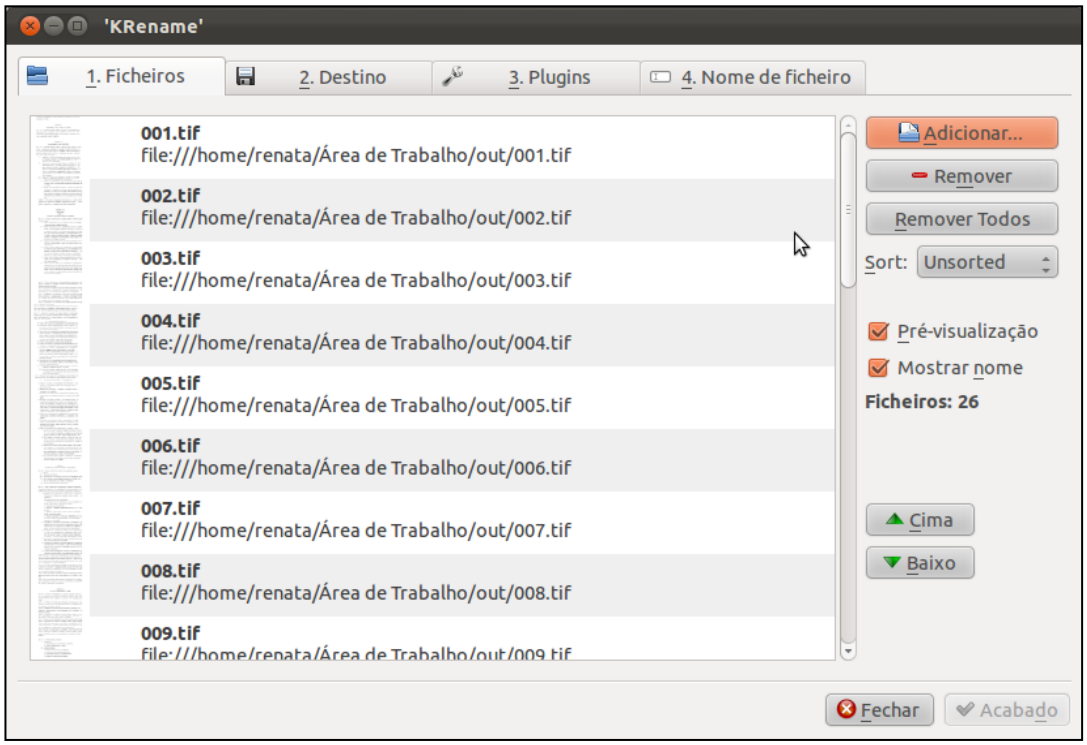

Figura 06 – *KRename* - Ficheiros
Autor: SEICHTER (2012)

Como pode ser visto na Figura 06, no KRename os arquivos são adicionados na aba “Ficheiros”, ao selecionar a aba uma lista de arquivos é apresentada. O software permite que os ficheiros sejam adicionados ou removidos da lista. O ideal é que o trabalho seja feito em grupos de arquivos, para que todos pertencentes ao mesmo grupo sejam nomeados conforme as especificações escolhidas. A aba “Destino”, conforme pode ser visualizada adiante na Figura 07, possibilita ao usuário escolher o local que os arquivos renomeados serão salvos, o usuário então pode escolher se quer somente renomear os ficheiros, movê-los para outro local ou fazer uma cópia dos mesmos para o caminho especificado.

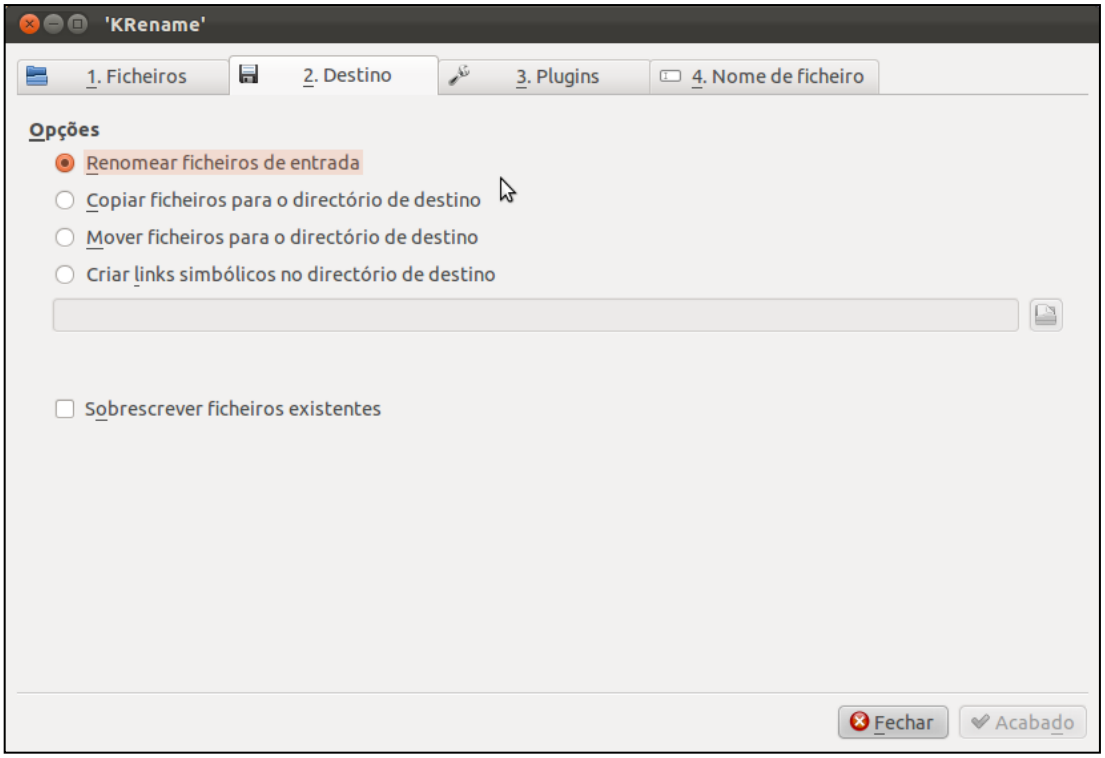

Figura 07 – *KRename* – Destino
Fonte: SEICHTER (2012)

A aba *plugins*, conforme demonstra a Figura 08, permite que o usuário instale ferramentas adicionais no software, como por exemplo um adicional de data e hora, ou itens especiais para arquivos de vídeo.

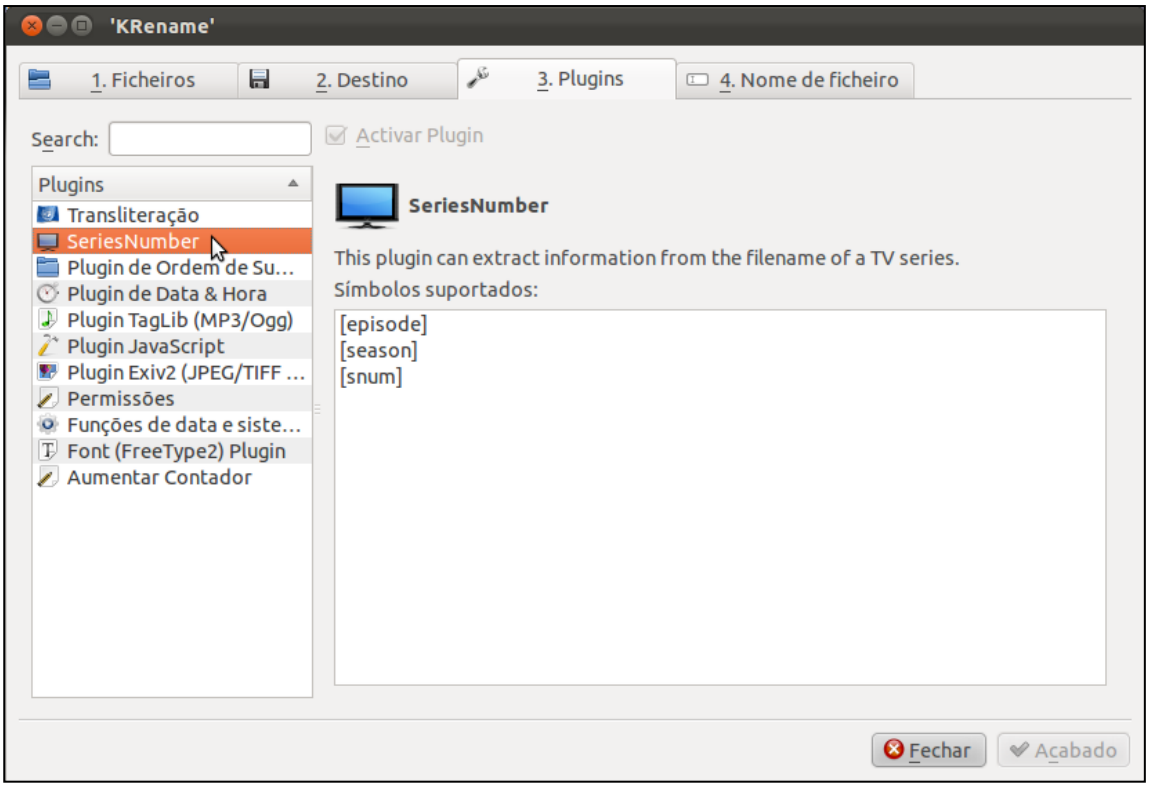

Figura 08 – *KRename – Plugins*
Fonte: SEICHTER (2012)

A aba “Nome do Ficheiro” é a que efetivamente possibilita que o usuário escolha como ele quer que os arquivos sejam renomeados. O usuário pode optar por colocar prefixos ou sufixos e o número seqüencial de índice dos arquivos. Conforme as edições são feitas, no campo abaixo vai aparecendo uma pré-visualização de como ficarão as modificações. Após terminar a edição o usuário seleciona o botão “Acabado” para salvar. Os arquivos são renomeados em massa e organizados, estando preparados para a próxima etapa do processo. A Figura 09 mostra a interface da aba “Nome de Ficheiro”.

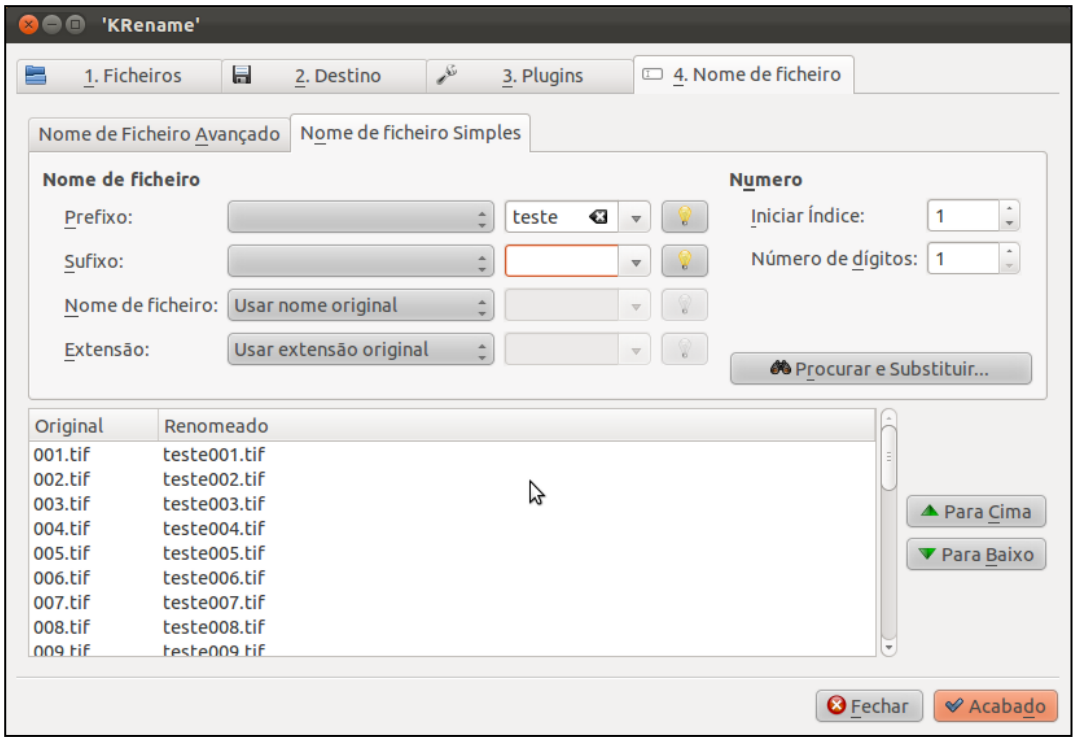

Figura 09 – *KRename* – Nome do Ficheiro
Fonte: SEICHTER (2012)

4.3 Tratamento da Imagem

Marques e Vieira (1999) citam que o principal objetivo das técnicas de tratamento de imagens é “processar certa imagem de modo que a imagem resultante seja mais adequada que a imagem original para uma aplicação específica”.

> “Algoritmos eficientes de processamento de imagens devem ser aplicados antes do reconhecimento de caracteres, procurando corrigir defeitos na imagem ou extrair dados não só inúteis para o reconhecimento, mas também que podem dificultar o processo.” (Alves, 2003)

Em uma abordagem mais técnica, Gomes (2007) a imagem digital pode ser entendida como um conjunto de *pixel*s. *Pixel* é a menor unidade da imagem digital. O autor cita que “uma imagem digital é uma matriz onde cada um de seus elementos é um número que representa a cor ou a intensidade do *pixel* de posição correspondente na imagem real”. Assim, a resolução da imagem é dada pelo número de *pixel* por unidade de distância. A resolução é um dos parâmetros básicos para a análise digital da imagem, deve ser considerada no processo de aquisição da

imagem, a escolha da resolução interferirá nas demais etapas de tratamento da imagem.

A quantização também é um ponto que deve ser analisado, ela corresponde ao “número máximo de níveis de intensidade ou cor que essa imagem pode apresentar”. (GOMES, 2007).

Ainda de acordo com Gomes (2007), principalmente para imagens em escala de cinza ou também conhecidas como monocromáticas, outros requisitos que devem ser considerados para a análise da imagem digital é o brilho e o contraste. Se os *pixel*s tiverem níveis de cinza baixos, próximo ao zero, a imagem fica muito escura, se os *pixel*s se aproximarem do branco ou do nível 255 a imagem ficará muito clara. O contraste deve ficar em um nível padrão, balanceando os níveis mínimos e máximos de intensidade dos *pixel*s.

O *software* livre escolhido para o tratamento de imagens em escala de cinza é o *Scan Tailor* para distribuições Linux. Ele pode ser instalado pela Central de Programas do próprio Ubuntu, ou através do download no site do desenvolvedor *(http://Scantailor.sourceforge.net/).*

Este *software* permite que a imagem seja balanceada na questão do brilho e do contraste, também é possível retirar os ruídos das imagens. Ruídos são pequenas impurezas que atrapalham a boa visualização da imagem. Além disso, o *Scan Tailor* permite que a imagem seja cortada, alinhada e corrigida quanto ao seu ângulo de inclinação e suas margens. É uma ferramenta ideal para imagens de texto advindas de digitalização de livros por exemplo, pois possibilita que as edições possam ser aplicadas em todas as páginas do arquivo. As ferramentas deste *software* são autoexplicativas. É recomendado que as imagens digitalizadas estejam no formato .tiff, assim é possível manter a qualidade da imagem. As mudanças de formato da extensão das imagens muitas vezes resultam na perda de sua qualidade.

A Figura 10 representa a *interface* inicial do *software Scan Tailor*:

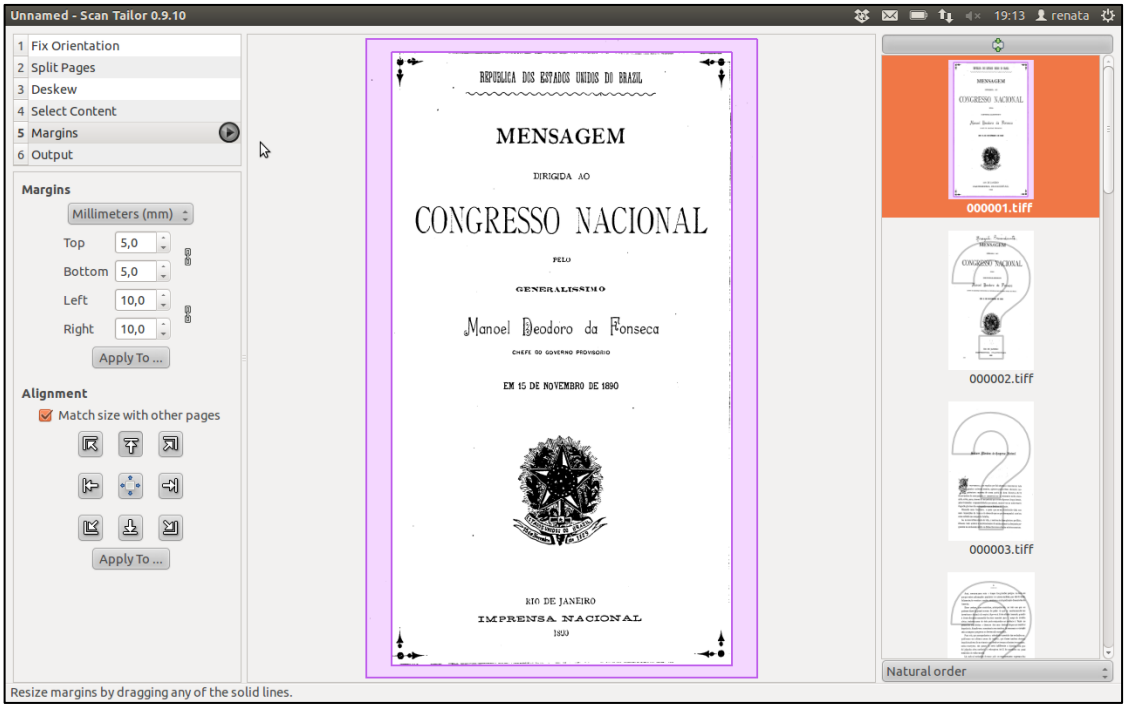

Figura 10 – *Scan Tailor – Interface* Inicial
Fonte: ARTSIMOVICH (2012)

A Figura 11 apresenta a tela da opção “Divisão de Páginas”. Nesta tela o usuário pode configurar o *layout* da página, para a direita, esquerda ou divido ao meio. Normalmente esta ferramenta vem marcada com a opção “Detectar automaticamente”, o que pode ser aplicado para a página atual ou para todas dentro do projeto.

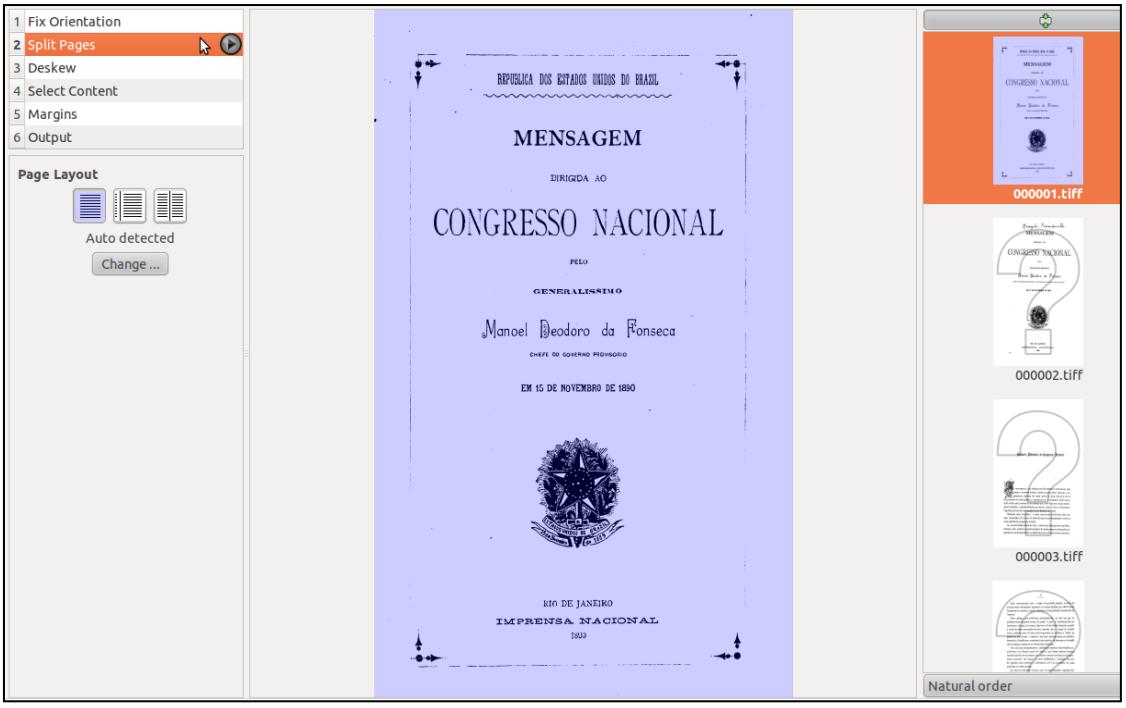

Figura 11 – *Scan Tailor* – Divisão de Páginas
Fonte: ARTSIMOVICH (2012)

A Figura 12 demonstra a etapa de alinhamento. Nesta etapa é possível editar o ângulo de alinhamento do documento. No momento da captura, o documento não fica exatamente na posição correta, o que acaba gerando uma imagem digitalizada fora do eixo. O fundo quadriculado e os eixos principais (x, y) facilitam o alinhamento do documento. A edição pode ser feita para uma página ou estendida para todas do conjunto.

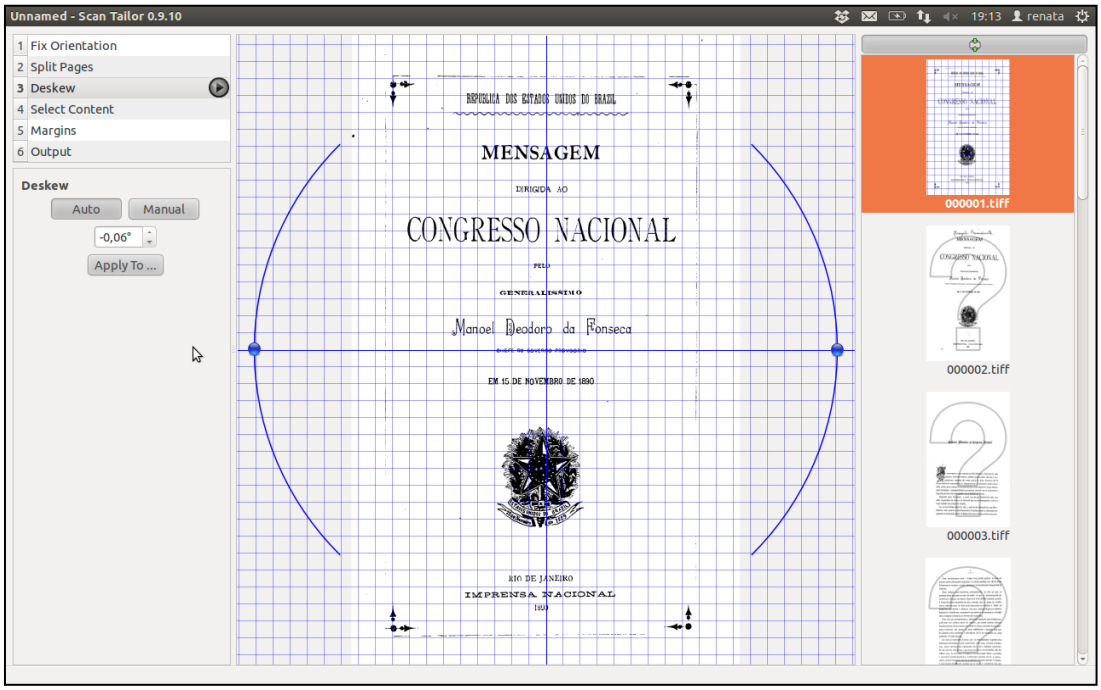

Figura 12 – *Scan Tailor* - Alinhamento
Fonte: ARTSIMOVICH (2012)

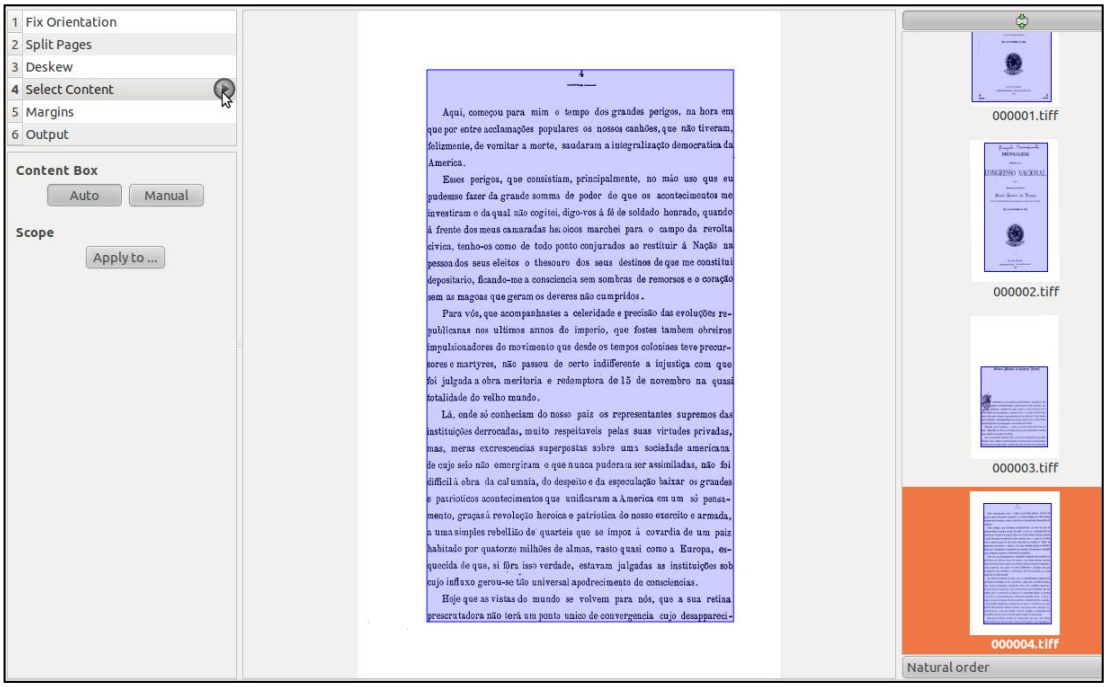

Figura 13 – *Scan Tailor* – Seleção de Conteúdo
Fonte: ARTSIMOVICH (2012)

A Figura 13 anteriormente vista, representa a etapa de “Seleção de Conteúdo”. Nesta etapa o usuário determina a área de conteúdo da imagem. É importante definir esta área, pois ela estabelecerá um centro para a construção das bordas, o que acaba influenciando no tamanho do documento de saída. Esta opção pode ser ativada no automático, assim, o programa seleciona sozinho a área de conteúdo para todas as páginas.

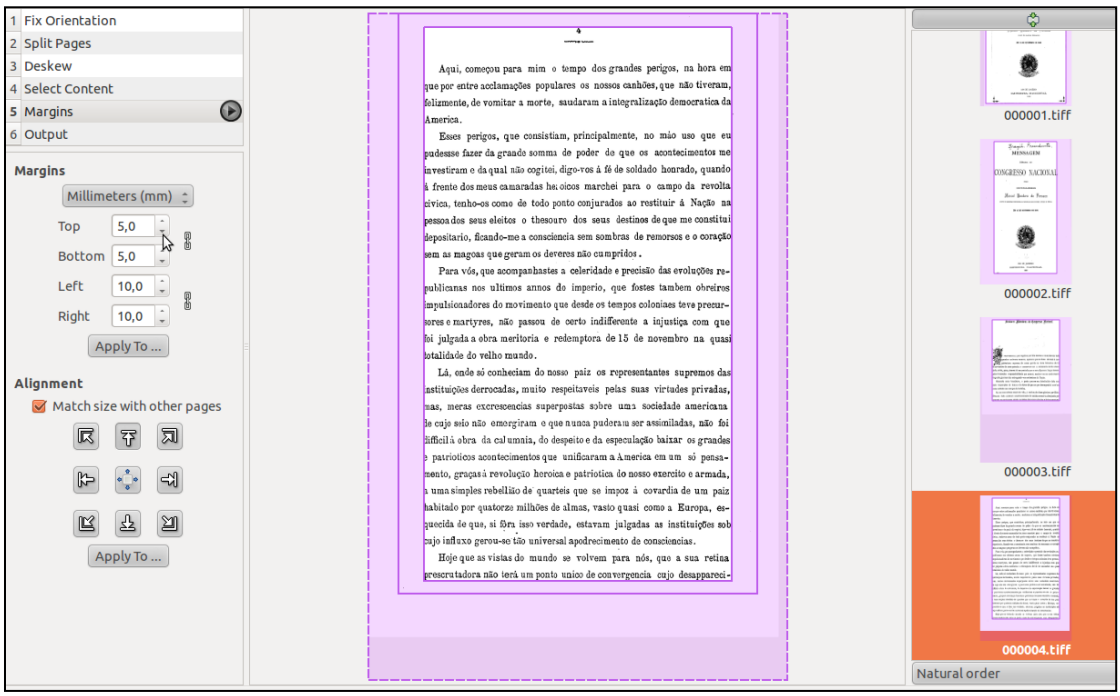

Figura 14 – *Scan Tailor* – Margens
Fonte: ARTSIMOVICH (2012)

A Figura 14 mostra a tela onde é possível configurar as margens e o alinhamento do documento digital. As margens podem ser editadas manualmente através de valores em milímetros. O alinhamento pode escolhido abaixo da opção da margem. A edição pode ser aplicada somente para o documento vigente ou para todos.

A Figura 15 a seguir apresenta a tela de saída. É possível escolher o modo preto e branco ou o modo escala de cinza para aplicação na imagem.

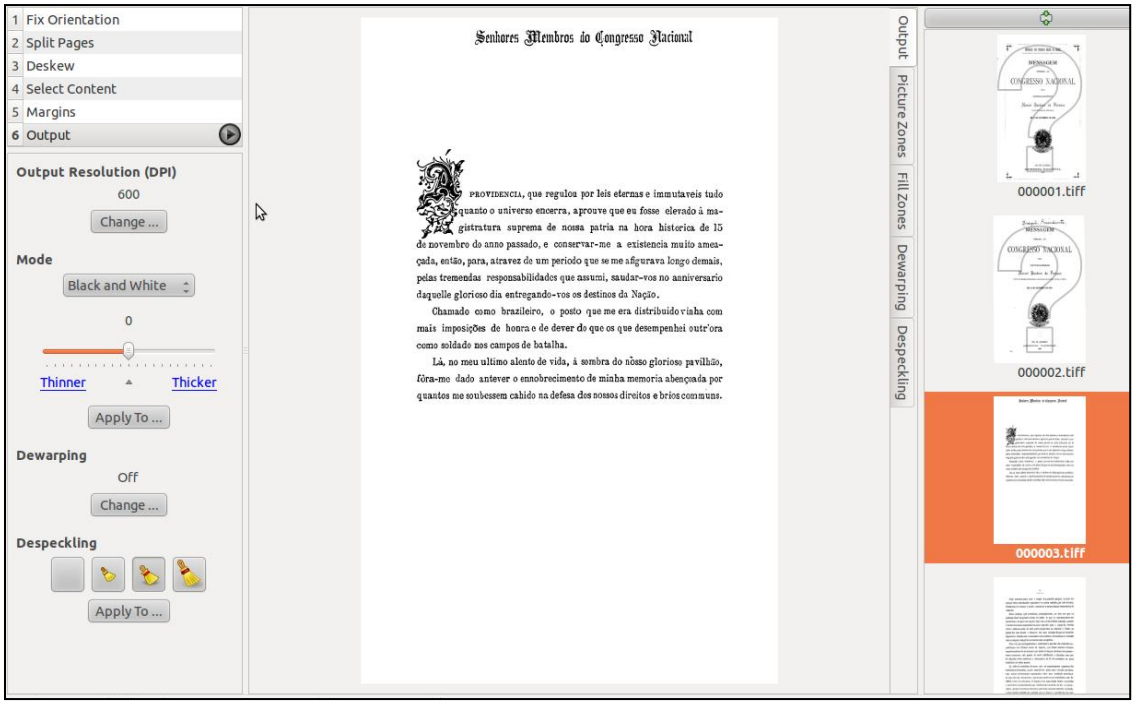

Figura 15 – *Scan Tailor* – Imagem de Saída
Fonte: ARTSIMOVICH (2012)

Os objetos digitais são salvos em 1 arquivo somente. Agora estão prontos para receberem a aplicação do OCR.

Quanto as imagens coloridas, conforme cita Gomes (2007), são formadas pela interposição de basicamente as três cores primárias, vermelho, azul e verde (ou imagem RGB – *Red, Green e Blue).* A diferença para as imagens na escala de cinza, é que aqui cada elemento é um vetor, e não um número escalar. Cada vetor representa um *pixel*, e cada *pixel* é formado pelos três elementos que armazena sua intensidade com base nas cores primárias – vermelho, verde e azul.

Vários são os problemas que devem ser corrigidos nas imagens, como por exemplo: não uniformidade na iluminação, inclinação em relação ao eixo, sujeiras que acabam aparecendo na imagem capturada, problemas no circuito da câmera ou do *scanner*, entre outros.

Gomes (2007) afirma que um procedimento que deve ser tomado como rotina na correção de qualquer imagem é a correção de fundo para depois atentar-se em corrigir os demais detalhes colocados em camadas posteriores na imagem.

O diferencial entre o tratamento de uma imagem em escala de cinza é que nesta as alterações podem ser feitas na imagem como um todo. Já na imagem colorida, as alterações devem ser feitas nas três matrizes primárias de cores: verde, vermelho e azul.

Para o tratamento de imagens coloridas existem diversos *softwares* no mercado, porém muitos sofrem limitações quanto as suas ferramentas ou quanto ao formato dos arquivos suportados. O *software* que ofereceu melhores ferramentas e melhor desempenho é o GIMP – GNU *Image Manipulation Program* na sua versão para Linux 2.8. O GIMP também possui versões para outros sistemas operacionais. GIMP é um *software* de uso profissional e livre, com uma *interface* completa e fácil de manipular que possui diversas funcionalidades, inclusive ele é passível de instalação de diversos *plug-ins* que os auxiliam em suas tarefas. Sua instalação pode ser feita pela Central de Programas do Ubuntu ou através do download e dos tutorias no site oficial do desenvolvedor (<http://www.gimp.org>).

A primeira coisa a se fazer ao abrir uma imagem, como Gomes (2007) recomenda, é editar seu fundo, corrigindo assim as falhas na iluminação decorrentes do processo de captura. As imagens também podem ser cortadas e as margens editadas para um melhor enquadramento. O *software* permite que sejam corrigidas as distorções, os efeitos negativos da não uniformidade da iluminação, o ajuste de contraste, de brilho, dos traços e a intensidade de cores, tudo isso contribui para melhorar o aspecto da foto. As imagens que contém somente textos devem ser bem preparadas para facilitar o reconhecimento de caracteres. A figura 16 mostra a *interface* básica do GIMP.

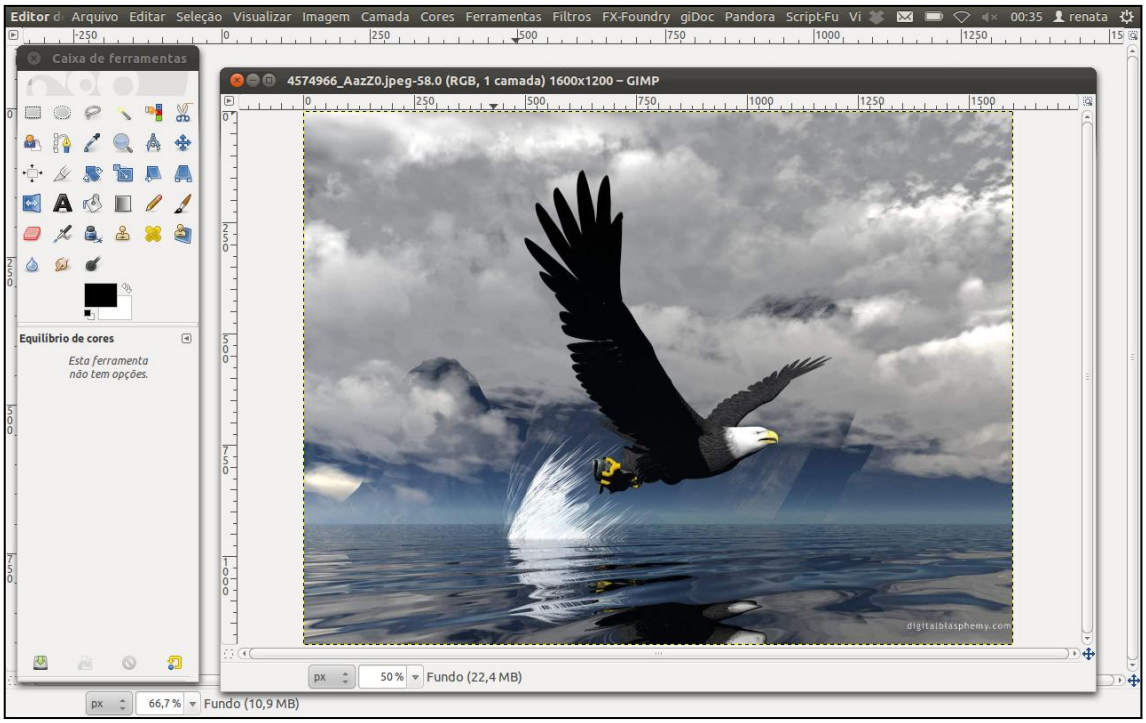

Figura 16 – GIMP - *Interface* Inicial
Fonte: NATTERER et al. (2012)

Conforme se pode verificar na Figura 17, o GIMP apresenta uma completa barra de ferramentas, o que permite que o usuário possa fazer vários ajustes em suas imagens. A seguir serão ilustrados alguns passos essenciais para a correção de imagens coloridas no GIMP.

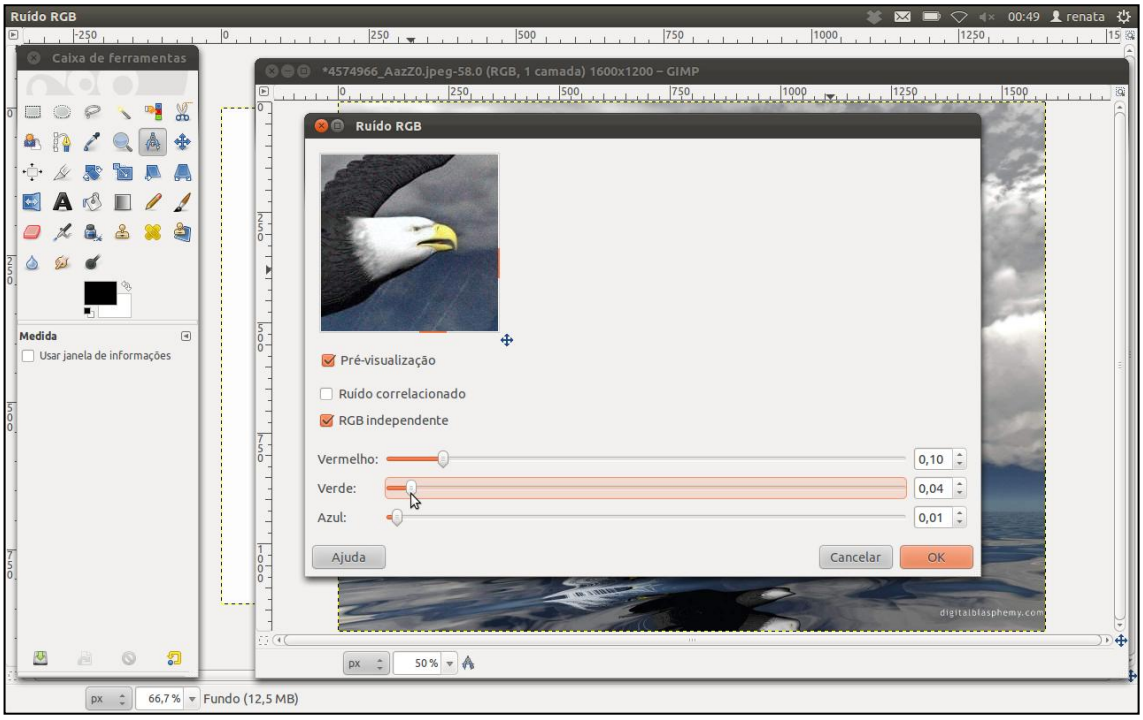

Figura 17 – GIMP – Correção de Ruídos
Fonte: NATTERER et al. (2012)

Como já exposto anteriormente, uma das primeiras coisas a serem feitas na imagem é a correção do fundo. É possível fazer a correção primeiramente de ruídos na imagem através da seleção da ferramenta na aba Filtro, em seguida selecionar a opção Ruído e por fim selecionar RGB. O RGB, como já foi visto, é a composição de cada pixel da imagem por três elementos primários, que são as cores vermelho, verde e azul. Na correção de ruídos é possível alterar a intensidade de cada um desses elementos, mudando a tonalidade do pixel, deixando-o mais uniforme e retirando suas imperfeições. A Figura 17 visualizada anteriormente mostra a interface para a realização desta tarefa. Conforme as alterações vão sendo feitas, uma pré-visualização é mostrada ao usuário.

O próximo passo é a correção da luminosidade, a qual também pode ser feita na aba Filtro, em seguida selecionar Efeitos de Iluminação. Pode ser realizada

a alteração do modo de luminosidade, ao lado aparece a pré-visualização. A figura 18 ilustra o procedimento.

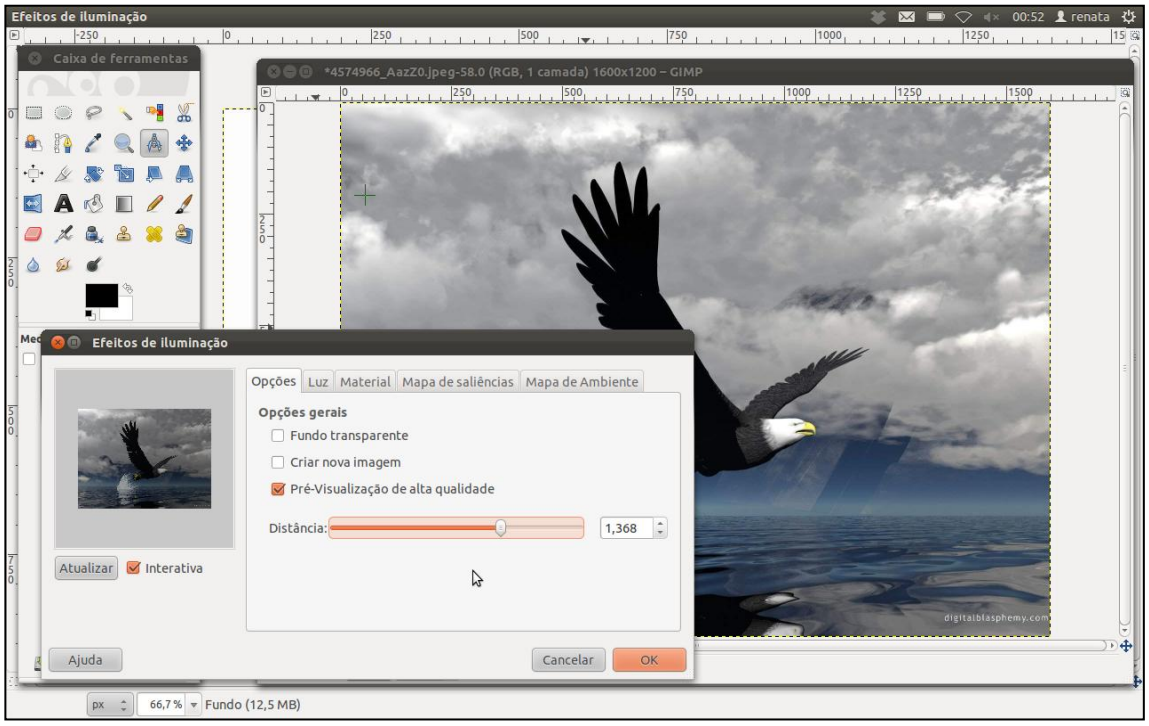

Figura 18 – GIMP – Efeitos de Iluminação
Fonte: NATTERER et al. (2012)

Na aba Filtro, ainda existem duas importantes ferramentas, a ferramenta de realce de cores e a de correção de distorções.

Para editar aspectos relacionados às cores da imagem, as ferramentas essenciais podem ser encontradas na barra chamada Cores, em seguida selecionar a opção Equilíbrio de Cores. Como pode ser verificado nas Figuras 19 e 20, elementos como brilho e contraste, saturação, equilíbrio de cores devem ser editados.

Quanto ao brilho e contraste, os níveis podem ser aumentados ou reduzidos e o resultado prévio é mostrado na imagem ao lado. Os procedimentos técnicos para a realização desta etapa foram demonstrados no início desta seção.

O equilíbrio de cores é muito importante, conforme Figura 20, é possível escolher a faixa que será alvo da edição, podendo ser: sombras, tons médios e tons claros. Os níveis de cores podem ser ajustados entre ciano, vermelho, magenta, verde, amarelo e azul. As modificações podem ser conferidas na pré-visualização exposta ao lado.

A Figura 21 demonstra a ferramenta de redimensionamento e aplicação de bordas e margens. Para isso, o usuário deve selecionar no menu a aba Ferramentas, em seguida, Ferramentas de Transformação, e por fim escolher as opções: redimensionamento, alinhamento, bordas, entre outras. No redimensionamento é possível escolher manualmente os novos valores.

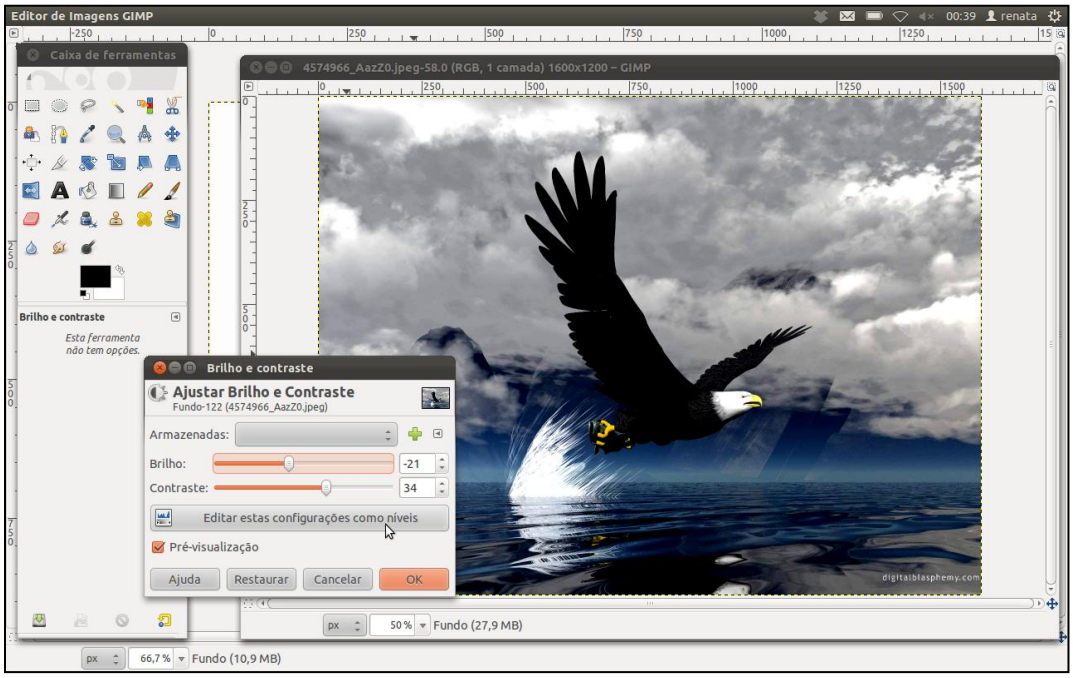

Figura 19 – GIMP – Brilho e Contraste
Fonte: NATTERER et al. (2012)

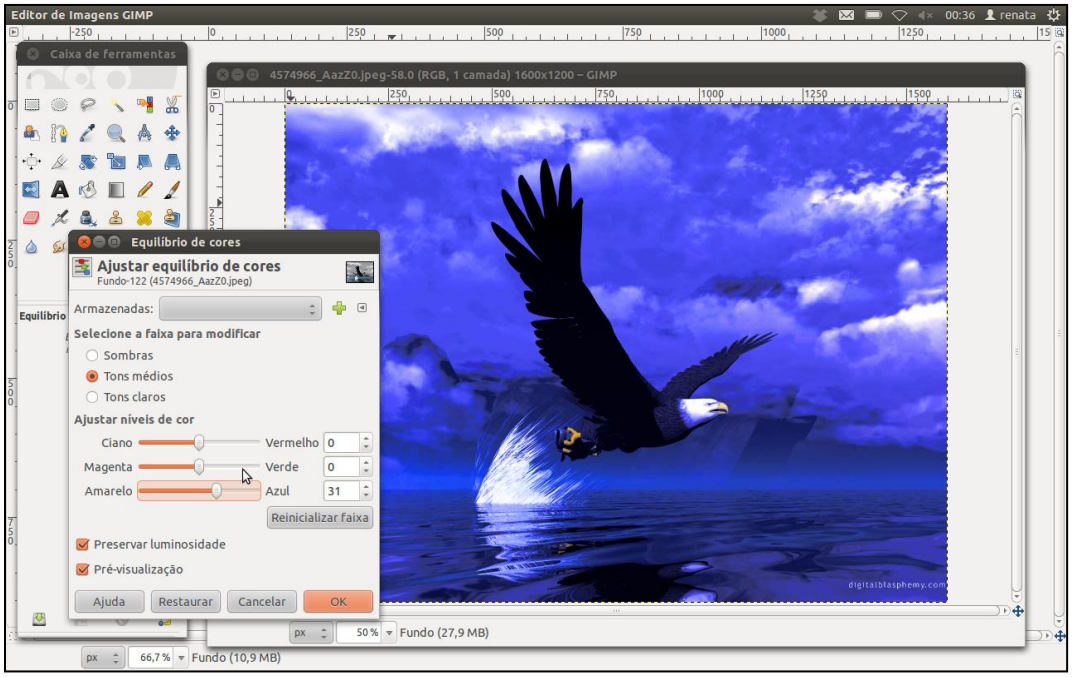

Figura 20 – GIMP – Equilíbrio de Cores
Fonte: NATTERER et al. (2012)

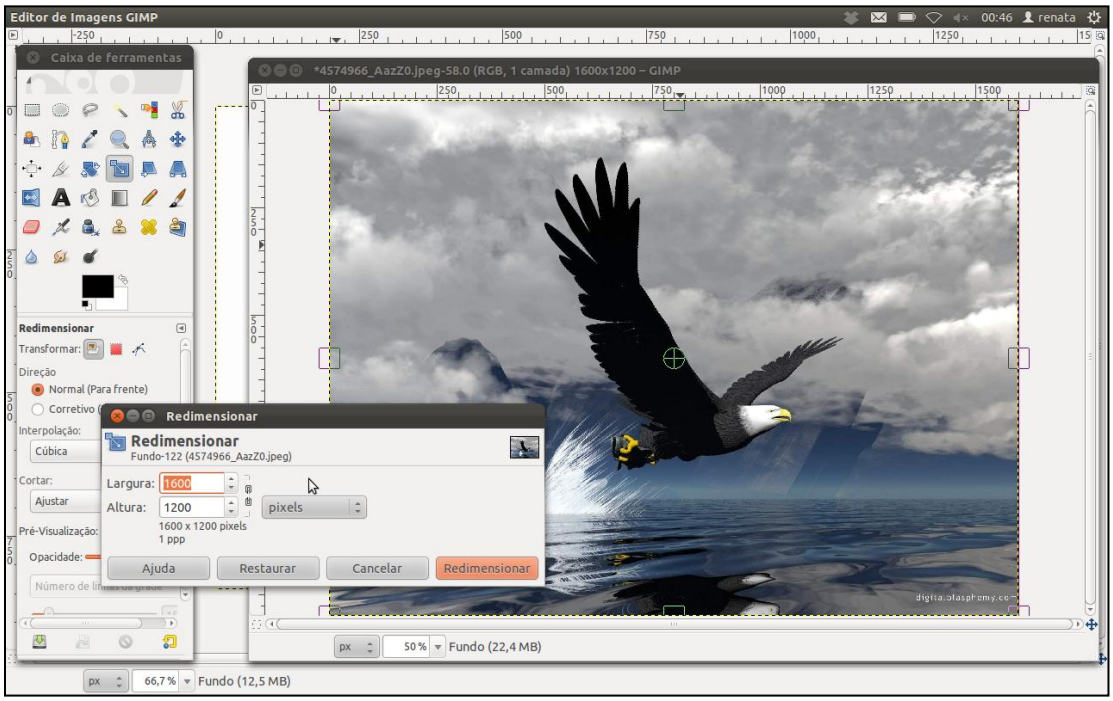

Figura 21 – GIMP – Redimensionamento
Fonte: NATTERER et al. (2012)

4.4 Reconhecimento Óptico de Caracteres (OCR)

O termo OCR é um acrônimo para o inglês *Optical Character Recognition.* De acordo com Alves (2003), “OCR’s são *softwares* capazes de extrair texto, caso estes existam, de imagens digitalizadas”. Ele “lê” páginas de imagens digitalizadas que contenham textos e extrai a camada de texto, a qual pode ser editada. O OCR não processa documentos que contenham somente imagens ou elementos gráficos.

A idéia de simular a leitura humana, como cita Osório (1991) surgiu em 1870 com o chamado *scanner* de retina inventado por Carey. Partindo disto, com a invenção do computador digital na década de 40, a idéia teve um grande impulso. Os primeiros OCR’s surgiram no ano de 1959 pela *Intelligent Machine Corporation*. Eles eram limitados a reconhecerem somente um tipo de fonte vindo de documentos específicos bancários. No decorrer do tempo a abrangência do reconhecimento dos caracteres foi aumentando e se aprimorando, ocasionando no aumento da utilização da ferramenta por empresas e organizações. “Em 1966, um padrão de fontes Americano, chamado OCR-A e um padrão Europeu chamado OCR-B foram desenvolvidos” (ALVES, 2003)

Na década de 1970 até 1980 a maioria dos OCR existentes eram baseados em padrões. Cada imagem de caractere era analisada e dela retirada todas as características dos traços, para assim comparar com a base de dados. Porém, textos muito claros acabavam sendo distorcido o que resultava num número muito baixo de acertos pela ferramenta. Com a evolução dos sistemas, na década de 1980 e 1990, o OCR já tinha a capacidade de analisar textos mais complexos e com misturas de fontes, como cita a autora: ”ao invés de utilizar padrões o sistema utilizava redes neurais, que são algoritmos capazes de aprender através de exemplos”. Atualmente são utilizados algoritmos especializados para o reconhecimento de caracteres, fazendo com que o OCR seja cada vez mais eficaz, conseguindo reconhecer até mesmo alguns caracteres danificados, (ALVES, 2003).

“Os sistemas de reconhecimento de caracteres podem ser desenvolvidos utilizando-se diferentes procedimentos tanto na aquisição dos dados como no processamento das informações.” (OSÓRIO, 1991). A Figura 22 mostra um esquema dos diferentes tipos de sistemas OCR:

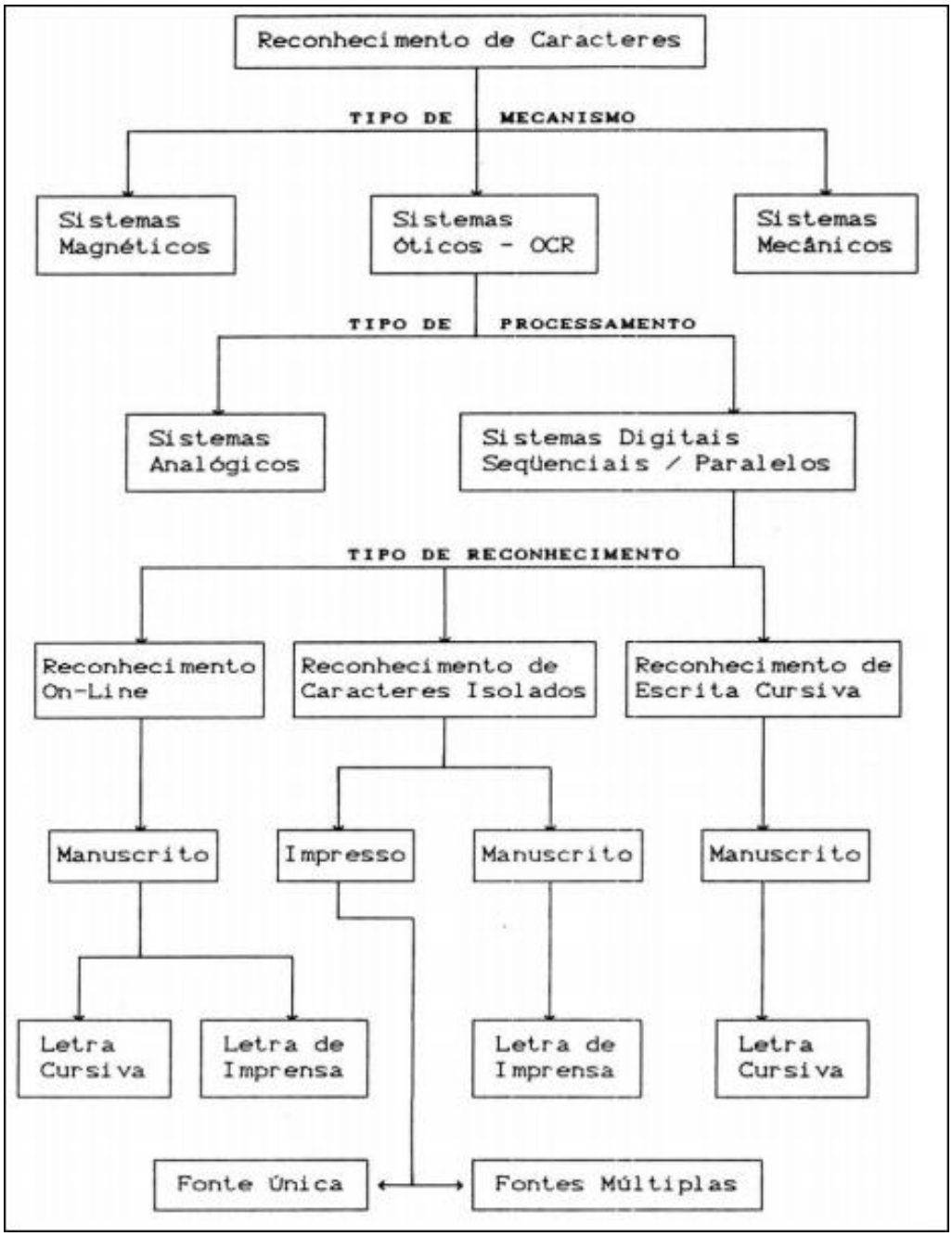

Figura 22 – Esquema de Tipos de Reconhecimento de Caracteres
Fonte: Osório (2001)

Este trabalho está focado no tipo de processamento do reconhecimento de caracteres dos Sistemas Digitais. Dentro desta categoria, pode-se observar que o reconhecimento pode ocorrer On-Line, Reconhecimento de Caracteres Isolados ou Reconhecimento de Escrita Cursiva.

Osório (1991) explica que, quanto ao reconhecimento On-Line possuem taxas de reconhecimento menores se comparado com os demais, sofrendo limitações quanto ao tamanho do arquivo ou a quantidade de reconhecimentos

diários. Eles ocorrem de forma interativa com o usuário e direto do navegador de internet Como exemplo de OCR's online pode ser citado o: *Document Conversion, Free-OCR, Google Docs*, entre outros.

Os sistemas de reconhecimento de caracteres isolados são os mais utilizados. Eles são ideais para textos digitados com letras misturadas ou não contínuas. O programa analisa cada caractere isoladamente, apresentando uma taxa de reconhecimento mais certa e como outro ponto positivo, este sistema é de fácil implementação (OSÓRIO, 1991).

O terceiro tipo de reconhecimento é o chamado: Reconhecimento de Escrita Cursiva. É avaliado como o mais complexo de todos os tipos de reconhecimento, considerando a dificuldade na implementação e a taxa de acertos. A complexidade é maior, devido ao fato de que cada pessoa tem estilos e características individuais em sua escrita e pelos caracteres não poderem ser separados, o reconhecimento torna-se mais complicado (OSORIO, 1991). Outro ponto que dificulta o reconhecimento de manuscritos é documentos que possuem escrita frente e verso, o OCR tem muita dificuldade em conseguir diferenciar cada lado. (ALVES, 2003).

O projeto está delimitado no Processamento de Sistemas Digitais, serão tratados adiante aspectos sobre o reconhecimento de escrita cursiva e o reconhecimento de caracteres isolados.

Quanto ao funcionamento básico de um OCR, a autora Alves (2003) faz um importante apontamento:

> O programa de OCR lê o bitmap (imagem) gerado pelo *scanner* e pondera as áreas de *pixel*s ativos e inativos da página, ou seja, mapeia o espaço em branco da página. Isso possibilita que o programa separe em blocos os parágrafos, colunas, títulos e partes gráficas. O espaço em branco entre as linhas de texto contidos num bloco define a base de cada linha, um detalhe essencial para o reconhecimento de caracteres no texto. (ALVES, 2003)

O programa de OCR guarda na "memória" as seqüências de *pixel*s que representam caracteres, podendo ser: letras, símbolos, números, pontuação, entre outros. Ao receber uma imagem, o OCR reconhece linha por linha, mapeando os *pixel*s ativos e comparando com sua memória interna.

Alves (2003) explica que o OCR "calcula a altura das letras do texto e analisa cada combinação das linhas retas, curvas e áreas preenchidas de cada caractere". Assim é possível reconhecer caracteres que possuem *pixel*s abaixo da linha base, como por exemplo, um "g" ou um "p".

O OCR é uma ferramenta que fornece a base para o reconhecimento ótico dos caracteres, mas o que realmente opera sobre o resultado é o chamado motor de OCR. Existem, atualmente, muitos motores que podem ser utilizados para o OCR, como por exemplo: *Ocropus, Cuneiform, Tesseract, Gocr, Ocrad,* entre outros. Cada motor OCR possui suas particularidades, o *Tesseract* é o mais popular, porém ele não faz reconhecimento de caracteres, requisito que o *Ocropus* faz. As maiorias dos OCR's apresentam uma alta taxa de erro e de rejeição. A rejeição difere-se do erro, porque nela, o caractere não é reconhecido e também não é confundido com nenhum outro. Para o reconhecimento de caracteres mais complexos, pode-se fazer uso de tecnologias mais inteligentes como redes neurais por exemplo.

O motor utilizado neste projeto será o *Ocropus*. Ele foi inicialmente escrito na linguagem C++, e, conforme foi se desenvolvendo, surgiram *script*s cada vez mais complexos, o que fez com que fosse migrado para a linguagem *Python*. Portanto, desde a versão 0.4, a "linguagem *script* principal para *Ocropus* é *Python*. (*OCROPUS*, 2012).

Como cita Breuel (2006), a arquitetura do OCR *Ocropus* contém três componentes principais, que são: análise do reconhecimento de linha de texto e linguagem de modelagem estática. A análise do (que é um diferencial deste motor) permite que o OCR detecte a disposição física do documento, identificando colunas, blocos de texto, imagens, e lendo-os todos em ordem. Este elemento separa o documento por regiões. O reconhecimento de linha de texto permite que o OCR identifique linhas que contenham textos e as leia. A modelagem estática, como cita o autor "permite o reconhecimento alternativos com conhecimento prévio sobre vocabulário, linguagem, gramática e domínio do documento". A modelagem estática então resulta na interpretação pelo próprio OCR quando estiver em casos de dúvidas ou ambigüidade quanto a algum caractere.

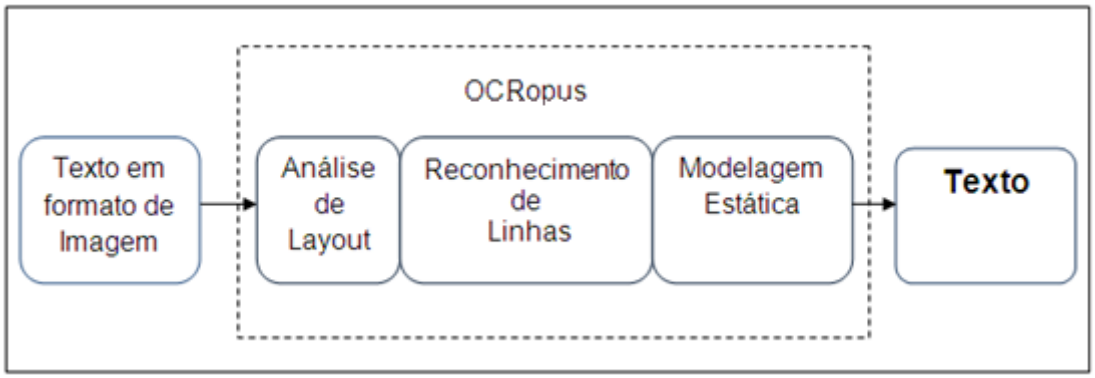

Figura 23 – Mecanismo *Ocropus*
Fonte: Autoria Própria

O *Ocropus* 0.4.4 foi instalado no sistema Ubuntu a partir do tutorial disponibilizado no site do desenvolvedor do *Ocropus*, disponível em: < *http://code.google.com/p/Ocropus/wiki/InstallTranscript>.* A instalação é feita por linha de comando. Junto com ele, devem ser baixados e posteriormente instalados os demais componentes: *iulib*, *openfst*, *ocroswig*, *ocropy*, ligações com *python*.

De acordo com o site do próprio desenvolvedor *OCROPUS* (2012), *Iulib* "fornece processamento de imagem básica, morfologia matemática, e algoritmos de aprendizagem de máquina". *Iulib* é uma biblioteca de estrutura de dados que facilita a implementação e o uso de funções e processos de entrada e saída relacionados com a utilização de imagens e vídeos. *Ocroswig* é uma extensão para *scripts* no *Ocropus*.

O *openfst* também é uma biblioteca que contém aplicações no "campo de reconhecimento de fala, tradução automática, reconhecimento ótico de caracteres, correspondência de padrões, processamento de string, aprendizado de máquina e extração e recuperação de informação". (OPENFST, 2012)

Oc*ropy* é o pacote onde estão alocados todos os *Ocropus* C + +, classes Python e suas funções. (*OCROPUS*, 2012).

O programa utilizado para a aplicação do *Ocropus* é o Gscan2pdf, *software* livre que também foi instalado pela central de programas do Ubuntu, porém, deve ser instalado juntamente com o *software*, os pacotes para a utilização do OCR (Os pacotes podem ser baixados através do *Synaptic* do Ubuntu). O *software* permite que as imagens possam ser importadas diretamente do equipamento de captura ou de um diretório qualquer. O Gscan2pdf compacta vários arquivos em um só, ideal para o caso de digitalização de livros, onde todas as páginas serão reunidas formando um arquivo único. O formato de arquivo de saída do Gscan2pdf mais

utilizado é o .pdf e o .djvu. No caso do projeto em questão, o formato padrão escolhido para os arquivos é o .djvu. De acordo com DJVU (2012) “Djvu (pronuncia-se "déjà vu") é um formato de documento digital, com tecnologia de compressão avançada e valor alto desempenho”. O formato Djvu foi desenvolvido com foco na utilização para web, pois tem uma redução considerável no tamanho do arquivo, o que facilita o *download* e o *upload* dos objetos digitais. “Um arquivo normal de 40MB, se transformado para o formato Djvu, pode reduzir seu tamanho para 8MB. (DJVU, 2012). Uma das principais características desta extensão é a possibilidade de gerar camadas de imagem e de texto acopladas, possibilitando a busca textual em imagens.

Muitos OCR’s e *softwares* ainda não trabalham com a extensão djvu. Por isso, outro motivo para a escolha do Gscan2pdf, o qual compacta, insere a camada de texto e disponibiliza o arquivo em formato Djvu. A figura 24 mostra como é a *interface* inicial do *software* Gscan2pdf. A coluna no lado esquerdo apresenta todas as páginas que formam o documento. As edições podem ser aplicadas somente na página em questão como em todas as demais. Ao salvar o documento pode-se escolher o formato (pdf, txt, djvu...) e então será gerado um único arquivo.

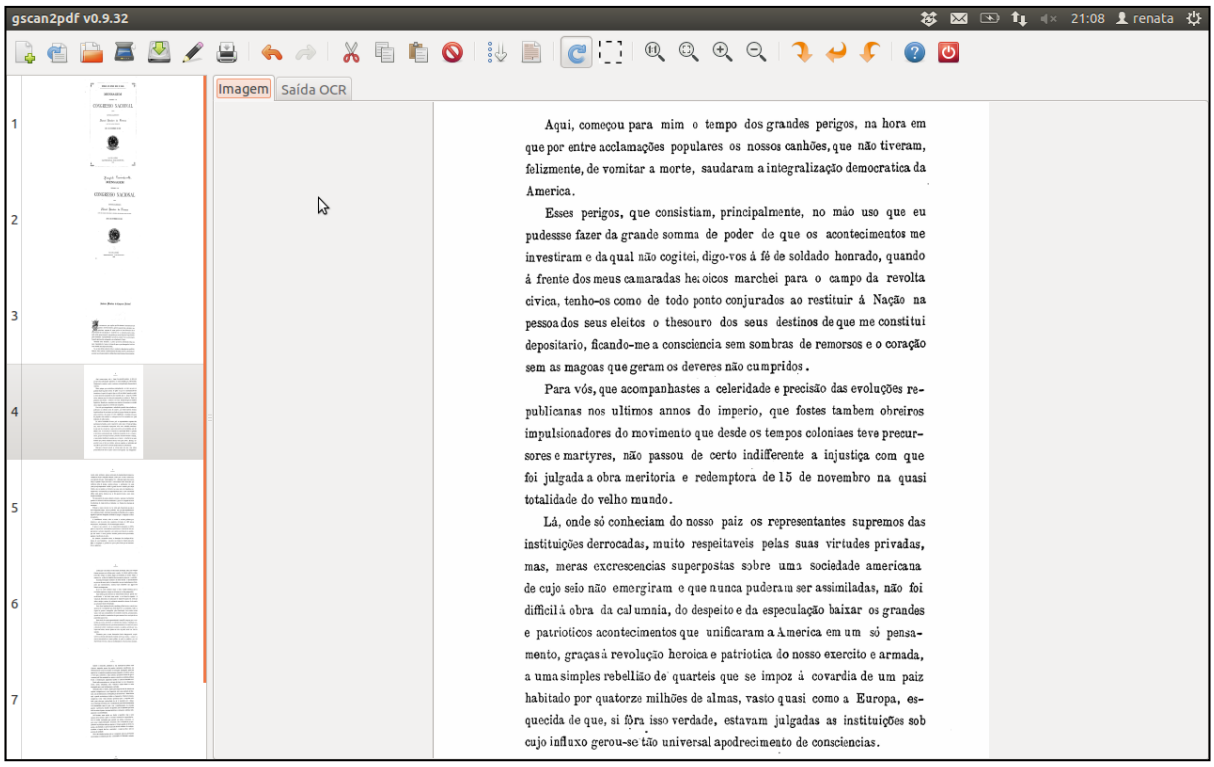

Figura 24 - *Gscan2pdf – Interface*
Fonte: RATCLIFFE (2012)

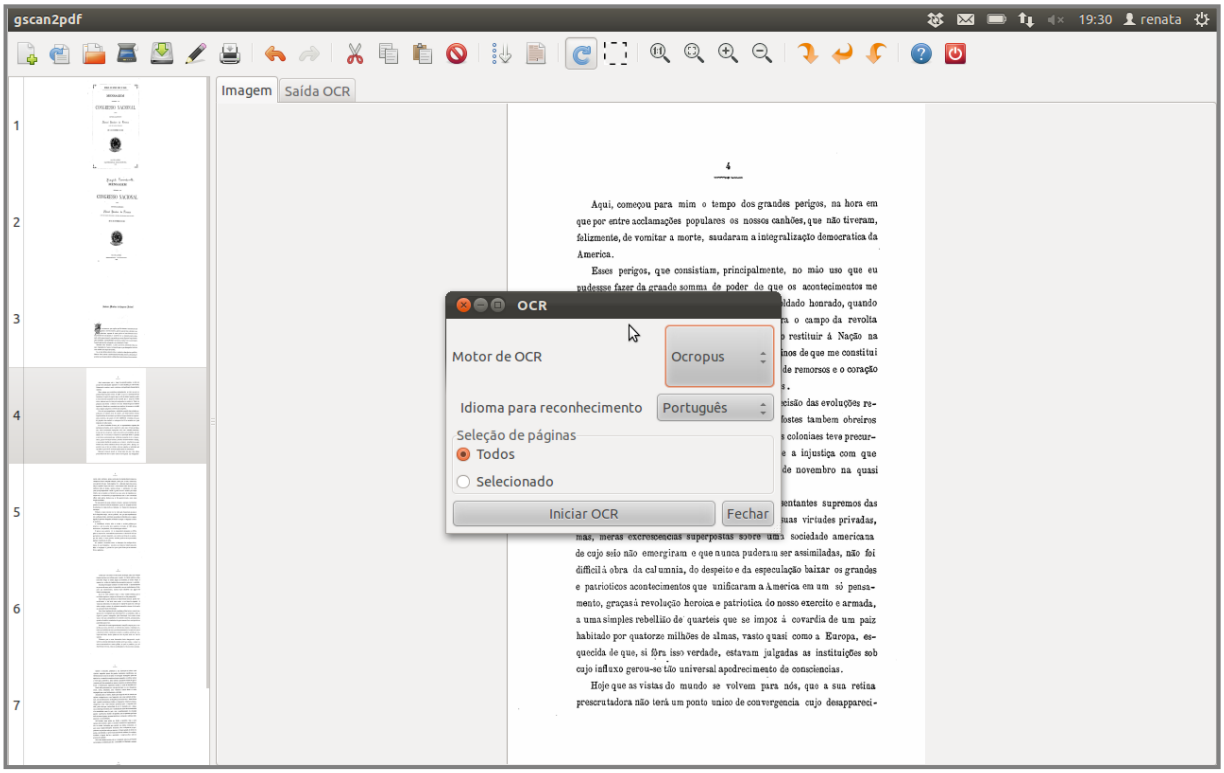

Figura 25 – *Gscan2pdf* – Aplicando OCR
Fonte: RATCLIFFE (2012)

Na figura 25 foi selecionada a ferramenta para aplicação do reconhecimento ótico dos caracteres. Dependendo de qual foram os motores instalados juntos com o *software*, através do gerenciamento de pacotes *Synaptic* do Ubuntu, os mesmos aparecerão como opções em “Motor de OCR”, para que o usuário escolha o mais conveniente. Em seguida é possível selecionar o idioma no qual o OCR irá trabalhar e em quais páginas será feita a aplicação do OCR.

A figura 26 a seguir, mostra a aba gerada após o reconhecimento do OCR. Como foi utilizado o *Ocropus*, podemos perceber que as linhas estão separadas, permanecendo o do documento. Esse é o mecanismo de trabalho do *Ocropus* como já foi visto anteriormente. Outros OCR que não possuem reconhecimento de acabam reconhecendo os caracteres e formam um grupo amontoado de palavras, perdendo a disposição original das mesmas.

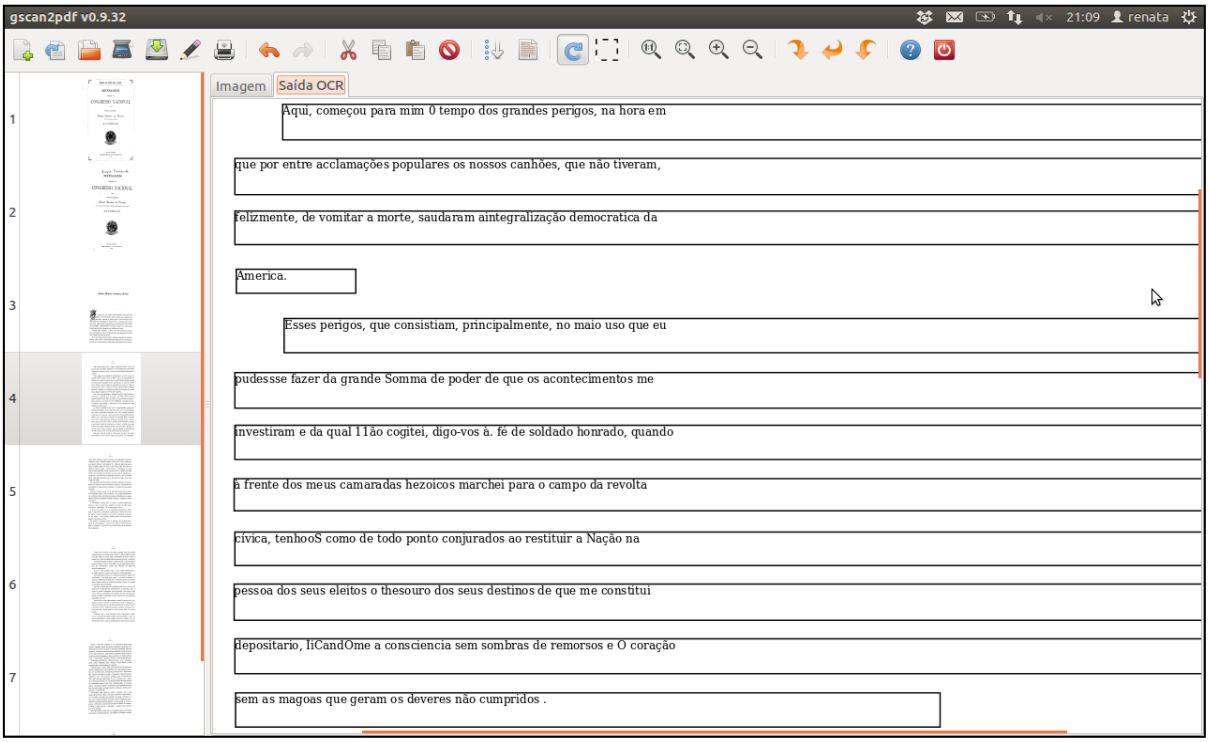

Figura 26 – *Gscan2pdf* - Tela OCR
Fonte: RATCLIFFE(2012)

Para melhores resultados do OCR o mesmo deve passar pelas fases de treinamento, quanto mais bem treinado, melhores resultados de reconhecimento. Esta é uma dificuldade encontrada, pois não foi encontrada nenhuma *interface* gráfica que permita o treinamento do OCRopus, outro ponto que dificulta é a falta de fontes confiáveis para pesquisa. Quase não há documentação acerca do treinamento do OCRopus.

4.5 Repositório para Armazenamento

Um dos principais objetivos da digitalização dos documentos, além da conservação das versões originais, é a disponibilização facilitada para pesquisas e consultas do público. A última fase do processo de construção de objetos digitais é o armazenamento confiável dos mesmos.

Martins et al. (2012) dispõe que “repositórios digitais são coleções de informação digital, que podem ser construídas de diferentes formas e com diferentes propósitos.”

Os repositórios podem ser desde forma simples como uma enciclopédia online categoria que recebe o nome de repositórios digitais, quanto mais complexos chamados de repositórios institucionais, que seguem termos de descrição de

arquivologia disponibilizando arquivos oficiais como, por exemplo, do governo. (MARTINS et al., 2012)

De acordo com Thomaz (2007), as empresas ou organizações “podem optar por desenvolver seu próprio repositório, para organização e acesso local ou contratar um provedor de serviço de preservação digital”. Para a autora um repositório digital não compreende somente o armazenamento e a administração dos objetos digitais, mas sim “que tenha a missão de fornecer acesso confiável, por longo prazo, aos recursos digitais administrados”. Os arquivos digitais são armazenados e manipulados, sendo possível sua recuperação e pesquisas posteriores, o repositório, portanto, deve ser sustentável e fiável.

Os repositórios digitais se diferenciam das bases de dados por algumas características, as quais citam Martins et al. (2012): os objetos digitais podem ser depositados pelo autor, proprietário ou terceiro, a arquitetura do repositório gera tanto o conteúdo quanto os seus possíveis meta dados (coisa que o banco de dados não faz) e, como última característica diferenciadora, o repositório oferece serviços básicos mínimos como controlar acesso, depositar, editar, pesquisar, entre outros.

> De fato, além das plataformas comerciais para repositórios gerais (como *Digital Commons*) ou das plataformas especializadas comerciais para repositórios de objetos de aprendizagem (como o *Blackboard Content System, Desire2Learn or The Learning Edge*) há diversas plataformas *Open Source* disponíveis. (MARTINS et al.,2012)

Algumas plataformas livres disponíveis no mercado são: *DSpace, Omeka, Ica-Atom, GreenStone e Archon.* Cada repositório possui suas particularidades. O projeto foi executado com a adoção do repositório Ica-Atom, por ser o repositório mais completo, estando de acordo com as regras internacionais de descrição arquivística.

Conforme ICA-ATOM (2012) a sigla ICA-AtoM significa, em português, “ICA”-Conselho Internacional de Arquivos e “ATOM” acesso à Memória. É um aplicativo com código aberto e fundamentado em ambiente web, baseado em padrões para a descrição arquivística suportando várias linguagens. É um ambiente multiarquivos e multirepositório, pois com o ICA-AtoM possibilita não somente que uma única instituição armazene seus arquivos, mas também que diversas instituições arquivísticas reúnam seus acervos.

O servidor utilizado é o *Apache*. Todas as interações do usuário com o sistema como criar, visualizar, pesquisar, atualizar e excluir ações, ocorrem através

do navegador web do usuário. Os usuários têm acesso a páginas HTML que ficam armazenadas no servidor web. Ao clicar em um botão ou link desencadeia um *script* PHP que envia um comando ao banco de dados e retorna o resultado como HTML voltando ao browser do usuário. (ICA-ATOM, 2012)

A base de dados utilizada é a *MySQL*, porém, conforme apresentado na página oficial do desenvolvedor ele “utiliza uma camada de abstração de dados e, assim sendo também compatível com *PostgreSQL, SQLite, SQLServer, Oracle*, entre outros”. O código PHP que gerencia os pedidos e respostas entre os clientes de internet A aplicação lógica e a aplicação de conteúdos estão armazenadas na base de dados symfony (*framework* que serve de biblioteca para aplicações mais complexas da web). O ICA-AtoM também utiliza o *Qubit*, *Open Information Management Toolkit*, desenvolvido pelo projeto ICA-AtoM e personalizado para desenvolver o aplicativo.

A Figura 27 mostra um esquema da arquitetura que envolve o repositório Ica- Atom.

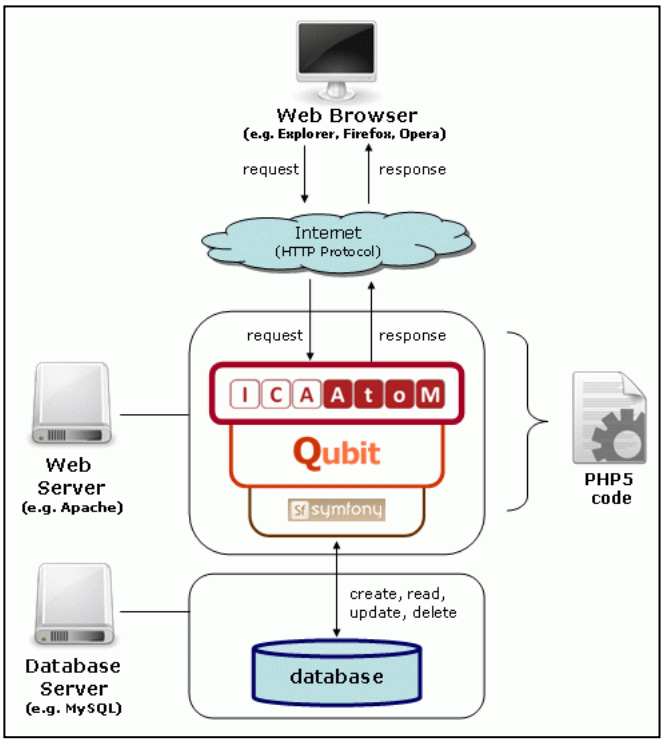

Figura 27 – Arquitetura *Ica-Atom*
Fonte: *ICA-AtoM* (2012)

De acordo com a própria documentação do desenvolvedor Ica-Atom (2012), a aplicação trabalha com as chamadas entidades, que são objetos com as quais o

sistema interage para organização dos dados. Existem quatro tipos de entidades: Descrição Arquivística, Autoridade de registros (atores), Instituições de arquivo e Termos. A descrição arquivística “fornece informação contextual sobre materiais de arquivos.” A entidade Autoridade de Registro fornece descrições dos atores que interagem com materiais de arquivos como criadores, custodiantes, entre outros, podendo estes serem entidades, pessoa ou famílias, eles estão ligados as respectivas descrições arquivísticas pelos chamados evento. As instituições de arquivos fornecem descrições de repositórios que preservam e provêem acesso a materiais de arquivo. E por último, a entidade Termos controla o vocabulário utilizado em todo o sistema.

O ICA-AtoM trabalha com permissões, as quais são concedidas dependendo da função que o usuário representa no sistema. Cada usuário pode ter uma ou mais funções. As funções compreendem em: pesquisador, contribuinte, editor, tradutor e administrador. O pesquisador é o usuário que interage fora do sistema, ele tem somente a permissão de visualização. O usuário contribuinte tem a permissão de pesquisar, navegar, criar, editar e exportar descrições. O Editor pode pesquisar, navegar, criar, editar, publicar, apagar e exportar descrições e editar vocabulário controlado termos. O tradutor pode pesquisar e navegar descrições e pode traduzir *interface* de usuário elementos e conteúdo do banco de dados. E por fim, o Administrador pode importar, exportar, criar, ler, atualizar, publicar e excluir qualquer registro no sistema, pode personalizar a aplicação às necessidades instituição específica. Somente ele pode gerenciar contas de usuário e perfis, atribuindo ou delimitando as funções dos usuários. (ICA-ATOM, 2012)

Cada objeto digital possui seu nível de descrição. Os níveis de descrição estão dispostos de forma hierarquizada, podendo ser: coleção, fundo, item, série ou subsérie. O usuário pode fazer upload de qualquer objeto digital, como exemplo: imagens digitalizadas, arquivos de som, arquivos de vídeo, entre outros. Cada objeto deve estar associado a uma descrição arquivística. A Figura 28 apresenta a *interface* inicial do ICA-AtoM.

Figura 28 – *ICA-AtoM - Interface* Inicial
Fonte: *ICA-AtoM* (2012)

Figura 29 – *ICA- AtoM* - Visualização do Objeto Digital
Fonte: *ICA-AtoM* (2012)

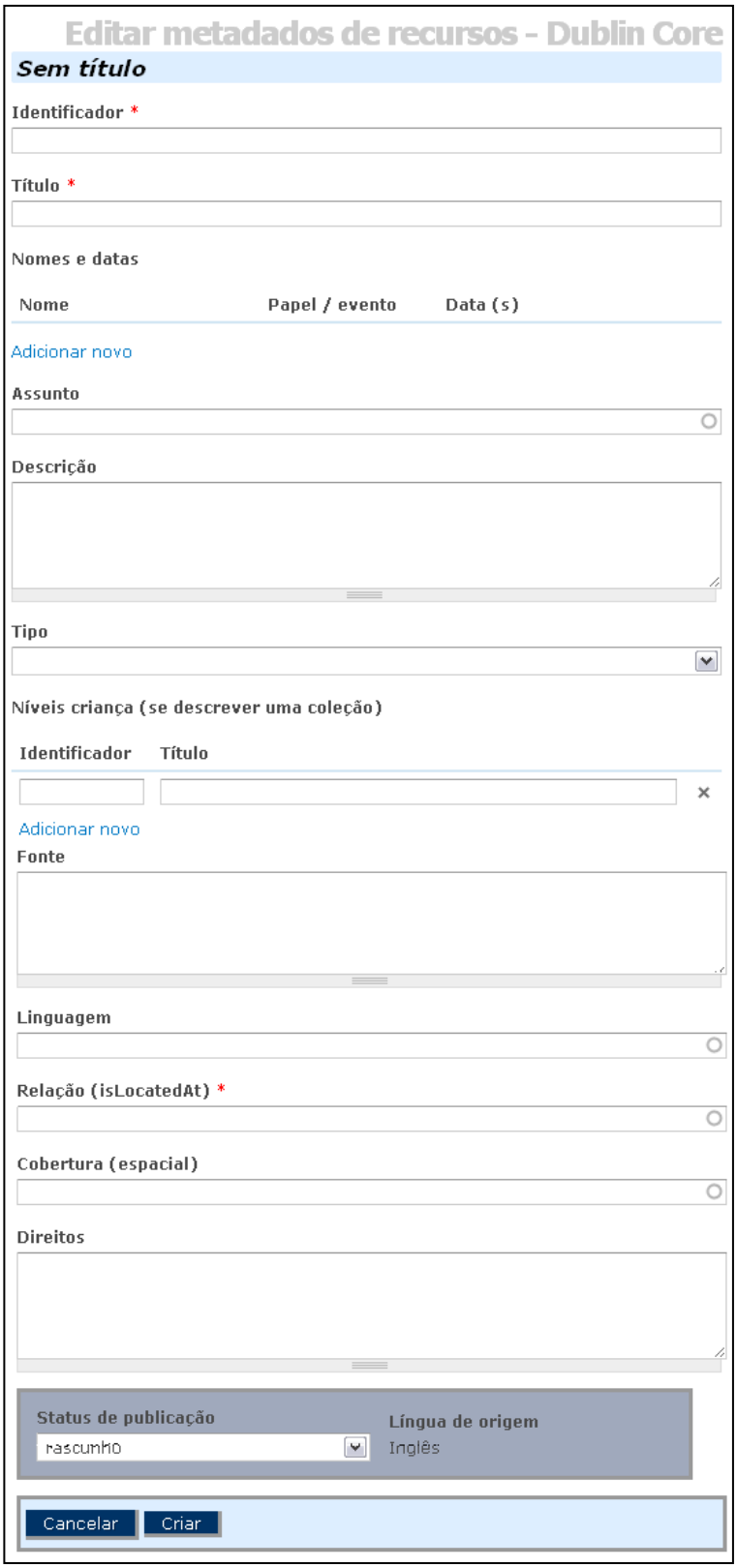

Figura 30 – *ICA-AtoM* – Nova Descrição Arquivística
Fonte: *ICA-AtoM* (2012)

A Figura 30 demonstra o formulário que o usuário deve preencher para adicionar uma nova descrição arquivística, assim é possível fazer a ligação com o objeto digital. Para realizar tal procedimento, o usuário deve ir ao menu Adicionar e selecionar a opção Descrição de Arquivo. Para criar uma nova Instituição Arquivística

ou uma nova autoridade de registro, o procedimento é o mesmo, o que mudos são os campos que devem ser preenchidos.

Figura 31 – *ICA-AtoM* - Menu
Fonte: *ICA-AtoM* (2012)

A Figura 31 apresenta a tela de visualização do objeto digital quanto o usuário está logado no sistema. Dependendo das suas permissões os botões no menu ficam disponíveis ou não. Por exemplo, nesta Figura o usuário tinha todas as permissões, podendo editar, deletar, anexar mais objetos digitais, adicionar novo registro, etc.

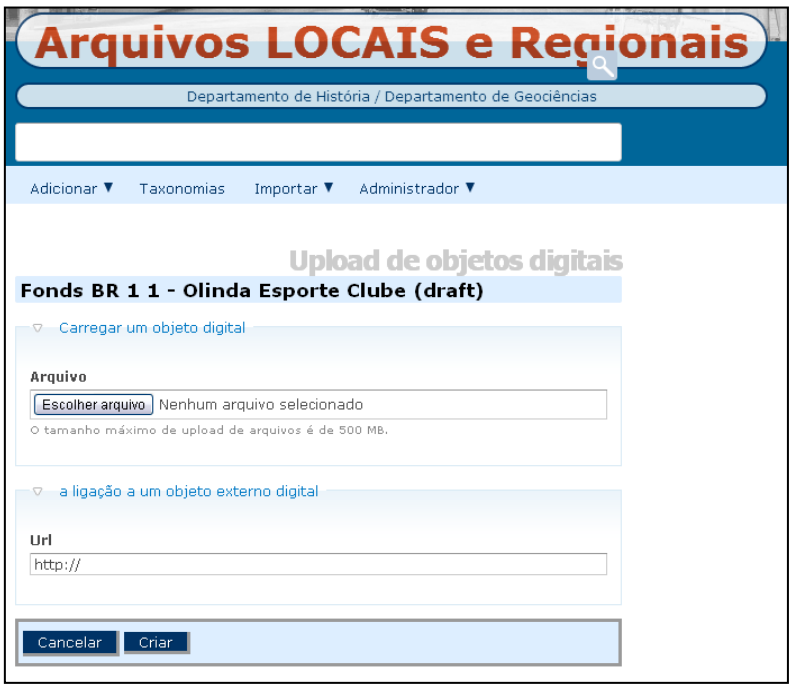

Figura 32 – *ICA-AtoM* – Ligação com um Objeto Digital
Fonte: *ICA-AtoM* (2012)

A Figura 32 representa a tela onde é possível fazer ligação com um objeto digital. Há um campo onde o usuário escolhe o caminho onde se encontra o objeto digital desejado e carrega, assim, ele ficará ligado a sua respectiva descrição arquivística.

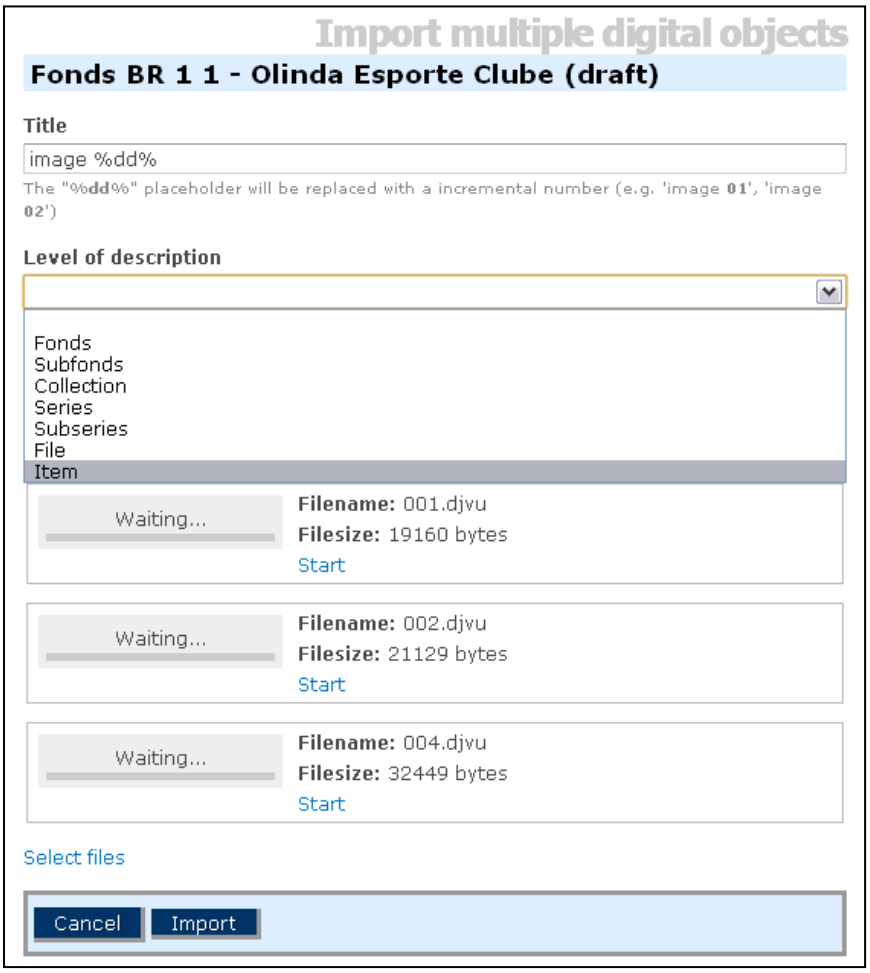

Figura 33 – ICA-AtoM – *Upload* de vários Objetos Digitais
Fonte: *ICA-AtoM* (2012)

Na Figura 33, o usuário pode carregar uma série de objetos digitais ao mesmo tempo. Os objetos digitais devem ser carregados na extensão djvu, pois, como já abordado anteriormente, este formato possui a mesma qualidade e com um tamanho significativamente menor em relação aos outros formatos, o que facilita o *download* e *upload* dos arquivos.

4.6 Modelo do *Workflow*

De acordo com Cruz (2000), os três elementos essenciais para a construção de um *workflow* são os chamados 3 r's, que do inglês significam: *roles* (papéis), *rules* (regras) e *rotes* (rotas), a partir disso, para ocorrer à implantação de um *workflow*, deve-se seguir um ciclo lógico, conforme mostrado na Figura 34.

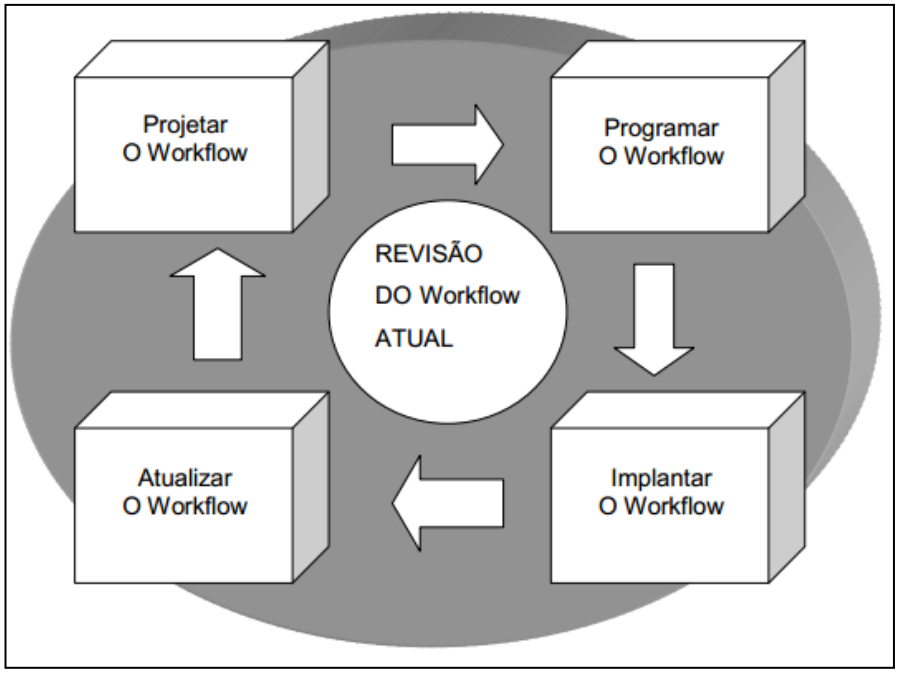

Figura 34 – Ciclo para Implantação do *Workflow*
Fonte: CRUZ (2000)

Cruz (2000) estabelece que antes de projetar o *workflow*, é necessária uma análise de como as atividades já estão sendo executadas, em seguida estipular as possíveis melhorias que podem ocorrer e enfim elaborar um novo modelo de processo. Cruz (2000) apresenta um roteiro para facilitar esta etapa, neste roteiro alguns pontos devem ser respondidos como, por exemplo:

1. Condições atuais do processo
2. Programar os tempos máximos e mínimos das atividades
3. Estabelecer pré-atividades para a pessoa
4. Descrever a execução, ferramentas, metodologias e técnicas para a execução de um item.
5. Descrever como as pessoas serão notificadas sobre um evento dentro do processo.
6. Descrever as pós-atividades

7. Descrever as autorizações dos membros participantes por questões de segurança.

8. Descrever os eventos e em que momento passarão por auditoria.

9. Descrever os casos em que o item terminará ou sofrerá interrupção.

Para que o *Workflow* possa ser programado, desde o começo devem ser bem estipulados requisitos como: o objetivo geral do procedimento, os participantes do *workflow* e qual papel representarão, a rota de informações e de documentos e formulários dentro do processo. (CRUZ, 2000)

Ainda de acordo com CRUZ (2000), de um modo adaptado, a estrutura de um *software* de *workflow* deve conter os elementos descritos a seguir:

- Definir a estrutura organizacional
- Definir o *design* do processo:
    - Criar as atividades
    - Desenhar o Fluxo do Processo
    - Definir as propriedades para cada Processo
    - Definir a estrutura de dados
    - Definir as Regras de negócio
    - Testar
- Carregar e Monitorar o processo por meio do Administrador
- Executar o *Workflow*

A terceira fase como na Figura 34 representa a implantação do *workflow*. Esta fase não será executada neste trabalho, aqui somente será estruturado o modelo de *workflow*, ficando a execução efetiva para trabalhos futuros.

É interessante lembrar, de acordo com Cruz (2004) que os vários processos da empresa não devem ser implantados de uma vez, e sim, começando do menor processo. Isto porque se deve deixar um tempo para que as pessoas que colaborarão com o *workflow* se acostumem. Existem dois tipos de pessoas que interagem com o *workflow*, os responsáveis pela implantação do *software* e os usuários, ambos devem ser treinados.

A quarta e última fase da Figura 34 – Ciclo para implantação de *Workflow* é a de atualização do *workflow*. É possível através de ferramentas disponibilizadas pelo próprio sistema que seja feita uma análise de desempenho das atividades, com esses demonstrativos é possível que se pense em melhorias que podem ser programadas e atualizadas no *workflow*. As melhorias, como estabelece Cruz

(2000), podem ser de ordem quantitativa (como em casos de fluxo de papéis, formulários, tempo de espera), como de ordem qualitativa (como o desempenho dos participantes no processo, as atividades executadas paralelamente, entre outros). A fase de atualização é uma das mais importantes, pois é nela que se tem o *feedback*, o que possibilita que o *workflow* seja cada vez mais otimizado.

O projeto se baseia na construção de um modelo de *workflow* que englobe todas as etapas pertencentes ao processo de digitalização de documentos. O tipo de *workflow* que se enquadra neste projeto é o *Workflow Ad Hoc,* as demais classificações podem ser visualizadas no capítulo 2 deste trabalho. Primeiramente será feita uma análise do fluxo de trabalho atual:

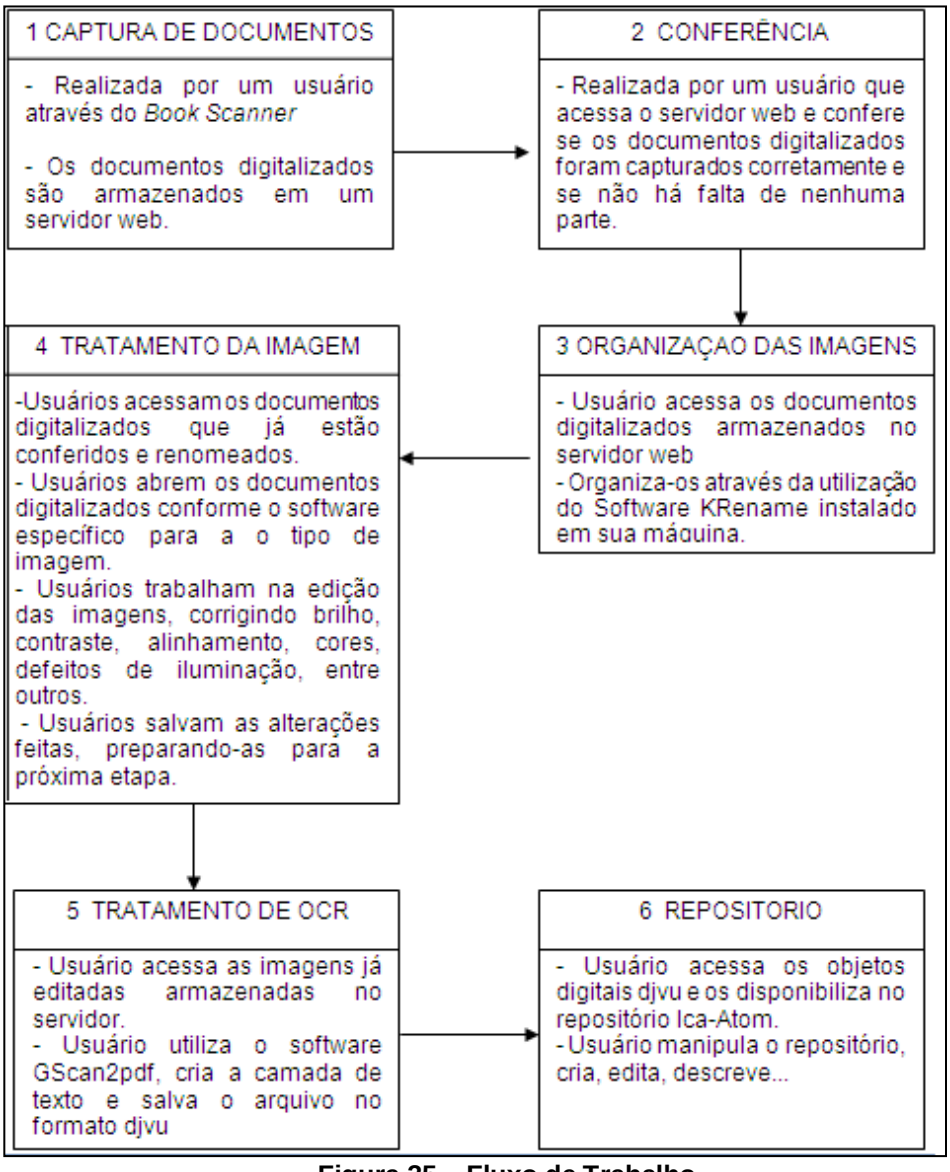

Figura 35 – Fluxo de Trabalho
Fonte: Autoria Própria

O Grupo de Atividades 1 no fluxo de trabalho visto da Figura 35, pode ser feito por mais de uma pessoa, a fim de agilizar o procedimento. A atividade compreende as ações de: posicionar corretamente o documento no *book scanner*, alinhá-lo, posicionar as câmeras digitais, verificar a iluminação e fazer a captura da imagem. O tempo desta atividade é relativo, pois depende do tamanho do documento que está sendo digitalizado, quanto mais páginas tiver, ou quanto mais frágil for seu manuseio mais tempo demorará. Enquanto estiver ocorrendo esta atividade, as demais podem estar sendo realizadas concomitantemente em outros documentos já digitalizados e armazenados no servidor web.

As imagens vão sendo salvas automaticamente na pasta do servidor web. O usuário da atividade 2 já vai realizando sua ação de conferir se as imagens estão completas e como deveriam. Qualquer erro encontrado é comunicado imediatamente aos usuários da atividade 1, que então corrigirão o erro de captura. Quando a captura estiver correta, o usuário enviará um comando de verificação correta. Ele transfere as imagens do diretório de captura para o diretório de imagens conferidas.

O usuário responsável pela atividade 3 é responsável por organizar e renomear os arquivos, atribuindo a arquivos do mesmo grupo o mesmo prefixo ou sufixo e o seu número de índice. Cada grupo de imagens pertencerá a um diretório dentro do sistema, o qual será nomeado com o título do documento que foi digitalizado e a data da digitalização.

Os usuários da atividade 4 trabalham com as imagens armazenadas nos diretórios que já estão em ordem, ou seja, que já passaram pela atividade 3. Esta atividade também tem um tempo relativo, pois depende do estado do documento digitalizado, alguns precisam de mais cuidados que outros. Os usuários podem trabalhar cada um na sua máquina e divididos em dois conjuntos, um tratando de imagens pretas e brancas e outro das imagens coloridas. Como as imagens possuem número de índice, é fácil para os usuários fazerem a distribuição do trabalho. Eles têm acesso à imagem armazenada no servidor web e cada um pode ir trabalhando em uma imagem e salvando as edições. As imagens devem ser tratadas, deixando-as o mais nítido possível, porém, sem se afastar da original, deve ser corrigido o fundo, contraste, brilho, cores, alinhamento, bordas, entre outras. Esta atividade pode acontecer concomitantemente com as demais. As imagens editadas são salvas no diretório “Preparados OCR”.

Os usuários da atividade 5 deverão acessar as imagens contidas nos diretórios preparados com o OCR e aplicá-las ao *software* Gscan2pdf, fazendo o reconhecimento de caracteres, conferindo o resultado, adicionando a camada texto, a camada imagem e salvando o resultado com a extensão djvu.

O último grupo compreende o(s) usuário(s) que coordenam o repositório. Eles possuem diversas atividades. Devem desempenhar suas atividades conforme o tipo de usuário cadastrado na aplicação. Mais detalhes sobre os tipos de usuários foram vistos anteriormente na seção de descrição do repositório. Nesta atividade, os usuários farão as descrições arquivísticas, acessarão os arquivos no servidor web e colocarão no repositório.

Cada usuário terá acesso no sistema, por questão de segurança, conforme seu papel desempenhado. Assim é mais fácil também de realizar o controle das atividades.

As atividades descritas anteriormente são constituídas por tarefas, e as tarefas são componentes essenciais do processo.

> As atividades em um fluxo de trabalho são executadas por papeis associados a cada atividade. Aos papéis são associados atores, que podem ser pessoas ou agentes automatizados. Atores executam as atividades determinadas para os papeis assumidos. (COSTA, 2009)

Assim, o modelo de *workflow* será construído a partir da representação gráfica dos elementos descritos anteriormente. Costa (2009) afirma que é importante, primeiramente elaborar o diagrama de caso de uso “porque será a base para formalizar as funcionalidades que o *workflow* deverá cumprir. Um caso de uso descreve as interações entre o sistema e os atores”, ou seja, discrimina como será o comportamento do *workflow*.

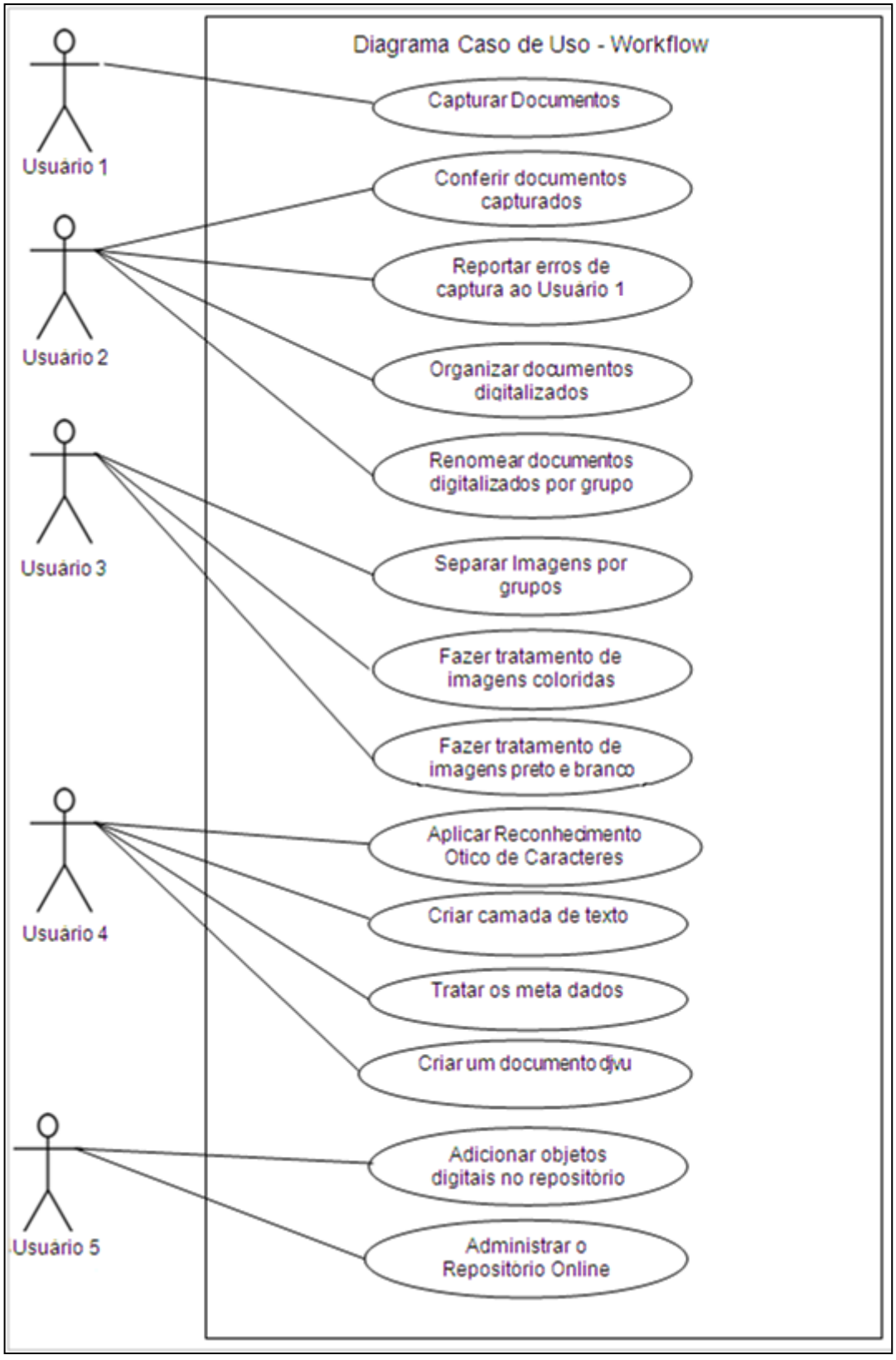

Figura 36 – Diagrama Geral de Casos de Uso – *Workflow*
Fonte: Autoria Própria

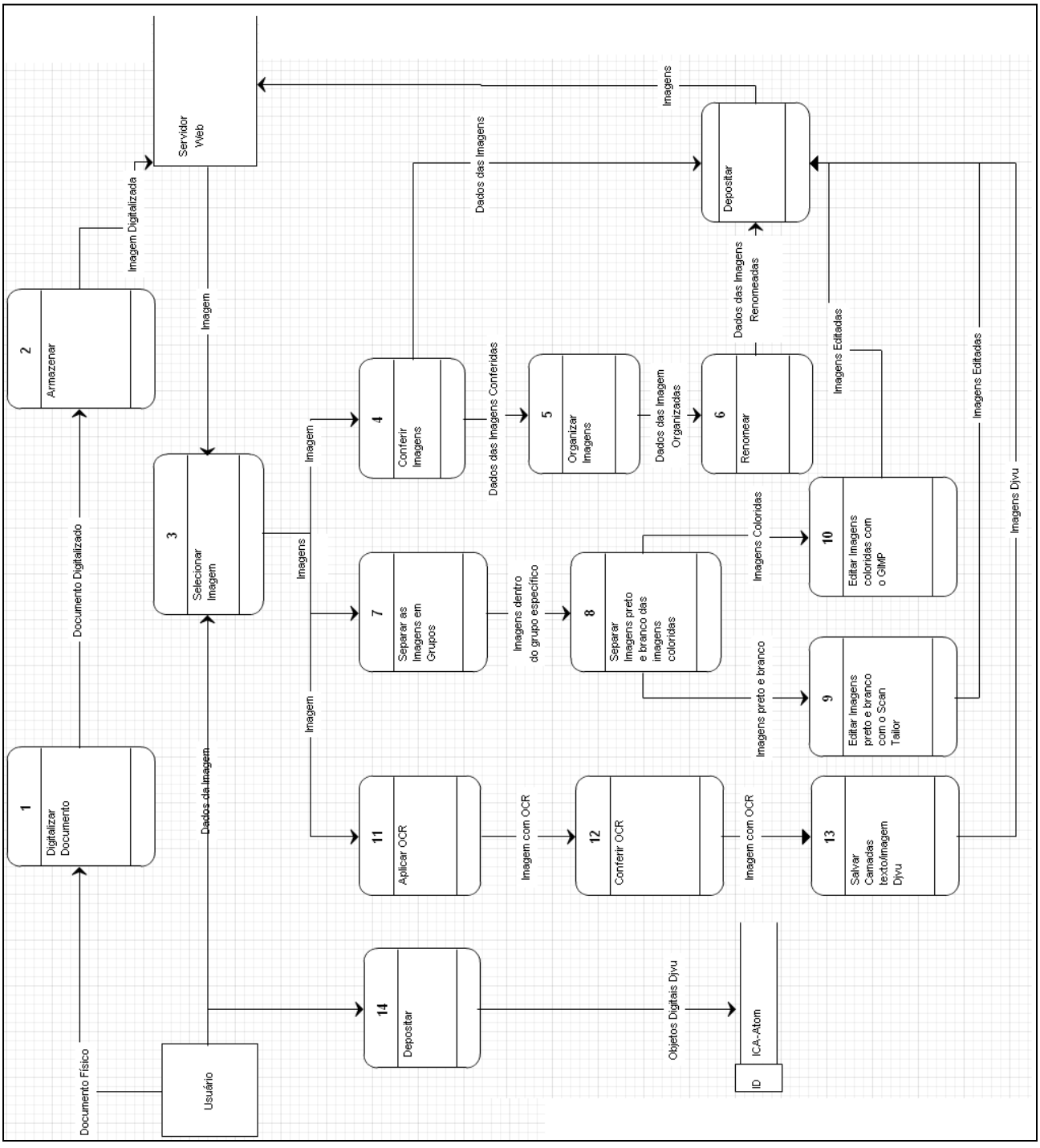

Figura 37 – Diagrama de Fluxo de Dados
Fonte: Autoria Própria

Após o levantamento de dados, é possível utilizar ferramentas para efetivamente construir o *workflow*. As maiorias dos sistemas possuem *interface* gráfica, onde o usuário pode estabelecer cada atividade e a seqüência de regras de roteamento. (COSTA, 2009)

Exemplo de ferramentas que realizam a implantação do *workflow*, conforme exposto por Mourão (2006):

- Bonita (versão 1.7.1): Ferramenta completa, com boas opções de gerenciamento das atividades e programa gráfico para criação de fluxos
- *JBoss jBPM* (versão 3.2 *alpha* 1): Possui uma linguagem de *workflow* própria chamado de *jBPM Process Definition Language* (JPDL), também oferece suporte a *Business Process Execution Language*. Ferramenta ágil na questão de criação e remoção de processos.
- OBE – *Open Business Engine* (versão 1.0 RC1): Pouca documentação e instalação complexa.
- *Enhydra Shark* (versão 2.0 beta 1): Faz uso de tecnologias pouco conhecidas, tais como o DODS (persistência objeto-relacional) e XMLC (tecnologia de *interface* com o usuário).
- *WfMOpen* versão 1.3.4: Bem documentada e aderente aos padrões *XPDL e Xforms.*
- *Yawl* versão beta 7: Linguagem de *workflow* própria e com bastante fonte para pesquisa acerca da ferramenta.

Dentre as ferramentas citadas acima, a que ofereceu melhor usabilidade e maior eficiência foi a *BonitaSoft.* Este *software* é muito completo, de fácil utilização e, principalmente, muito bem documentada, o que acaba auxiliando muito os novos usuários. A figura 38 mostra a interface do *BonitaSoft:*

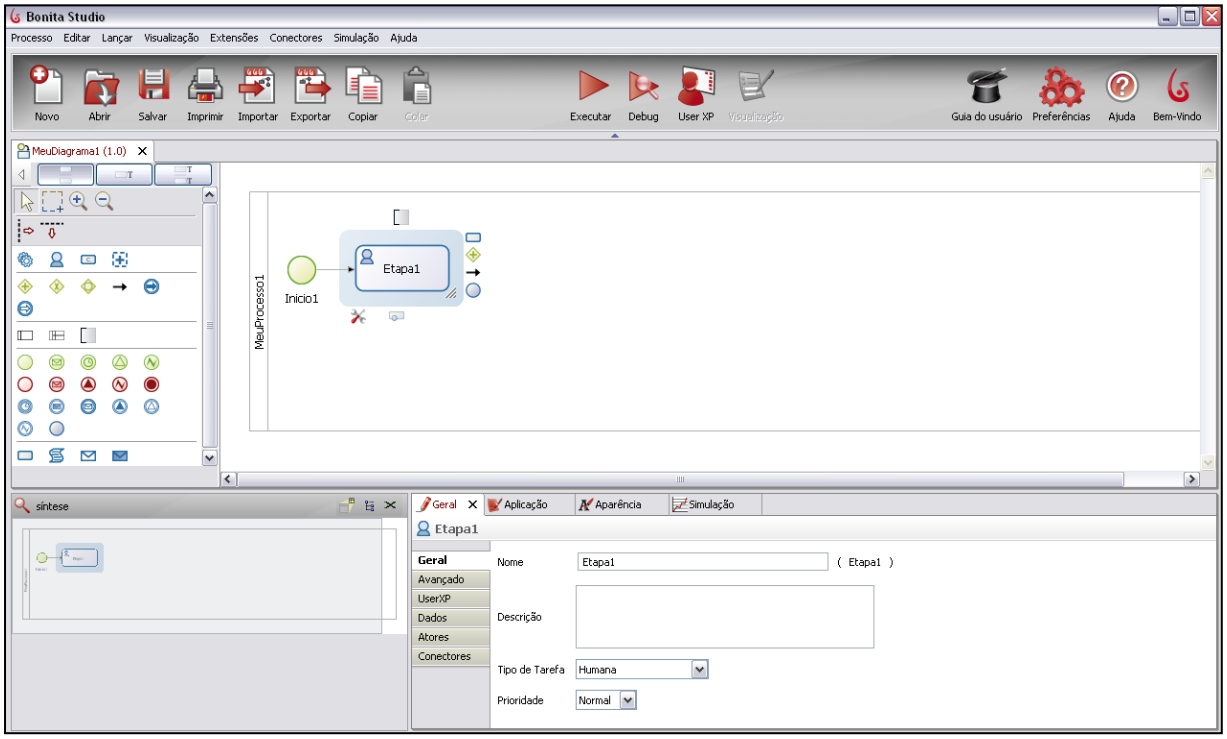

Figura 38 – *Bonita Soft – Interface*
Fonte: *BONITA SOFT* (2012)

A parte superior esquerda contém uma paleta dos elementos que podem ser usados para editar o *workflow*. Ao lado, na parte superior direita, o espaço em branco é destinado para a montagem do fluxo de trabalho, os elementos são colocados neste espaço. A parte inferior esquerda mostra a visão geral do *workflow,* quando o *workflow* está com um tamanho consideravelmente complexo, este espaço auxilia o usuário para encontrar o ponto específico que está procurando dentro do todo. A parte inferior direita é destinada para a configuração de detalhes, editando os elementos de forma individual.

Cada elemento pode ser configurado de forma bem específica. Por exemplo, a configuração de uma das etapas do *workflow,* como pode ser visualizado na Figura 39 – Configuração Etapa – *Workflow,* possui diversos campos a serem editados como: nome, descrição, tipo de tarefa (humana, receber, enviar, abstrata...), nível de prioridade, sumário da etapa, etiquetas, atores, conectores e demais dados.

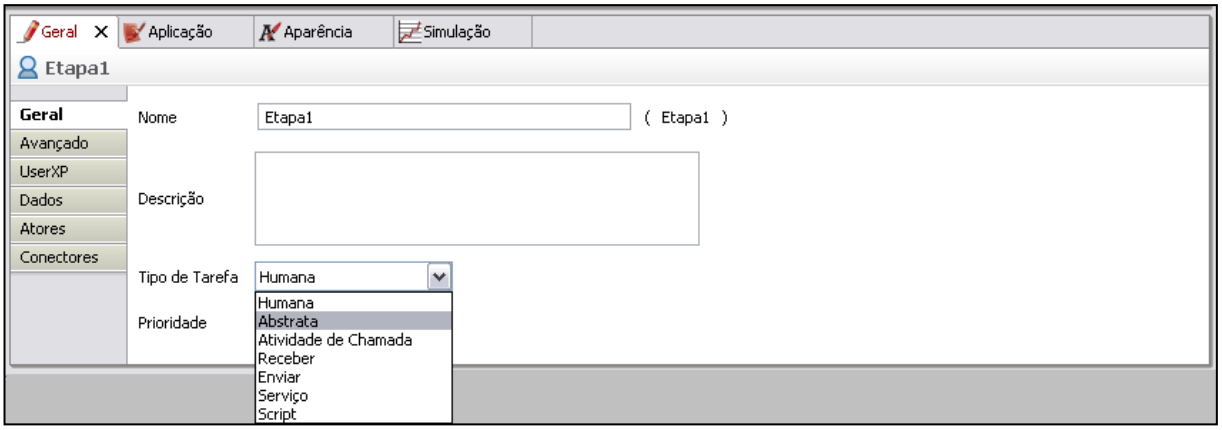

Figura 39 – *Bonita Soft* – Configuração de Etapa
Fonte: *BONITA SOFT* (2012)

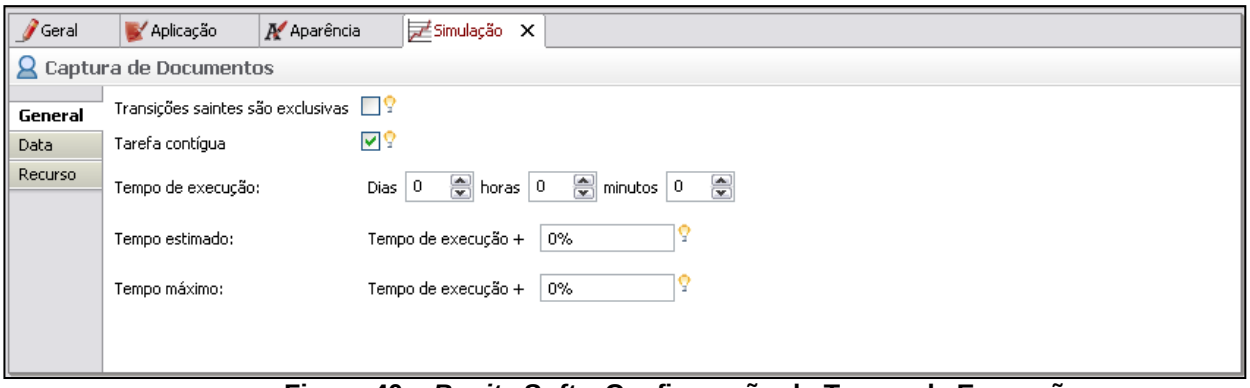

Figura 40 – *Bonita Soft* – Configuração do Tempo de Execução
Fonte: *BONITA SOFT* (2012)

A figura 40 exemplifica o que foi falado anteriormente relativo ao tempo de execução. Cada etapa pode ter seu tempo de execução pré-estabelecido, inclusive, a ferramenta possibilita fazer a simulação de como as tarefas estão acontecendo.

A figura 42 apresenta um exemplo de uma utilização simples de um subprocesso. A ferramenta possibilita a inserção de processos e de etapas (subprocessos). Pode-se criar vínculo de dependência e estabelecer a ordem de ação de cada uma das etapas do processo. Os subprocessos também são descritos detalhadamente. O *BonitaSoft* também permite que sejam criadas variáveis de entrada e saída de dados e condições de transição .

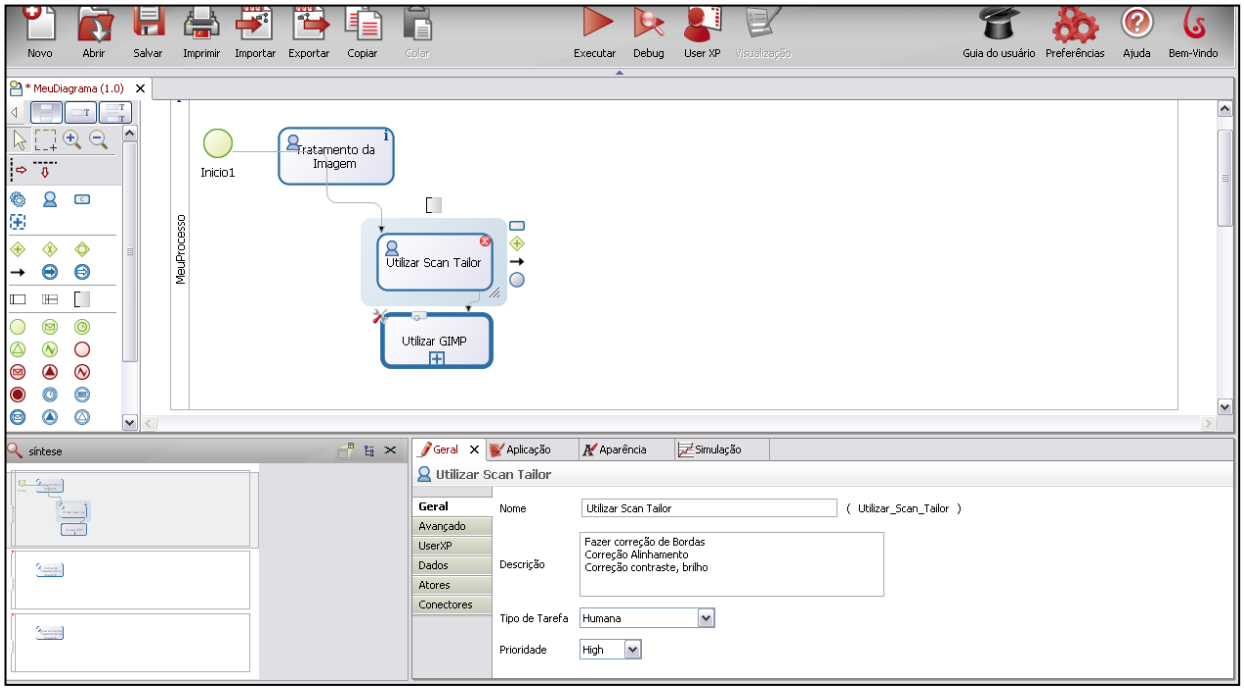

Figura 41 – *Bonita Soft* - Utilização Subprocesso
Fonte: *BONITA SOFT* (2012)

Por fim, com a documentação deste trabalho é possível implementar o *workflow* na ferramenta *BonitaSoft.* O *workflow* criado atenderá as características e os benefícios que foram citados no capítulo 2, assim otimizando e gerenciando as etapas do processo de digitalização de documentos. A figura 42 mostra a aplicação do Diagrama de Fluxo de Dados (Figura 32) na ferramenta *BonitaSoft.*

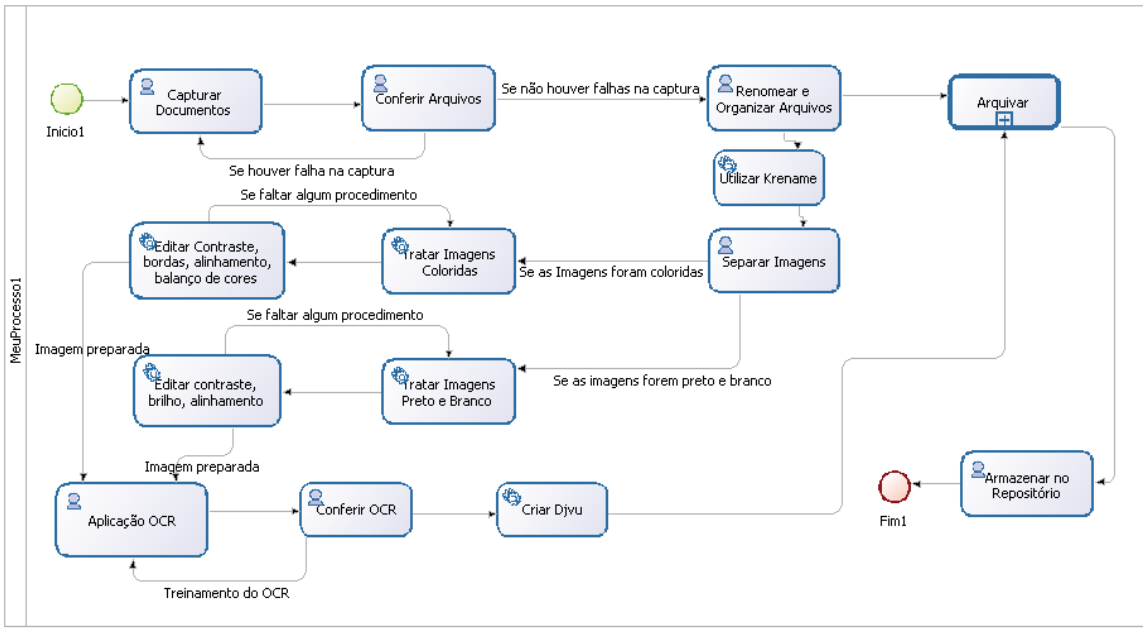

Figura 42 - *BonitaSoft* – Diagrama Fluxo de Dados
Fonte: Autoria Própria

“Após a configuração do *workflow*, é recomendável que seja desenvolvido um processo piloto para a validação do *workflow* implantado”. (COSTA, 2009)

O próprio software escolhido já permite que se faça uma simulação antes de efetivamente implementar o modelo de *workflow* construído.

5. CONCLUSÃO

O estudo realizado neste trabalho atingiu aos objetivos propostos. Como pôde ser verificado no decorrer do trabalho, a implantação do Gerenciamento Eletrônico de Documentos traz diversos benefícios consideráveis, e, neste rol, destacam-se a possibilidade de uma melhor conservação dos documentos originais e, principalmente, uma facilitação no acesso do público para consultas e pesquisas nos acervos.

O GED é uma ótima solução, porém, muitas instituições não têm acesso a esta ferramenta por ela ainda ser de alto custo. Assim, este trabalho foi desenvolvido fazendo uso somente de *softwares* livres. Para alcançar a tecnologia GED, é necessário que seja estudada e entendida cada uma das fases que compõe o processo de produção de objetos digitais.

Seguindo esta linha de raciocínio, foram estabelecidas as seguintes etapas do processo de digitalização de documentos: captura de documentos físicos, conferência, organização, tratamento de imagem, aplicação do OCR, formação e armazenamento dos objetos digitais no repositório *online.* Pode-se constatar que as etapas formam uma seqüência de dependência, caso uma das etapas seja realizada de modo insatisfatório, todas as subseqüentes serão prejudicadas, e, conseqüentemente, a produção final do objeto digital também.

As etapas foram estudadas e detalhadas, resultando assim na descrição das atividades pertinentes a cada fase. Para uma melhor explanação do conteúdo, a teoria foi exemplificada através do estudo de caso, com descrições das atividades e ilustração das telas referentes a utilização dos *softwares*.

Os *softwares* utilizados foram escolhidos através de uma seleção, onde foram analisados segundo uma lista de requisitos: usabilidade, confiabilidade de resultado, eficiência, entre outros. Os que apresentaram melhores resultados foram: para a organização de arquivos: *KRename*; para o tratamento de imagens preto e branco: *Scan Tailor*; para o tratamento de imagens coloridas: *GIMP*; para a aplicação de OCR: *GScan2pdf*; e para o armazenamento dos objetos digitais: *Ica-Atom*.

Por fim, foi apresentado um estudo sobre a tecnologia *Workflow*, abordando seus conceitos, classificações e aplicações. Esta parte do estudo serviu de base para a modelagem de um *workflow* aplicado ao processo de digitalização de

documentos. Foi construído um diagrama de fluxo de dados e um diagrama de atividades que formam a base para a possível implementação da tecnologia.

Também foi apresentada uma lista de indicações de *softwares* livres que realizam esta implementação, e alguns exemplos da utilização da ferramenta *BonitaSoft,* que foi a mais adequada para utilização. Com a documentação é possível implementar efetivamente o *workflow* para o processamento e a construção de objetos digitais. O *workflow* acaba otimizando e automatizando as tarefas executadas entre os participantes.

Este estudo pode ser utilizado por instituições para a implementação de uma tecnologia de digitalização de seus acervos, e, como foi visto, esta tecnologia traz muitos benefícios, abrange diversas áreas e vem sendo cada vez mais aplicada.

O trabalho contribuiu para uma sistematização do processo de digitalização de documentos, aspecto que é cada vez mais utilizado, uma vez que o uso da informática está se tornando essencial para qualquer organização.

6. TRABALHOS FUTUROS

Este trabalho servirá de referência para trabalhos futuros na área de digitalização de documentos. O *workflow*, quando aplicado, poderá otimizar todas as etapas de digitalização de documentos. Pode-se pensar na implantação de um web Server que comporte todas as aplicações, possibilitando assim, que um grupo de pessoas trabalhem em conjunto na digitalização de um livro, por exemplo, sem terem que acessar cada um dos programas isoladamente.

Pode também ser realizados trabalhos acerca de um aprimoramento do reconhecimento de caracteres, inclusive para aplicações em documentos manuscritos. Existem poucas fontes e materiais sobre o treinamento de motores de OCR.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, Neide F. **Estratégias Para Melhoria Do Desempenho De Ferramentas Comerciais De Reconhecimento Óptico De Caracteres,** 108f. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Pernambuco, 2003. Disponível em: http://www.liber.ufpe.br/teses/arquivo/20040603161933.pdf, Acesso em: 15 de abril de 2012.

ARAUJO, R. M.; BORGES, M. R. S., **Sistemas de *Workflow***, XX Jornada de Atualização em Informática, Congresso da SBC - 2001.

ARTSIMOVICH, Joseph. **SCAN TAILOR.** Disponível em: <http://scantailor.sourceforge.net/> Acesso em: 19 jun. 2012.

AALST, Wil.V.D. ; HEE, Kees. V.; **Gestão de *Workflow*s: Modelos, métodos e sistemas.** Coimbra: Editora Imprensa da Universidade de Coimbra, 2009. 316p.

BRASIL. Lei n. 8.159, de 8 de janeiro de 1991. Dispõe sobre a política nacional de arquivos públicos e privados e dá outras providências. **Diário Oficial, Brasília, 9 jan. 1991.**

BELLOTTO, Heloísa Liberalli. **Arquivística: objeto princípios e rumos.** São Paulo: AASP, 2002

BONITA SOFT. Disponível em: <http://www.bonitasoft.com> Acesso em: 20 jun. 2012

BREUEL, Thomas M. **The *Ocropus* Open Source OCR System.** Kaiserslautern, Germany: DFKI and U. Kaiserslautern, 2006.

CRUZ, Tadeu. **Sistemas de Informações Gerenciais: Tecnologias da Informação e a Empresa do Século XXI.** 2ª Ed. São Paulo: Atlas, 2000. 245p.

CRUZ, Tadeu. ***Workflow* II: A Tecnologia que Revolucionou Processos.** Rio de Janeiro: E-Papers Serviços Editoriais Ltda., 2004. 212p.

CONSELHO NACIONAL DE ARQUIVOS - CONARQ. **Modelo de requisitos para sistemas informatizados de gestão arquivística de documentos**; e-Arq, CONARQ, 2010. Disponível em: http://www.conarq.arquivonacional.gov.br/media/publicacoes/recomenda/recomendaes_para_digitalizao.pdf. Acesso em: 20 de março de 2012.

COSTA, Lourenço. **Formulação de uma metodologia de Modelagem de Processos de Negócio para implementação de *workflow***. 2009. 126 f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2009.

DJVU. Disponível em: <http://djvu.org/> Acesso em: 05 mai. 2012

GESTÃO ELETRÔNICA DE DOCUMENTOS. Disponível em: http://www.ged.netbr> Acesso em: 20 jun. 2011

GOMES, Otávio F. M. **Microscopia co-localizada: novas possibilidades na caracterização de minérios**, 104f. Tese (Doutorado em Ciência dos Materiais e Metalurgia) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

HADDAD, Samir R. **GED - Uma Alternativa viável na Gestão da Informação Estratégica,** 39f. Monografia (Especialização em Informática Pública) - Prodabel e Universidade Católica de Minas Gerais, 2000.

HOLLINGSWORTH, D.. **The Workflow Reference Model.The Workflow Management Coalition Specification TC00-1003**, Workflow Management Coalition, Hampshire, UK, 1995.

ICA-ATOM. Disponível em: <http://ica-atom.org/ > Acesso em: 02 mai. 2012

JUNIOR, Alfredo Luiz Santos. **Quem mexeu no meu sistema.** Rio de Janeiro: Brasport, 2008, 193 p.

FILHO, Ly Freitas. **Sistemas *workflow* em processos empresariais, aplicando técnicas de processos produtivos fabris e modelagem em redes de petri.** 36f. **Dissertação** (Curso de Pós Graduação em Engenharia Elétrica) – Universidade Mackenzie, 2000.

MARQUES FILHO, Ogê; VIEIRA NETO, Hugo. **Processamento Digital de Imagens**. Rio de Janeiro: Brasport, 1999

MARQUES, C. A.; APARECIDO, E. L.; OKABE, N. **A Problemática Da Conservação Da Informação Nos Acervos Bibliográficos E Suas Formas De Conversão Para Mídia Eletrônica**. Disponível em: <http://www.pesquisas.unicoc.edu.br/arquivos/A_PROBLEMATICA_DA_CONSERVACAO_DA_INFORMACAO_NOS_ACERVOS_BIBLIOGRAFICOS.pdf> . Acesso em: 25 jun. 2011.

MARTINS, Ana B.; RODRIGUES, Eloy; NUNES, Manoela B. **Repositórios de informação e ambientes de aprendizagem:** Criação de espaços virtuais para a promoção da literária e da responsabilidade social. Redes de Bibliotecas Escolares: Newsletter n.03. Disponível em: <http://www.rbe.min-edu.pt/news/newsletter3/repositorios.pdf>. Acesso em: 02 mai. 2012.

MOURÃO, Walter I. **Avaliação do uso de ferramentas de *workflow* em processos típicos de engenharia de *software*.** Belo Horizonte: Arcadian Tecnologia S/A, 2006.

NASICMENTO, Anna C. A. A.; PERES, Cristiane V.; OLIVEIRA, Maria J.; SILVA, Karla I. C. **Guia Para a Digitalização de Documentos Versão 2.0.** 2ª Ed. Brasília: Embrapa STC, 2006.43p.

NATTERER, Michael; NEUMANN, Sven, et al. **GIMP – GNU Image Manipulation Program.** Disponível em: <http://www.gimp.org > Acesso em: 20 jun. 2012.

***OCROPUS*.** Disponível em: <http://code.google.com/p/*Ocropus*/>. Acesso em: 20 de abr. de 2012.

***OPENFST*.** Disponível em: <http://www.openfst.org/>. Acesso em: 01 de mai. de 2012.

OSÓRIO, Fernando S. **Um Estudo sobre o Reconhecimento Visual de Caracteres através de Redes Neurais.** 303f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul, 1991.

PENA, M. G.; SILVA, A.C. **A Digitalização De Documentos Históricos E A Gestão Eletrônica De Documentos Para Disponibilização On Line.** Saber Digital: Revista Eletrônica do CESVA, Valença, v. 1, n. 1, p. 85-102, mar./ago. 2008

PEREIRA, Luiz A. M.; CASANOVA, Marco A. Sistemas de Gerência de *Workflow*s: Características, Distribuição e Exceções. **PUC – Rio Inf.MCC.** Rio de Janeiro V.11 n.03 Mar. 2003. Disponível em: <ftp://ftp.inf.puc-rio.br/pub/docs/techreports/03_11_pereira.pdf> Acesso em: 10 fev. 2012

RATCLIFFE, Jeffrey. **GSCAN2PDF**. Disponível em: http://gscan2pdf.sourceforge.net/ Acesso em: 20 jun. 2012

RONCAGLIO, Cynthia; SZVARÇA, Décio Roberto; BOJANOSKI, Silvana de Fátima. **GESTÃO DE DOCUMENTOS E INFORMAÇÃO** Enc. BIBLI: R. Eletr. Bibl. Ci. Inf., Florianópolis, n. esp., 2º sem. 2004

RONDILELI, Roseli C. **Gerenciamento Arquivísticos de Documentos Eletrônicos: uma abordagem teórica.** 1ª Ed. Rio de Janeiro: Editora FGV, 2002. 160 p.

RUIZ, Duncan D.; de OLIVEIRA, José P. M.; NICOLAO, Mariano. *Workflow*: Conceitos e Abrangência. **Acta Scientiae - Revista de Ciências Naturais e Exatas.** Canoas, Vol.3, nº1/2, p. 49 – 64, jan./dez. 2001.

SANTANA, Jonh W. S. de. **Sistemas *Workflow*: Uma aplicação ao IC,** 20f. Monografia (Bacharelado em Ciência da Computação) – Universidade Federal de Alagoas, 2006.

SEICHTER, Dominik. **KRENAME.** Disponível em: <http://www.krename.net> Acesso em: 20 jun. 2012.

SILVA, Flávio L. O. e. **Gerenciamento Eletrônico de Documentos (GED): Natureza, Princípios e Aplicações**, 72f. Monografia (Bacharelado em Ciência da Computação) – Universidade Federal do Mato Grosso, 2001. Disponível em: < http://www.arquivar.com.br/espaco_profissional/sala_leitura/teses-dissertacoes-e-monografias/GED_natureza_principios_aplicacao.pdf/view> Acesso em: 26 set 2011.

THOMAZ, Kátia P. **Repositórios Digitais Confiáveis e Certificação.** Arquivística.net Rio de janeiro, v.3, n.1, p. 80-89, jan./jun.2007

TRISTESSE. **GPRENAME.** Disponível em: <http://gprename.sourceforge.net/> Acesso em: 20 jun. 2012.

VIEIRA, Tatiana Almeida S. C. **Execução Flexível de *Workflow*s.** 2005. 259 f. **Tese** (Doutorado em Informática). Centro Técnico Científico da PUC – Rio, Rio de Janeiro, 2005.