UNIVERSIDADE FEDERAL DE SANTA CATARINA

Centro de Ciências da Educação

CURSO DE GRADUAÇÃO EM BIBLIOTECONOMIA

Alexandre Pedron Martins

O FORMATO PDF/A COMO MEIO DE PRESERVAÇÃO DIGITAL NO PROCESSO DE DIGITALIZAÇÃO DAS TESES E DISSERTAÇÕES DA BIBLIOTECA CENTRAL DA UFSC

Florianópolis

2012

ALEXANDRE PEDRON MARTINS

O FORMATO PDF/A COMO MEIO DE PRESERVAÇÃO DIGITAL NO PROCESSO DE DIGITALIZAÇÃO DAS TESES E DISSERTAÇÕES DA BIBLIOTECA CENTRAL DA UFSC

Trabalho de Conclusão do Curso de Graduação em Biblioteconomia, do Centro de Ciências da Educação da Universidade Federal de Santa Catarina, como requisito parcial à obtenção do título de Bacharel em Biblioteconomia. Orientação de: Prof. Dr. Marcio Matias.

Florianópolis

2012

Acadêmico: Alexandre Pedron Martins

Título: O formato PDF/A como meio de preservação digital no processo de digitalização das teses e dissertações da Biblioteca Central da UFSC

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Biblioteconomia, do Centro de Ciências da Educação da Universidade Federal de Santa Catarina, como requisito parcial à obtenção do título de Bacharel em Biblioteconomia, aprovado com nota 8,5

Florianópolis, 18 de Dezembro de 2012.

Marcio Matias, Dr. UFSC
Professor Orientador

Adilson Luiz Pinto, Dr. UFSC
Membro da Banca Examinadora

William Barbosa Vianna, Dr. UFSC
Membro da Banca Examinadora

Ficha catalográfica elaborada por Alexandre Pedron Martins, graduando de Biblioteconomia da Universidade Federal de Santa Catarina.

M379f Martins, Alexandre Pedron, 1976 –

O formato PDF/A como meio de preservação digital no processo de digitalização das teses e dissertações da Biblioteca Central da UFSC / Alexandre Pedron Martins. - - 2012.
93 f. : il. ; 30 cm

Orientador: Márcio Matias.

Trabalho de Conclusão de Curso (Graduação em Biblioteconomia) – Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Florianópolis, 2012.

1. Biblioteconomia. 2. Preservação digital. 3. PDF/A. 4. Digitalização. I. Título.

CDD – 025.84

AGRADECIMENTOS

Ao professor Márcio Matias, meu orientador, obrigado pela ajuda, confiança e recomendações na elaboração do trabalho.

Aos professores do CIN/UFSC, em especial a Araci Isaltina de Andrade Hillesheim e Estera Muszkat Menezes.

Ao bibliotecário Rafael Cobbe pela ajuda e sugestões.

A Claudiane Weber, pela competência e dedicação como bibliotecária e professora.

Aos meus colegas de turma, que sempre ajudaram a manter um clima de alegria ao longo do curso, em especial a João Paulo, Denise e Suelen.

A minha família, em especial a minha mãe e minha irmã pela força.

A minha namorada Francielli, por me acompanhar nesta longa caminhada. Obrigado pela atenção, risos, paciência e parceria.

A todas as pessoas que, direta ou indiretamente, colaboraram para a realização desta pesquisa.

MARTINS, Alexandre Pedron. **O formato PDF/A como meio de preservação digital no processo de digitalização das teses e dissertações da Biblioteca Central da UFSC**. 93 f. 2012. Trabalho de conclusão de curso (Graduação em Biblioteconomia) – Centro de Ciências da Educação, Universidade Federal de Santa Catarina, Florianópolis, 2012.

RESUMO

Este trabalho baseou-se na literatura dos processos de digitalização e da temática da preservação digital e teve como objetivo analisar a utilização do formato PDF/A inserido num processo de digitalização de documentos impressos. Fez-se uma revisão bibliográfica sobre a preservação digital e os formatos de documentos eletrônicos utilizados para armazenamento digital da informação. Descrevem-se os aspectos conceituais e técnicos da digitalização de documentos impressos. Apresenta uma comparação entre os formatos TIFF e PDF/A nas suas características técnicas e em testes realizados em diferentes configurações. Conclui que o formato PDF/A atende as necessidades de armazenamento digital de documento eletrônico e estabelece um nível de segurança para que os documentos eletrônicos possam ser acessados a longo prazo.

Palavras-chave: Biblioteconomia. Preservação digital. PDF/A. Digitalização. Documento eletrônico.

MARTINS, Alexandre Pedron. **O formato PDF/A como meio de preservação digital no processo de digitalização das teses e dissertações da Biblioteca Central da UFSC**. 93 f. 2012. Trabalho de conclusão de curso (Graduação em Biblioteconomia) – Centro de Ciências da Educação, Universidade Federal de Santa Catarina, Florianópolis, 2012.

ABSTRACT

This work was based on the literature the digitization and digital preservation issue and aims to analyze the use of PDF/A inserted in a process of scanning paper documents. There was a review of the literature on digital preservation and electronic document formats used for storing digital information. We describe the conceptual and technical aspects of digitization of printed documents. Presents a comparison between the formats TIFF and PDF/A in its technical characteristics and tests in different settings. It concludes that the PDF/A format meets the needs of digital storage and electronic document provides a level of security for electronic documents that can be accessed over the long term.

Key-words: Librarianship. Digital preservation. PDF/A. Digitization. Electronic document.

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTAS DE ABREVIATURAS E SIGLAS

ABNT - Associação Brasileira de Normas Técnicas

ALA - *American Library Association*

CCITT - *International Telegraph and Telephone Consultative Committee*

CONARQ – Conselho Nacional de Arquivos

DICOM - *Digital Imaging and Communications in Medicine*

DPI – *Dots Per Inch*

G4 – Compressão Grupo 4

HTML - *HyperText Markup Language*

IBGE - Instituto Brasileiro de Geografia e Estatística

IFLA - *International Federation of Library Associations and Institutions*

ISO - *International Organization for Standardization*

JPEG – *Joint Photographic Experts Group*

LZW - Lempel-Ziv-Welch

OCR – *Optical Character Recognition*

PDF - *Portable Document Format*

PDF/A - *Portable Document Format for Archiving*

RLE – *Run Length Encoding*

UFSC - Universidade Federal de Santa Catarina

UNESCO - *United Nations Educational, Scientific and Cultural Organization*

TIFF - *Tagged Image File Format*

XML - *Extensible Markup Language*

XPS - *XML Paper Specification*

SUMÁRIO

1 INTRODUÇÃO

Prolongar a existência de seus registros foi o objetivo de vários povos na história da humanidade. O desejo de preservar a história esteve inserido no âmbito de muitos povos ao redor do planeta, nem todos conseguindo atingir este objetivo.

Conforme a escada da evolução avança um novo degrau, manter registrado o presente significa dar sequência à corrente evolutiva até a próxima geração, como pondera Mello (1979, p. 32) “o Homem não para, avança sempre, pois o que estaciona morre. E o Homem é imortal, porque sempre um sucede a outro, indefinidamente”.

Manter registros da evolução do Homem por períodos tão longos de tempo foi possível devido ao tipo de suporte onde foram gravados. Estes registros não tiveram apenas o tempo como inimigo natural, conflitos entre povos também contribuíram para a destruição. Suportes utilizados por povos da antiguidade podem ser encontrados nos dias atuais devido à resistência, como bem lembrados por Mello (1979, p 52) “a escrita na pedra é que resistiu ao tempo, à água, à chuva e ao fogo.”.

O processo de evolução contribuiu para o desenvolvimento dos povos, o que consequentemente refletiu na maneira de preservar seus registros. Paes (2007, p. 16) lembra que:

> Logo que os povos passaram a um estágio de vida social mais organizado, os homens compreenderam o valor dos documentos e começaram a reunir, conservar e sistematizar os materiais em que fixavam, por escrito, o resultado de suas atividades politicas, sociais, econômicas, religiosas e até mesmo de suas vidas particulares. Surgiram, assim, os arquivos, destinados não só a guarda dos tesouros culturais da época, como também a proteção dos documentos que atestavam a legalidade de seus patrimônios, bem como daqueles que contavam a historia de sua grandeza.

Partindo deste ponto, o nível de desenvolvimento da humanidade passou cada vez mais a gerar informações que necessitavam serem armazenadas de alguma maneira, seja para uso posterior ou com um intuito histórico. Os aspectos de legalidade, agregados ao cotidiano da sociedade,

passaram a produzir um número significativo de registros formais armazenados em arquivos para eventual consulta futura.

Os arquivos ganharam reconhecimento pela sociedade. Assim, também cita Paes (2007, p. 19) que “as definições antigas acentuavam o aspecto legal dos arquivos, como depósitos de documentos e papeis de qualquer espécie, tendo sempre relação com os direitos das instituições ou indivíduos”.

Do período que compreende a invenção da prensa tipográfica por Gutenberg até o inicio do século XX, a humanidade teve um aumento significativo na quantidade de informação registrada. Esta evolução relaciona-se em sua essência, ao Renascimento e a necessidade do homem moderno de armazenar informação e conhecimento.

Conforme aumentava a quantidade de conhecimento armazenado, passou-se a buscar outras maneiras de cuidar destes registros. Siqueira (2010, p. 58) corrobora descrevendo o cenário:

> No final do século XIX, com o aumento da produção bibliográfica, da pesquisa científica e o surgimento de novos suportes houve a necessidade do desenvolvimento de outras técnicas para organização e administração da informação, já que a Bibliografia não dava mais conta de tais necessidades. Enquanto as bibliotecas públicas projetavam suas atenções à educação da massa trabalhadora, produzida pela Revolução Industrial, e os arquivos procuravam se institucionalizar e resolver seus problemas de organização informacional, a Documentação abriu espaço no século XX.

Cada vez mais inserido no cotidiano da sociedade, os documentos tornaram-se indispensáveis para o funcionamento de governos, corporações e organizações. Quanto à era moderna da humanidade, Levy (1999, p. 22) salienta que “é impossível separar o humano de seu ambiente material, assim como dos signos e das imagens por meio dos quais ele atribui sentido a vida e ao mundo.”.

Com relação às mudanças impostas pela constante evolução tecnológica no universo das bibliotecas, Spellman e Holley (2011, p. 24, tradução nossa) advertem:

> Num mundo onde a digitalização cada vez mais esta se tornando popular e onde cada vez mais aumenta o número de documentos existentes somente na forma digital, bibliotecas precisam se adaptar

> a esta situação e estarem dispostas a trabalhar com os usuários para criarem um futuro viável.

Ainda no âmbito das bibliotecas, o processo de digitalização de acervos passou a ser uma ação comum devido a diversos fatores. Lapotin (2006, p. 273, tradução nossa) diz que a questão de acesso e preservação de materiais são as duas principais motivações para as bibliotecas conduzirem projetos de digitalização. Estas questões conferem com as citadas pelo Conselho Nacional de Arquivos (CONARQ, 2010, p. 4) como recomendação do processo de digitalização:

> A digitalização de acervos é uma das ferramentas essenciais ao acesso e a difusão dos acervos arquivísticos, além de contribuir para a sua preservação, uma vez que restringe o manuseio aos originais, constituindo-se como instrumento capaz de dar acesso simultâneo local ou remoto aos seus representantes digitais como os documentos textuais, cartográficos e iconográficos em suportes convencionais, objeto desta recomendação.

Sendo a facilidade de acesso ao acervo um dos objetivos a serem atingidos, as bibliotecas universitárias passaram a dar importância aos projetos de digitalização de seus acervos. Com o intuito de oferecerem o serviço de biblioteca digital, as bibliotecas universitárias iniciaram um processo de mudança como destacado por Marcondes (2006, p. 176):

> As bibliotecas começam a ser transformar: nota-se uma preocupação crescente em atender o usuário com o máximo de rapidez e eficiência, maior preocupação com o acesso a informação em detrimento da posse do documento, minimizando-se as limitações de tempo e espaço na busca da informação. As coleções e os serviços foram complementados com novos formatos e novas versões, tudo isso, certamente, facilitado pela utilização das novas tecnologias.

Assim, percebe-se que é as novas tecnologias que permitem as bibliotecas oferecerem estas novas formas de acesso ao acervo para seus usuários, indo de encontro aos ideais que formam um profissional da informação.

A busca para oferecer novas formas de acesso também foi um dos catalisadores para que a Biblioteca Central da Universidade Federal de Santa Catarina iniciasse um projeto para digitalizar e disponibilizar eletronicamente seu acervo de teses e dissertações à comunidade universitária em geral.

1.1 JUSTIFICATIVA

A justificativa para a realização desta pesquisa tem sua origem em questões pessoais, científicas e sociais. A questão pessoal é referente ao fato de o autor da presente pesquisa, estar inserido dentro de um projeto de digitalização de obras junto à Biblioteca Central da Universidade Federal de Santa Catarina. As técnicas e ferramentas utilizadas neste projeto geraram interesse sobre os benefícios que as bibliotecas têm ao digitalizarem seus acervos e de que forma o problema da preservação digital afeta o resultado deste trabalho.

Sob o ponto de vista científico, a justificativa esteve relacionada à vontade de se aprofundar nas estratégias existentes que visam permitir a preservação e acesso futuro dos documentos digitais criados na atualidade, sejam aqueles nascidos digitalmente ou aqueles convertidos a partir de outros meios.

No âmbito social, o tema da pesquisa, justificou-se pelo fato de que não há sociedade evoluída sem que sua história seja preservada para as futuras gerações. Prolongar a existência e o acesso a importantes registros é imperativo para o profissional da informação no âmbito atual.

1.2 DEFINIÇÃO DO PROBLEMA

Atualmente uma parte considerável dos documentos é criada em meio digital. Porém a quantidade de documentos impressos armazenados e as limitações de acesso impulsionam cada vez mais as organizações a digitalizarem estes documentos.

O simples ato de converter um documento impresso em digital não assegura a preservação para futuros acessos e usos. Desta forma, surgem dúvidas sobre quais seriam as melhores opções para armazenar de forma eletrônica os documentos digitalizados. Qual é o formato que apresenta as características mais adequadas para garantir a integridade, a preservação e o

acesso às teses e dissertações da Biblioteca Central da Universidade Federal de Santa Catarina?

1.3 OBJETIVOS

Os objetivos deste trabalho dividiram-se em objetivo geral e objetivos específicos, apresentados a seguir.

1.3.1 Objetivo Geral

Analisar o uso do formato PDF/A no processo de digitalização das teses e dissertações da Biblioteca Central da UFSC.

1.3.2 Objetivos Específicos

Os objetivos específicos foram:

- Levantar as questões que envolvem a preservação digital em longo prazo.

- Identificar as características de preservação digital do formato PDF/A.

- Comparar o formato PDF/A e o formato TIFF como meio de armazenamento de documentos eletrônicos.

2 REFERENCIAL TEÓRICO

2.1 PRESERVAÇÃO DE DOCUMENTOS

O conceito de documento costuma variar de acordo com o contexto de onde é aplicada a palavra. Normalmente qualquer suporte físico que carregue consigo alguma informação é descrito como um documento. Estes suportes variam desde livros a mapas, passando por filmes, fotografias, manuscritos e cartazes, entre outros tipos.

Como lembrado por Marcondes (2010, p. 10), por meio de sua função informativa, documentos viabilizam, de forma mediada, a transferência de conhecimento.

Transportar uma informação em si é a função básica de um documento. Mas a informação tem que possuir uma utilização, um valor para a sociedade. Ortega (2010) salienta que:

> A capacidade de um documento ser informativo implica o aspecto pragmático do objeto informacional a medida que revela o caráter social e simbólico da informação e, consequentemente, os ambientes e as situações concretas de uso.

O valor que um documento tem está intrinsicamente relacionado à informação contida nele e sua relação com o mundo ao seu redor. Este valor é que proporcionará autonomia para o processo de conservação do próprio documento, determinando o esforço em mantê-lo integro pelo maior tempo possível.

As decisões sobre estas questões ficam a cargo dos profissionais responsáveis pelos documentos. Para Hazen (2001, p. 7), “a preservação ocupa posição de destaque entre os principais problemas dos bibliotecários”. Portanto, é crucial analisar todos os aspectos que envolvem a preservação de um determinado acervo.

Há diversos fatores que influenciam escolhas no universo da preservação de documentos. Questão de custos é um dos fatores decisórios neste ponto, como salienta Conway (2001, p. 12):

> A preservação tradicional, como forma responsável de resguardar essas informações, funciona somente quando a prova tem uma forma física, quando o seu valor é superior aos custos de sua manutenção, e quando os papeis desempenhados pelos seus criadores, responsáveis por sua guarda e usuários são mutuamente reforçados.

A preservação de documentos como processo recente nas organizações também tem suas características. Hazen (2001, p. 7) separa as atividades de preservação da seguinte forma:

> A preservação pode ser entendida como o agrupamento de três tipos principais de atividade. O primeiro tipo concentra-se nos ambientes de biblioteca e nas maneiras de torná-los mais apropriados a seus conteúdos. O segundo incorpora esforços para estender a vida física de documentos através de métodos como desacidificação, restauração e encadernação. O terceiro tipo envolve a transferência e de conteúdo intelectual ou informativo de um formato ou matriz para outro.

Para os profissionais bibliotecários e arquivistas, a atividade de transferir a informação de um formato para outro se tornou complexa e sujeita a falhas que somente são percebidas posteriormente. Decidir qual o suporte a ser utilizado para armazenar o conteúdo é crucial.

Apesar de utilizado desde a década de 1920, a microfilmagem ainda é utilizada como principal mecanismo de armazenagem e preservação de grandes volumes de documentos.

Tradicional meio utilizado para armazenar e preservar documentos, a microfilmagem acaba por trazer empecilhos para o usuário devido estrutura necessária para acesso. Como destacado por Waters (2001, p. 14), “diferentemente de um livro, que pode ser carregado e utilizado em qualquer lugar, o microfilme obriga o usuário a utilizar um equipamento especial de projeção, em um local especifico”. Somado a isto, Waters (2001, p. 14) complementa dizendo que um microfilme é difícil de ser folheado e lido, pois não há a mesma facilidade de uso em relação a copia de papel.

Não sendo considerada uma substituição ao processo de microfilmagem, mas sim um recurso complementar, a digitalização passou a ser uma opção escolhida por organizações que visam ter os benefícios de um acervo em formato eletrônico.

Pearson (2002) acredita que há uma necessidade urgente para que as bibliotecas adotem ações em conjunto e encontrem um mecanismo que coloque a digitalização e a estratégia para atingi-la como alta prioridade.

Já Rikowski (2011, tradução nossa) expõe três preocupações dos projetos de digitalizações: de que os textos eletrônicos de hoje sejam legíveis daqui alguns anos; em como convencer os administradores do valor ganho com a preservação digital; e que o acervo digital seja divisor de público, principalmente em países em desenvolvimento.

Cavalcante (2007, p. 159) deixa o cenário mais nebuloso quando diz que "a preocupação em preservar aquilo que se produz ou se transfere para o ciberespaço e pode ser considerado como patrimônio digital possui ainda muitas questões e poucas respostas".

Estas questões mostram o quanto à preservação de documentos eletrônicos é tão ou mais complexo que a preservação de documentos físicos.

2.2 O PROCESSO DE DIGITALIZAÇÃO

O termo digitalização pode ser aplicado em diferentes contextos e com relação a diferentes materiais, mas basicamente se refere ao ato de transformar um objeto físico num objeto digital. A utilização mais comum do termo esta relacionada à operação de escanear material impresso, criando arquivos digitais que os representem num sistema eletrônico.

Neste cenário, Puglia (2000, tradução nossa) descreve o processo assim:

> Digitalização converte uma imagem numa serie de elementos de imagem ou pixels, pequeno quadrados que são preto ou branco (binário), um tom especifico de cinza (escala de cinza) ou colorido. [..] Os pixels são arranjados dentro de uma matriz bidimensional chamada bitmap. Isto é referido como uma imagem rasterizada.

Desta abrangência, iremos lidar aqui com a transformação de material impresso em objetos digitais, ou seja, a transformação de documentos impressos em documentos digitais. Representante digital é o termo utilizado pelo CONARQ (2010, p. 5) para identificar este objeto digital resultante da digitalização de documentos não digitais.

Com relação à digitalização, o CONARQ (2010, p. 5) descreve este como um processo que converte os documentos arquivísticos para o formato digital, que consiste em unidades de dados binários.

As razões para que um documento seja convertido do suporte físico para o formato digital são diversas. Para Rikowski (2011, p. XII, tradução nossa):

> Muitas das vantagens da digitalização delineadas incluem a habilidade de ser possível pesquisar, navegar e comparar uma variedade de materiais; desenvolvendo substitutos digitais de raros e frágeis objetos originais; trazendo coleções unidas numa virtual forma digitalizada, que do contrário não poderia ser adquirida junto e elevando o perfil e prestigio de uma organização.

Existem diversos fatores que levam uma organização a digitalizar seus documentos. A informação arquivada em papel dificulta seu acesso em termos de pessoas autorizadas, localidade e tempo, restringindo a disseminação, pois ela está em um único local fisicamente. Esta mesma informação convertida em formato digital, permite um acesso quase irrestrito ao seu conteúdo. Como lembra Lévy (1999, p. 34), a transmissão de informações digitais pode ser feita por todas as vias de comunicação imagináveis.

Governos e organizações passaram a utilizar os documentos em formato digital, aproveitando as vantagens como a acessibilidade. Esta transição gerou também mudança no escopo dos profissionais que lidam com a informação dentro das organizações.

Neste cenário de mudança, Aquino (2004, p. 9) destaca:

> As conexões da informática com a telemática têm sido responsáveis pelo surgimento da informação em diferentes formatos de acesso e uso dessa informação. A passagem da cultura impressa para a cultura digital afetou não só os ambientes do papel, exigindo-lhes não só sua adequação aos novos formatos, mas impondo a aquisição de novas competências e habilidades para o desenvolvimento dos serviços informacionais.

A evolução da telemática, que pode ser entendida como a junção das tecnologias de informática e telecomunicações, adicionou novos desafios aos profissionais que trabalham com a organização e difusão da informação. Novos campos de estudo como protocolos de comunicação e banco de dados foram acrescentados ao ambiente de formação dos cursos da área da ciência da informação.

Esta mudança do suporte impresso ao digital não é apenas quanto aos meios de produção. É também relacionado ao armazenamento da informação, pois há uma constante urgência em digitalizar os documentos em suporte de papel não visando somente à preservação, mas também a acessibilidade desta informação.

Cunha (2008, p. 7) é direto quanto à acessibilidade dos arquivos digitais: “o armazenamento digital amplia as possibilidades de pontos de acesso a um determinado documento”. Esta ideia de facilidade também compartilhada por Ferreira (2006, p. 17), quando ele coloca que a simplicidade pela qual um material digital é criado e disseminado pelas modernas redes de comunicação.

Outro aspecto relevante se refere às possibilidades de edição do documento digital. Como salientado por Tammaro e Salarelli (2008, p. 13):

> Um dos pontos fortes do documento digital, quando comparado com o documento tradicional, é a possibilidade de ser formalmente manipulado, de ser desmontado e remontado em mil combinações diferentes sem jamais perder a possibilidade de manter intato o original.

O processo de captura digital da imagem do material é chamado de escaneamento. O equipamento utilizado para esta tarefa varia conforme o tipo de suporte do material que será digitalizado. Conforme elencado pelo CONARQ (2010, p. 9), os equipamentos de captura digital dividem-se em:

a) *Scanner* de mesa (*flat bed*);
b) *Scanners* planetários;
c) Câmeras digitais;
d) Equipamento de digitalização de negativos e diapositivos;
e) Equipamento para digitalização de microformas;
f) *Scanners* de produção e alimentação automática.

Além dos tipos de *scanners* listados pelo CONARQ, existem outros tipos variados, entre eles o *scanner* para grandes formatos (chamado também de *scanner* A0).

Para cada equipamento e *software* utilizado no processo de digitalização, são necessários cuidados com as configurações escolhidas. Estas são decididas com base no material que será scaneado.

Kenney e Chapman (2001, p. 7) destacam estes cuidados:

> As maneiras de se determinar os requisitos de qualidade da imagem variam com a gama de documentos a serem convertidos e com os processos utilizados para o escaneamento. Diferentes tipos de documentos requerem diferentes abordagens.

Assim, para que sejam produzidos representantes digitais o mais fiel possível aos equivalentes físicos, são necessários capturas de qualidade que possam ser classificadas como matrizes digitais. Para isso, a correta relação do tipo de documento com as configurações de *software* e *hardware* é essencial.

2.3 PRESERVAÇÃO DIGITAL A LONGO PRAZO

Após o surgimento do computador na década de 1970 e a sua utilização cada vez maior em setores públicos e privados da sociedade, acreditou-se que muitos problemas relacionados à organização e manutenção da informação estariam solucionados. A infinidade de tipos de *hardware*, *softwares*, redes e formatos de arquivos digitais davam a falsa impressão de que a informação armazenada em forma de bits estaria protegida da ação do tempo.

Porém, com o advento da tecnologia da informação, novos problemas surgiram, criando desafios talvez maiores que os anteriores já enfrentados pelos profissionais da área da Ciência da informação. Como ressalta Rondinelli (2005, p. 23) “de fato, a teia construída pela tecnologia da informação tem implicações econômicas politicas, sociais e culturais que a explicam, ao mesmo tempo em que geram novas implicações econômicas, politicas, sociais e culturais”.

Surge assim à problemática da preservação digital. Em seu artigo “*A Digital Dark Ages*” publicada durante a conferência da *International Federation of Library Associations* (IFLA) no ano de 1997, Terry Kuny lançava então um aviso à comunidade de profissionais da informação, sobre as consequências da falta de padrão no arquivamento e manuseio do suporte eletrônico da informação. Num cenário onde a evolução tecnológica começava a dar passos mais rápidos, Kuny (1997, p. 1, tradução nossa) já alertava “Um destes impactos é como nós estamos preservando os registros históricos numa era

eletrônica onde mudança e velocidade são valorizadas mais do que conservação e longevidade".

A Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) também trata do assunto no manifesto intitulado *Guidelines for the Preservation of Digital Heritage* (traduz-se como *Diretrizes para a preservação da herança digital*). No capitulo sete deste manifesto a preservação digital é definida da seguinte forma: "Preservação digital é usado para descrever o processo envolvido em manter informações e outros tipos de herança que existem em formato digital.".

Sobre a fragilidade e integridade da informação em formato digital, Conway (2001, p. 24) alerta:

> Os meios de armazenagem digital devem ser manuseados com cuidado, mas eles sobreviverão por mais tempo do que a capacidade dos sistemas de recuperar e interpretar os dados neles armazenados. Como não poderemos prever ao certo quando um sistema se tornará obsoleto, bibliotecas devem estar preparadas para fazer migrar dados importantes de imagens, índices e programas para as futuras gerações de tecnologia.

Conway (2001, p. 28) ainda adiciona outro aspecto sobre obsolescência tecnológica quando diz que "[...] o compromisso do distribuidor de dar suporte e manter um sistema antigo é inversamente proporcional a sua capacidade de vender um novo sistema.".

O termo preservação digital é recente nos meios da ciência da informação, sua descrição também é pouco tratada. Marcondes et al (2006, p. 21) tratam o termo preservação digital da seguinte maneira: "Conjunto de ações técnicas, gerenciais e administrativas destinadas a manter a integridade e a acessibilidade de objetos digitais de valor continuo, pelo tempo que transcenda as mudanças tecnológicas.". Outra definição mais simples é dada pela *American Library Association* (ALA, 2008): preservação digital combina politicas, estratégias e ações que garantam acesso ao conteúdo digital ao passar do tempo.

O tema da preservação digital ainda não possui a atenção necessária e isto se traduzirá em problemas futuros. Marcondes et al (2006, p. 114) afirmam, que a compulsão em produzir informações digitais é infinitamente superior a nossa capacidade de preservar e oferecer acesso a estas informações.

A própria noção errada de preservação interfere na compreensão comum da sociedade, sobre o que é preservar um objeto que não existe fisicamente. Esta concepção errônea é confirmada por Besser (2010):

> Para prevenir novas perdas, precisamos entender os problemas de longevidade do mundo digital. Precisamos pensar que a preservação no mundo digital difere daquela a que estamos acostumados no mundo analógico. No mundo analógico, todos os esforços de preservação enfocam a obra, o objeto como artefato. Quando começamos a nos engajar na preservação da informação em formato digital, foi preciso dar um salto conceitual, mudando o enfoque na preservação física do objeto para a preservação do conteúdo informativo, que pode estar completamente dissociado de qualquer artefato físico.

Desta questão, parte a ideia incorreta de que o simples ato de armazenar digitalmente a informação em algum suporte moderno é o suficiente para assegurar seu futuro acesso.

A variedade de formatos de arquivos eletrônicos existentes é outro agravante. Para Ferreira (2006, p. 39):

> Existe um leque variado de opções no que diz respeito a formatos para representação de imagens bidimensionais (e.g. BMP, GIF, JPEG, PNG, TARGA). Se durante o processo de ingestão, todas as imagens digitais forem convertidas para um único formato, futuras intervenções ao nível da sua preservação poderão ser realizadas de forma mais simples e, Consequentemente, mais económica.

Neste âmbito, o profissional bibliotecário se situa como um dos principais atores no cenário da preservação digital, em paralelo a outros como profissionais da tecnologia da informação.

As diferenças nestes papeis são exemplificadas por Márdero Arellano (2006):

> Preservação digital tem diferentes significados dependendo do contexto, para os profissionais da informação, por exemplo, pode ser a infraestrutura e o comprometimento institucional necessário para proteger a informação representada digitalmente, em quanto que para os especialistas das ciências da computação ela seria uma maneira de atenuar a obsolescência tecnológica e aumentar a memória humana.

A concepção de preservação digital também possui situações que seguem na mesma direção dos ideais de preservação. Manter inalterado o conteúdo do objeto digital parece ser o principal objetivo de todo o processo. Sob o ponto de vista de Marcondes (2006, p. 118):

> O próprio sentido conceitual da preservação, no contexto da informação digital, esta imerso em um paradoxo: tradicionalmente preservar algo significa mantê-lo imutável e intacto; entretanto, no ambiente digital, preservar significa, na maioria dos casos, mudar, recriar, renovar: mudar formatos, renovar mídias, hardware e software.

Portanto há uma busca por diretrizes que possibilitem ao profissional da ciência da informação conhecer os limites, para alterar o suporte digital sem o risco de alterar a informação contida dentro dele.

Sobre os documentos digitais, Lacombe e Silva (2007, p. 115) acreditam que o grande desafio é garantir a produção de documentos confiáveis e que haja manutenção da autenticidade e acesso a longo prazo.

2.4 O FORMATO PDF/A

O formato de arquivo *Portable Document Format* (PDF), que traduz-se como Formato de Documento Portátil, foi desenvolvido pela *Adobe Systems* na década de 1990. Sua principal característica foi a de representar documentos de maneira independente em relação à plataforma utilizada (neste caso o conjunto *software* e *hardware*). Isto ocorre devido a cada arquivo em PDF conter em si uma descrição completa do conteúdo, como seu layout, texto, gráficos, fontes e outras informações que forem necessárias para a sua visualização. Após alguns anos de desenvolvimento progressivo e evolução, o formato virou praticamente um padrão dentro das corporações e organizações, se tornando inclusive no ano de 2008 um padrão aberto conforme a norma ISO 32000.

Identificado como *Portable Document Format for Arquiving* (Formato de Documento Portátil para Arquivamento), o formato PDF/A foi desenvolvido pela empresa Adobe Systems como um subconjunto do formato padrão PDF. Publicado em 2005 como a norma ISO 19005 para padrão de arquivamento para preservação em longo prazo, o formato PDF/A possui especificações para a criação, visualização e impressão de documentos PDF com o objetivo de manter embutidos os arquivos necessários.

Como advertido pela própria Adobe (ADOBE, 2010), o formato PDF/A não define nenhuma estratégia de arquivamento ou objetivos de um sistema de arquivamento.

Em paralelo ao formato PDF/A, há outras variações do formato PDF visando diferentes funções. No quadro 1 temos uma tabela com os diferentes padrões atuais de PDF e suas características.

Quadro 1 – Relação dos padrões existentes de PDF.

PADRÕES DE PDF		
Especificação	**Proposito**	**Descrição**
PDF ISO 32000	**O padrão de cobertura** PDF "Convencional" para uma ampla gama de uso.	Contém a especificação PDF completa e substitui a edição 1.7 da referencia Adobe PDF. Este padrão se tornara a fundação para todas as futuras gerações dos padrões derivados.
PDF/A ISO 19005	**Arquivamento** Gestores de arquivos, arquivistas, gerentes de conformidade.	Fornece especificações para a criação, visualização e impressão de documentos digitais usados para preservação a longo prazo. PDF/A preserva e documentos finais de registros como arquivos autônomos. Ele não permite referencias a conteúdos externos, pois estes itens podem não existir no futuro. PDF/A1 é baseado na versão 1.4 do PDF.
PDF/E ISO 24517	**Engenharia** Arquitetos, engenheiros, profissionais da construção, equipes de fabricação de produtos.	Fornece especificações para a criação, visualização e impressão de documentos usados em fluxos de trabalho de engenharia. PDF/E facilita o intercambio de documentações e desenhos para compartilhar com outros dentro da uma cadeia de fornecimento ou linha de analise e conferencia. Ele especifica configurações de PDF adequadas para fluxos de montagem e fabricação, fluxos de trabalhos geoespaciais e suporta mídias interativas, incluindo animações e 3D. PDF/E é baseado na versão 1.6 do PDF.

PDF/X ISO 15930	**Produção de impressão** Profissionais de impressão, designers gráficos, profissionais de criação.	Fornece especificações para a criação, visualização e impressão de paginas de arquivos finais de prensa. PDF/X fornece diretrizes para configurações de PDF que afetam aspectos críticos de impressão, como espaço de cor e captura. Ele também restringe outros conteúdos – como multimídia integrada – que não serve diretamente a produção de impressão de alta qualidade.
PDF Healthcare	**Área da Saúde** Fornecedores da área da saúde e consumidores.	Fornece melhores praticas e diretrizes de implementação para facilitar a captura, intercambio, preservação e proteção de informação da área da saúde. Seguinte a estas diretrizes é fornecido um contêiner eletrônico seguro que pode armazenar e transmitir informações médicas incluindo documentos pessoais, dados XML, dados e imagens DICOM, notas clinicas, relatórios de laboratórios, formulários eletrônicos, imagens escaneadas, fotografias, raio-x digital e eletrocardiograma.
PDF/UA ISO 14289	**Acesso universal** Pessoas com deficiências, Gestores de TI em organizações governamentais ou comerciais, gestões de conformidade.	Fornece um conjunto de diretrizes para a criação de arquivos em PDF que sejam universalmente acessíveis. Arquivos PDF/UA aumentam a legibilidade de documentos para pessoas com deficiência, comprometimento da visão ou locomoção limitada. Estas diretrizes podem ser usadas em conjunto com uma larga variedade de outras configurações para a criação de arquivos PDF.
PDF/VT ISO 16612-2	**Impressão variável e transacional** Profissionais de impressão.	Fornece especificações para a criação, visualização e impressão de extratos bancários e faturas de negócios, usados dentro da uma indústria de impressão variável e transacional.

Fonte: Adobe Systems Incorporated, 2012.

Apesar de compartilhar algumas características com os outros padrões baseados no formato PDF, o formato PDF/A é o único voltado especificamente para o arquivamento e preservação digital a longo prazo.

O formato PDF/A possui a vantagem de ser um padrão aberto. Para o CONARQ (2011, p. 35) “a adoção de formatos digitais abertos configura-se, adicionalmente, como medida de preservação recomendável e necessária.”.

O padrão PDF/A é dividido em partes, todas de acordo com as normas correspondentes liberadas pela ISO. A parte 1 liberada em 2005 com a norma

ISO 19005-1 possui dois níveis de conformidades, enquanto que a parte 2 liberada no ano de 2011 e a parte 3 liberada no ano de 2012 possuem três níveis de conformidades cada uma respectivamente No quadro 2 temos a relação das partes disponibilizadas e algumas características.

Quadro 2 – Relação das partes do PDF/A.

Partes do PDF/A			
Nome	**PDF/A-1**	**PDF/A-2**	**PDF/A-3**
Parte	Parte 1	Parte 2	Parte 3
Norma ISO	ISO 19005-1:2005	ISO 19005-2:2011	ISO 19005-3:2012
Ano de liberação	2005	2011	2012
Versão do PDF	Baseado na versão 1.4	Baseado na versão 1.7	Baseado na versão 1.7
Níveis de conformidade	PDF/A-1a PDF/A-1b	PDF/A-2a PDF/A-2b PDF/A-2u	PDF/A-3a PDF/A-3b PDF/A-3u

Elementos de compatibilidade	- Fontes precisam estar embutidas. - Conteúdo de áudio e vídeo proibidos. - Encriptação proibida. - Links externos proibidos. - Compressão LZW e JPEG2000 proibidos. - Transparência e camadas proibidas. - Javascript e arquivos executáveis são proibidos.	- Fontes OpenType pode ser embutidas. - Disponibilidade de uso de assinatura digital do padrão PAdES. - Permitido uso de compressão JPEG2000. - Permitido uso de transparências e camadas. - Possibilidade de embutir arquivos PDF/A.	- Possibilidade de embutir diversos arquivos (como planilhas, documentos de texto, XML, CAD, CSV,..).

Fonte: Dados da pesquisa (2012)

O padrão PDF/A-1 possui dois níveis de conformidade estipulados pela norma ISO 19005-1:2005. Estes dois níveis são chamados Nível A e Nível B de conformidade.

A ISO (2005, tradução nossa) define que os arquivos de conformidade Nível A tem que atender a todos os requerimentos que determinam a representação de um documento eletrônico PDF/A-1. Já os arquivos de Nível B de conformidade tem a pretensão de assegurar a aparência visual do documento, permitindo que o documento deixe de atender a requerimentos como estrutura do PDF e caracteres Unicode.

As demais diferenças entre os dois níveis de conformidade podem ser vistos no quadro 3, a seguir.

Quadro 3 – Características dos níveis do PDF/A-1.

	ISO 19005-1:2005: PDF/A-1a (Nível A)	**ISO 19005-1:2005: PDF/A-1b (Nível B)**
Conformidade	Conformidade completa do PDF/A	Conformidade restrita do PDF/A
Objetivo	Produzir arquivos em PDF com acesso completo a todo o conteúdo.	Produzir arquivos em PDF que somente garantem reprodução visual.
Versão do PDF	PDF 1.4	

Identificação de PDF/A	Usuários são obrigados a declarar a identificação de PDF/A e o nível de conformidade.	
Metadados	Especificações como autor, título do documento, data de criação e o programa fonte precisam ser compatíveis com padrão XMP.	
Estrutura lógica	Estrutura e acessibilidade precisam ser criados com o uso de marcação e descrição de imagens alternativas e declaração da linguagem utilizada.	Não há requerimentos explicito de estrutura logica.
Encriptação	Configurações de segurança são proibidas. Tem que ser possível abrir/processar o arquivo PDF em questão sem requisitar uma senha.	
Cores	Todas as cores precisam estar identificadas. Sistema de cores dependente de equipamento precisa ser identificado com o proposito de saída.	
Transparência	Não permitido.	
Camadas do PDF	Não permitido.	
Compressão	Compressão LZW não permitida. Compressão JPEG2000 não permitida.	
Fontes	Todas as fontes precisam (pelo menos como subconjuntos) estarem disponíveis (embutidas) diretamente dentro do documento PDF em questão.	
	Mapeamento dos códigos de caracteres para glifo devem ser inequívocos.	
	Cada letra precisa ter um Unicode equivalente.	-
Anotações	Comentários na forma de som ou filmes não são permitidos. Anotações no estilo texto/rotulo são permitidas.	
Referencia de conteúdo	Conteúdo de imagens ou páginas referenciadas (não embutidas) não são permitidas.	
Imagens alternativas	Imagens alternativas (para monitores com tela de baixa resolução) não são permitidas.	

Linguagens de programação	*Javascript* embutido não é permitido.
Ações	Certas ações, como abrir vídeos, sons ou enviar e resetar formulários não são permitidos.
Formulários	Permitido, mas com restrições.

Fonte: Drummer, Oettler, Von Seggern, 2007.

Apesar das características direcionadas ao proposito de arquivamento e preservação em longo prazo, o formato PDF/A ainda tem recebido uma adoção lenta dentro das organizações. Numa pesquisa realizada no ano de 2010 por Doug Miles (2010), diretor da organização *AIIM Market Intelligence*, com 144 pessoas da área de arquivos, foi questionado quais eram os formatos de arquivos eletrônicos mais utilizados para armazenar uma proporção significativa dos registros das suas respectivas organizações. O resultado foi de que o formato PDF/A foi o menos utilizado, em comparação com os outros formatos presentes na pesquisa como PDF, TIFF, JPEG e HTML.

2.5 O FORMATO TIFF

O formato TIFF teve suas especificações publicadas inicialmente ainda no ano de 1986. Desenvolvido pela empresa *Aldus Corporation*, o formato TIFF tem sua identificação da nomenclatura *Tagged Image File Format* (que se traduz como Formato de Arquivo de Imagem Marcado) e tem como função primaria armazenar imagens. No ano de 1994, a empresa *Aldus Corporation* acabou por se fundir com a *Adobe Systems Incorporated*.

A definição feita pela ADOBE (1992, p. 4, tradução nossa) sobre o formato TIFF elenca os seguintes pontos:

a) TIFF descreve dados de imagem que tipicamente originam-se de *scanners*, digitalizadores de vídeo e programas de retoque de pinturas e fotos.

b) TIFF não é uma linguagem de impressão ou de descrição de páginas. O proposito do TIFF é descrever e armazenar dados rasterizados de imagem.
c) A meta primaria do TIFF é fornecer um ambiente rico no qual as aplicações podem trocar dados de imagem. Esta riqueza é necessária para tomar vantagem das variáveis capacidades dos *scanners* e outros dispositivos de imagem.
d) Embora o TIFF seja um formato complexo, este pode facilmente ser usado por *scanners* e aplicações simples, bem como devido ao baixo número de requerimentos.
e) TIFF será melhorado numa base continua conforme novas necessidades em imagem surjam. Uma alta prioridade tem sido dada para estruturar o TIFF com o proposito de que futuras melhorias possam ser adicionadas sem causar dificuldades desnecessárias aos desenvolvedores.

Apesar de o formato TIFF ter tido duas variações que se tornaram norma ISO, nenhuma delas é utilizada no processo de digitalização de documentos. Estes dois padrões são o TIFF/EP (norma ISO 12234, publicada no ano de 2001) utilizado para fotografia eletrônica e o TIFF/IT (norma ISO 12639, publicada em 2004) utilizada para sistemas de pré-impressão de alto nível.

Mesmo sem uma padronização especifica para armazenamento, o formato TIFF é utilizado como padrão de saída dos *scanners* disponíveis no mercado.

Dentro dos governos, empresas e organizações, são utilizados vários tipos de formatos de arquivo para o processo de armazenamento de documentos eletrônicos. Muitos são utilizados de maneiras inadequadas devido a falta de características de preservação digital. No quadro 4 é possível observar uma comparação das características de alguns formatos utilizados para arquivamento digital, entre eles os formatos PDF/A e TIFF.

Quadro 4 – Comparação de formatos de arquivo para armazenamento.

	PDF/A	**XPS**	**TIFF-G4**	**JPEG**	**DOC (Word)**
Padrão ISO para arquivamento	Sim	Não	De fato um padrão, porém não um	Não	Não

			padrão oficial.		
Qualidade de fac-símile	Sim	Sim	Sim	Sim	Não
Segurança de fonte	Sim. Nível máximo de segurança graças a restritas especificações definidas dentro do padrão PDF/A.	Sim. Fontes são embutidas.	Não existem fontes, desde que os arquivos são imagens em pixels.	Não existem fontes, desde que os arquivos são imagens em pixels.	Não. A exibição de fontes pode variar em diferentes computadores. Usuários não são avisados sobre fontes perdidas. Usuários não são avisados sobre fontes substituídas.
Texto pesquisável	Sim. Ativado por meio de OCR, inclusive para textos em paginas escaneadas.	Sim	Possível se o texto pesquisável é gerado usando OCR. Não há procedimento padrão para armazenar o texto dentro do arquivo TIFF-G4.	Não	Sim
Cores consistentes	Sim. Cores consistentes são exigidas pelo padrão.	Possível	Não. Produz mapa de bits em preto e branco.	Possível	Não
Imagens e gráficos são partes fixas dos documentos	Sim	Sim	Sim. Eles são incorporados dentro da imagem pixelizada.	Sim. Eles são incorporados dentro da imagem pixelizada.	Não. Não pode ser sempre manuseado com segurança.
Dados estruturados	Sim – com PDF/A-1a com PDF marcado.	Sim. Com XML.	Não	Não	Possível
Multi-plataforma	Sim	Não	Sim	Sim	Somente com restrições (problema com fontes)
Visualizador gratuito	Sim. PDF/A é sempre exibido da mesma maneira.	Sim (atualmente somente para Windows).	Sim (mas não largamente utilizado).	Sim. Maioria dos navegadores de internet podem exibir JPEG.	Sim. Há alternativas gratuitas ao Office, mas os documentos podem ser exibidos diferentemente.

Fonte: DRUMMER; OETTLER; VON SEGGERN, 2007.

O formato TIFF é ainda um dos mais utilizados para a tarefa de arquivamento de documentos, como destacado por Drummer; Oettler; Von Seggern (2007, p. 9, tradução nossa):

> Por um longo tempo, muitas autoridades publicas e companhias que precisam armazenar grandes quantidades de correspondências, registros, faturas, contratos e informações similares em arquivos digitais tem usado o formato de imagem em pixel TIFF (Tagged Image File Format).

A questão de qualidade de imagem do formato TIFF também é colocada pelo CONARQ (2010, p. 13) como uma razão para adoção do formato:

> O formato mais utilizado para os representantes digitais matrizes é o formato TIFF (Tagged Image File Format), que apresenta elevada definição de cores sendo amplamente conhecido e utilizado para o intercâmbio de representantes digitais entre as diversas plataformas de tecnologia da informação existentes.

O termo matriz digital usado pelo CONARQ refere-se basicamente a imagem original que é capturada do documento físico. É utilizada a denominação matriz, pois a partir desta captura poderá ser recriado digitalmente o documento em outros formatos e configurações, de acordo com as necessidades.

3 PROCEDIMENTOS METODOLÓGICOS

A opção por pesquisar um determinado assunto ocorre quando não se encontra facilmente as informações sobre o tema ou estas informações são insuficientes para oferecer uma noção ampla. Mas o processo de pesquisa necessita da uma metodologia para ter forma.

Segundo Marconi e Lakatos (1992, p. 83):

> Todas as ciências caracterizam-se pela utilização de métodos científicos; em contrapartida, nem todos os ramos de estudo que empregam métodos são ciências. Dessas afirmações podemos concluir que a utilização de métodos científicos não é da alçada exclusiva da ciência, mas não há ciência sem o emprego de métodos científicos.

A metodologia é um dos aspectos essenciais de qualquer pesquisa a ser realizada. Sem sua determinação prévia, o progresso do trabalho da pesquisa ficará comprometido.

3.1 TIPO DE PESQUISA

O presente estudo é uma pesquisa de ordem exploratória, sendo o procedimento metodológico escolhido o de pesquisa bibliográfica. De acordo com Gil (2002, p. 44), a pesquisa bibliográfica é desenvolvida com base em material já elaborado, constituído principalmente de livros e artigos científicos. Esta escolha deu-se devido ao intuito de buscar na literatura da área os conceitos do tema preservação digital, ajudando a compreender a questão abordada. Outro ponto relevante na escolha do tipo de pesquisa foi à dificuldade em encontrar literatura sobre o formato PDF/A.

3.2 ANÁLISE DOS DADOS

A análise dos dados foi de ordem quali-quantitativa. A pesquisa qualitativa tem o ideal de descrever a codificar diferentes significados do material analisado. Segundo Godoy (1995):

> De maneira diversa, a pesquisa qualitativa não procura enumerar e/ ou medir os eventos estudados, nem emprega instrumental estatístico na análise dos dados. Parte de questões ou focos de interesses amplos, que vão se definindo a medida que o estudo se desenvolve. Envolve a obtenção de dados descritivos sobre pessoas, lugares e processos interativos pelo contato direto do pesquisador com a situação estudada, procurando compreender os fenômenos segundo a perspectiva dos sujeitos, ou seja, dos participantes da situação em estudo.

Ainda Godoy (1995), afirma que “os pesquisadores qualitativos estão preocupados com o processo e não simplesmente com os resultados ou produto.”.

Quanto à pesquisa quantitativa, Silva e Menezes (2005, p. 20) a definem assim:

> [...] considera que tudo pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las. Requer o uso de recursos e de técnicas estatísticas (percentagem, média, moda, mediana, desvio-padrão, coeficiente de correlação, análise de regressão, etc.).

Para realizar a análise comparativa dos formatos TIFF e PDF/A, foram escolhidas três obras dentro do projeto de digitalizações da Biblioteca Central da Universidade Federal de Santa Catarina para serem escaneadas e convertidas para ambos os formatos. Estas três obras escolhidas são compostas de uma dissertação bitonal (preto e branco) com prominência de textos e tabelas, uma dissertação colorida com texto e diversas fotografias/figuras e outra sendo um mapa em formato A0.

Após o processo de digitalização, as obras foram convertidas através do uso do *software Abbyy FineReader 11 Corporate Edition*. Alguns recursos e características do *software Abbyy FineReader 11 Corporate Edition* estão apresentados em manual contido no apêndice desta pesquisa.

Finalizado a conversão das obras, dados referentes ao tamanho dos arquivos eletrônicos foram tabelados e calculados para obterem-se informações precisas.

4 RESULTADOS

Neste capitulo foi analisado os dados da pesquisa com base em informações levantadas no projeto de digitalização das teses e dissertações da Biblioteca Central da Universidade Federal de Santa Catarina. Para isso inicialmente foi selecionado três tipos diferentes de obras inseridas no acervo de teses e dissertações, que pudessem representar de forma geral o universo das obras digitalizadas. Também foi realizado um levantamento da produção do projeto de digitalização, a fim de selecionar um período mensal e criar uma media nos critérios de consumo de armazenamento de dados.

4.1 SELEÇÃO DAS OBRAS

Dentre todas as obras que compõem o acervo de teses e dissertações da Biblioteca Central da UFSC, foi escolhido três tipos de obras que pudessem representar a variedade de material do acervo. Estes três tipos de obras são listadas abaixo:

a) Obra composta de textos e tabelas, com impressão bitonal (preto e branco).
b) Obra composta de textos, figuras e fotografias, com impressão colorida.
c) Obra composta de mapas em grandes formatos, impressão mista.

Com bases nestas características, foram selecionadas as três obras para serem escaneadas em diferentes tipos de configurações de captura. Os detalhes destas três obras são detalhados a seguir:

a) Obra 01: dissertação confeccionada no ano de 1999; possui 217 páginas; composta de texto e 96 tabelas; impressão bitonal (preto e branco).
b) Obra 02: tese confeccionada no ano de 2002; possui 151 páginas; composta de texto e 72 fotografias; impressão colorida.

c) Obra 03: mapa geológico do município de Florianópolis, na escala de 1:50000, confeccionado no ano de 1990 pelo IBGE (SC).

4.2 CAPTURA E CONVERSÃO

O processo de captura foi realizado num *scanner* de produção modelo i4600 fabricado pela empresa *Kodak*, em conjunto com o *software* da mesma empresa, intitulado *Kodak Capture Desktop*. As configurações de capturas foram selecionadas com base na tabela de recomendações de digitalização publicada pelo CONARQ (2010, p. 17).

As configurações de captura selecionadas para cada uma das obras de exemplo são descritas assim:

a) Obra 01: densidade de 100dpi a 600dpi; tonalidade bi-tonal (1 bit); compressão Group 4 (G4).
b) Obra 02: densidade de 100dpi a 600dpi; tonalidade colorida (24 bit); compressão JPEG.
c) Obra 03: densidade de 150dpi a 600dpi: tonalidade bi-tonal, escala de cinza e colorida; compressão RLE.

A conversão destas capturas para o formato PDF/A foi realizado através do *software FineReader 11 Professional Edition* (compilação 11.0.102.583), produzido pela empresa *Abbyy*.

Com o objetivo de realizar uma comparação mais ampla, além do formato PDF/A, as capturas no formato TIFF também foram convertidas para o formato PDF ISO 32000.

As diferentes características entre os formatos TIFF, PDF e PDF/A estão listadas no quadro 5.

Quadro 5 – Relação dos formatos TIFF, PDF e PDF/A.

Comparativo dos formatos TIFF, PDF e PDF/A			
	TIFF	**PDF**	**PDF/A**
Versão mais recente	6.0 (Suplemento 2) 2002	1.7 Extensão nível 8 2011	PDFA Parte 3 2012

Padrão aprovado pela ISO	Possui, mas não utilizados em digitalização.	ISO 32000	ISO 19005
Visualizador gratuito	Sim	Sim	Sim
Acessível em ambiente multi-plataforma	Sim	Sim	Sim
Texto pesquisável	Não. Somente através de *software* especifico que realize OCR no arquivo.	Sim	Sim
100% das informações incorporadas	Não	Não	Sim
Fontes embutidas	Não	Sim (não obrigatório)	Sim (obrigatório)
Livre de licenças	Sim	Sim	Sim
Documento estruturado	Não	Sim	Sim
Encriptação	Não	Sim	Não (Proibido)
Metadados	Sim (simples)	Sim. Usa padrão XMP.	Sim. Usa padrão XMP.
Compressão	JPEG, CITT, RLE, LZW	JPEG, JPEG2000, CCITT, LZW	JPEG2000 (somente no padrão PDF/A-2)

Fonte: Dados da pesquisa (2012)

A configuração de conversão utilizada no *software FineReader* foi a de manter intacta a resolução original da captura, sem ajustes de páginas ou ruídos. A compressão utilizada foi a Group 4 (G4) para as capturas bitonais e JPEG para as capturas em escala de cinza e coloridas.

4.3 ESPAÇO DE ARMAZENAMENTO

Realizou-se a analise do espaço de armazenamento ocupado pelas obras de exemplo usando-se os formatos TIFF, PDF e PDF/A. Com a finalidade de atingir o objetivo proposto, verificaram-se as diferenças geradas

por cada uma das escolhas. Optou-se por usar também dados do formato PDF ISO 32000 como ponto de referencia para o formato PDF/A.

Adverte-se aqui que para o formato PDF/A será utilizado o padrão de Nível 1b (conversão somente da imagem da capturada, sem OCR, estrutura e marcações).

No quadro 6 temos a comparação de espaço de armazenamento da obra 01, nos formatos TIFF, PDF e PDF/A-1b.

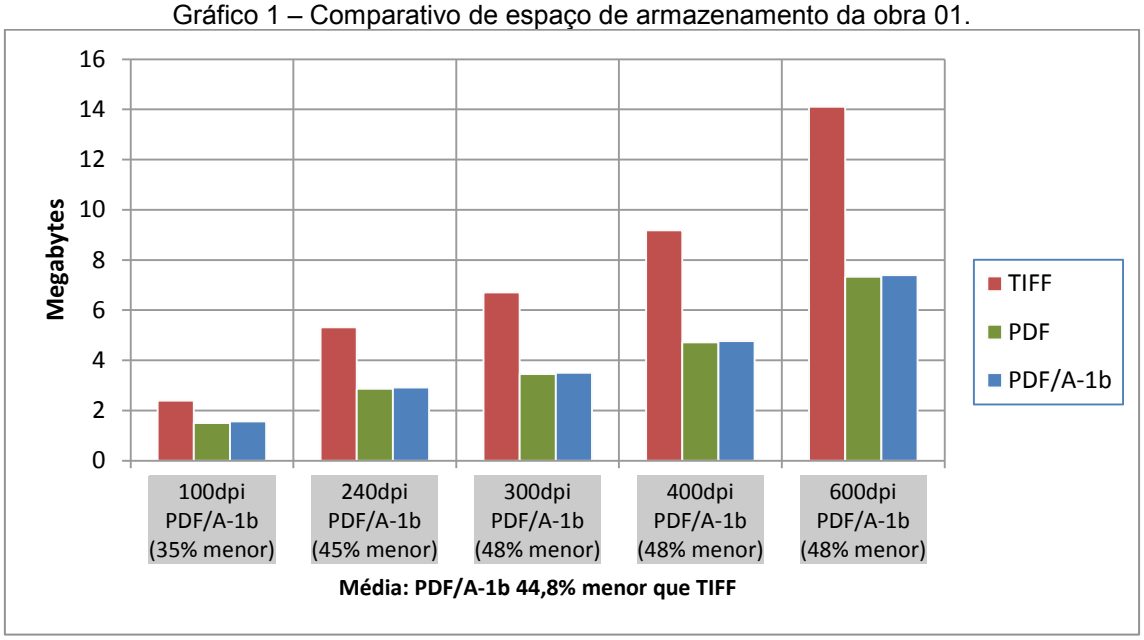
Quadro 6 – Relação de espaço de armazenamento da obra 01.

Obra 01 (A4 Bitonal) – TIFF/PDF/PDFA-1b			
Qualidade da captura	**Tamanho em TIFF**	**Tamanho em PDF**	**Tamanho em PDF/A1-b**
Densidade: 100dpi Tonalidade: bitonal (1 bit) Compressão: G4	2,40 MB	1,50 MB	1,56 MB
Densidade: 240dpi Tonalidade: bitonal (1 bit) Compressão: G4	5,31 MB	2,86 MB	2,92 MB
Densidade: 300dpi Tonalidade: bitonal (1 bit) Compressão: G4	6,70 MB	3,45 MB	3,50 MB
Densidade: 400dpi Tonalidade: bitonal (1 bit) Compressão: G4	9,18 MB	4,71 MB	4,77 MB
Densidade: 600dpi Tonalidade: bitonal (1 bit) Compressão: G4	14,1 MB	7,33 MB	7,39 MB

Fonte: Dados da pesquisa (2012)

No quadro 6 nota-se de imediato que os arquivos no formato PDF e PDF/A1b ocuparam um espaço menor com relação ao formato TIFF, apesar de todos os três formatos utilizarem a compressão de padrão G4. Esta diferença fica mais evidente ao olharmos o gráfico 1.

Gráfico 1 – Comparativo de espaço de armazenamento da obra 01.



Fonte: Dados da pesquisa (2012)

A partir dos dados obtidos, constatou-se que, a partir da captura em 100dpi de densidade, o arquivo gerado no formato PDF/A ficou 35% menor em relação ao formato TIFF. Já partindo da densidade em 240dpi, o formato PDF/A ficou 45% menor. Para as densidades em 300dpi, 400dpi e 600dpi, a diferença foi identifica: arquivos 48% menores em PDF/A em relação ao formato TIFF. Em média, o formato PDF/A ficou 44,8% menor em relação ao formato TIFF.

Realizaram-se os mesmos processos de conversão na obra 02, percebeu-se que as diferenças obtidas na obra 01 inverteram-se, resultado num maior espaço de armazenamento nos formatos PDF e PDF/A em comparação com o formato TIFF. Estes números podem ser vistos no quadro 7.

Quadro 7 – Relação de espaço de armazenamento da obra 02.

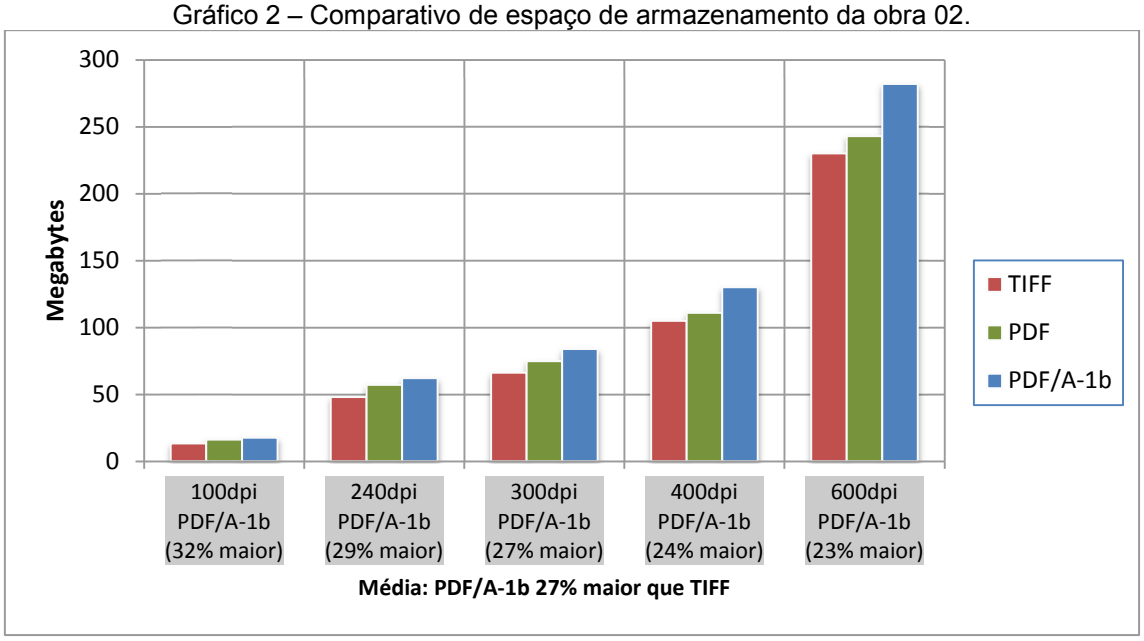
Obra 02 (A4 Colorido) – TIFF/PDF/PDFA-1b			
Qualidade da captura	**Tamanho em TIFF**	**Tamanho em PDF**	**Tamanho em PDF/A-1b**
Densidade: 100dpi Tonalidade: colorido (24 bit) Compressão: JPEG	13,3 MB	16,2 MB	17,6 MB

Densidade: 240dpi Tonalidade: colorido (24 bit) Compressão: JPEG	48,1 MB	57 MB	62 MB
Densidade: 300dpi Tonalidade: colorido (24 bit) Compressão: JPEG	66,1 MB	74,9 MB	83,8 MB
Densidade: 400dpi Tonalidade: colorido (24 bit) Compressão: JPEG	105 MB	111 MB	130 MB
Densidade: 600dpi Tonalidade: colorido (24 bit) Compressão: JPEG	230 MB	243 MB	282 MB

Fonte: Dados da pesquisa (2012)

Ao olharmos estas diferenças no gráfico 2, nota-se que no caso de obras capturadas em cores o espaço ocupado tanto pelo PDF quanto pelo PDF/A são maiores. Outro ponto a ser percebido é de que quanto maior a densidade da imagem, maior fica a diferença entre os três formatos abordados.

Gráfico 2 – Comparativo de espaço de armazenamento da obra 02.



Fonte: Dados da pesquisa (2012)

Comparado lado a lado, verificou-se que no caso da obra 02 capturada em 100dpi, o formato PDF/A-1b ficou 32% maior do que em formato TIFF. Em 240dpi de densidade, a diferença foi de 29% maior no formato PDF/A. Para

capturas em 300dpi, 400dpi e 600dpi, as diferenças ficaram em 27%, 24% e 23% respectivamente. Em média, o formato PDF/A-1b ficou 27% maior em relação ao formato TIFF.

Aqui se analisou as diferenças de tamanho de armazenamento para a obra 03, nos formatos TIFF, PDF e PDF/A-1b. No quadro 8 é demonstrada a relação entre os três formatos, com capturas escaneadas em duas densidades diferentes (300dpi e 600dpi) e em dois padrões de tonalidades diferentes: escala de cinza (8 bit) e colorido (24 bit). Não foi realizada a captura em tonalidade bitonal devido à extrema perda de informação visual que ocorreu com a imagem do mapa cartográfico.

Quadro 8 – Relação de espaço de armazenamento da obra 03.

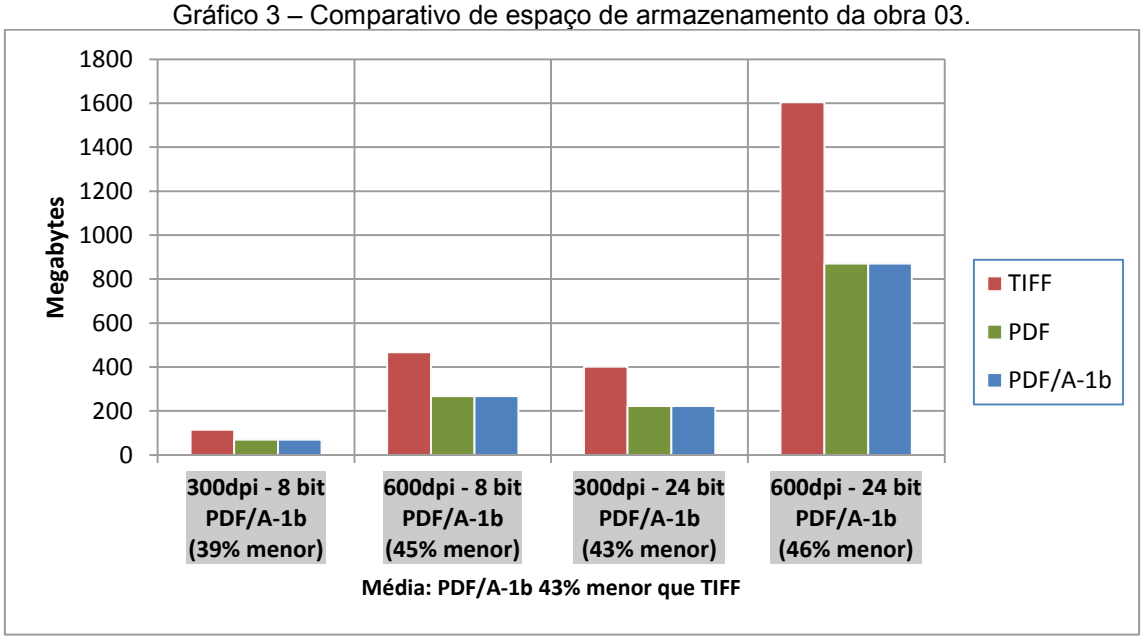
Obra 03 (Mapa A0) – TIFF/PDF/PDFA-1b			
Qualidade da captura	**Tamanho em TIFF**	**Tamanho em PDF**	**Tamanho em PDF/A-1b**
Densidade: 300 dpi Tonalidade: escala de cinza (8 bit) Compressão: RLE	115 MB	70,2 MB	70,2 MB
Densidade: 600 dpi Tonalidade: escala de cinza (8 bit) Compressão: RLE	467 MB	268 MB	268 MB
Densidade: 300 dpi Tonalidade: colorido (24 bit) Compressão: RLE	402 MB	223 MB	223 MB
Densidade: 600 dpi Tonalidade: colorido (24 bit) Compressão: RLE	1.604 MB	871 MB	871 MB

Fonte: Dados da pesquisa (2012)

Através do quadro 8 percebe-se que uma diferença significativa de tamanho entre o formato TIFF e PDF/A-1b. Esta condição dá-se devido às diferenças entre a compressão RLE do *scanner* A0 e a compressão JPEG adotada para o formato PDF/A-1b.

Uma visualização mais adequada da diferença pode ser observada no gráfico 3.

Gráfico 3 – Comparativo de espaço de armazenamento da obra 03.



Fonte: Dados da pesquisa (2012)

Utilizando-se uma captura em 300dpi de densidade, o formato PDF/A ficou 39% menor em tonalidade de cinza (8 bit) e 45% em tonalidade colorida (24 bit). Com relação à captura em 600dpi, o formato PDF/A ficou 43% menor em tonalidade cinza (8 bit) e 46% menor em tonalidade colorida (24 bit). Na média de todas as configurações, o PDF/A-1b ficou 43% menor que o formato TIFF.

4.4 DIFERENÇAS ENTRE OS NÍVEIS A E B DO PDF/A-1

Nesta etapa produziu-se uma comparação de espaço de armazenamento somente entre arquivos do formato PDF/A, nas suas duas variantes, as de Nível A e Nível B. O interesse neste aspecto foi o de mensurar o impacto no armazenamento dos arquivos, ao se escolher entre os dois níveis.

Para a obra 01, nota-se no quadro 9 que a maior quantidade de elementos obrigatórios no PDF/A-1a traduziu-se em arquivo de maior tamanho em relação ao formato PDF/A-1b.

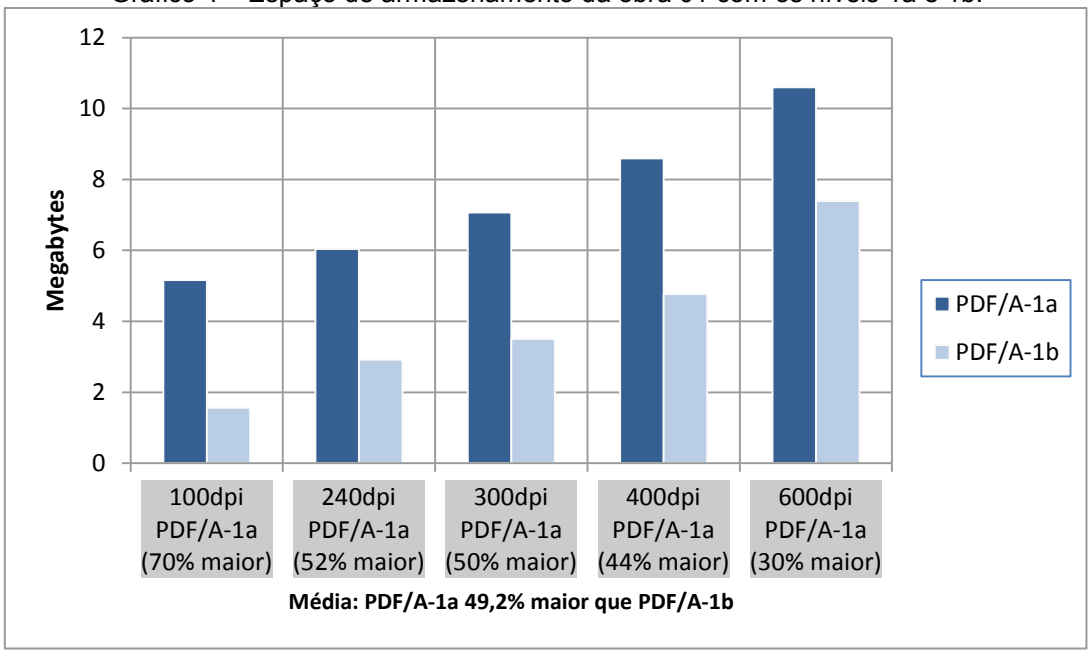
Quadro 9 – Espaço de armazenamento da obra 01 nos níveis 1a e 1b.

Obra 01 (A4 Bitonal) – PDFA-1a / PDFA-1b		
Qualidade da captura	**Tamanho em PDF/A-1a**	**Tamanho em PDF/A-1b**
Densidade: 100dpi Tonalidade: bitonal (1 bit) Compressão: G4	5,16 MB	1,56 MB
Densidade: 240dpi Tonalidade: bitonal (1 bit) Compressão: G4	6,04 MB	2,92 MB
Densidade: 300dpi Tonalidade: bitonal (1 bit) Compressão: G4	7,07 MB	3,50 MB
Densidade: 400dpi Tonalidade: bitonal (1 bit) Compressão: G4	8,59 MB	4,77 MB
Densidade: 600dpi Tonalidade: bitonal (1 bit) Compressão: G4	10,6 MB	7,39 MB

Fonte: Dados da pesquisa (2012)

O contraste de tamanho ocupado pelos arquivos nos dois níveis apresenta-se claramente no gráfico 4, onde se percebe que quanto maior a densidade da captura do documento, também menor fica a diferença entre os níveis 1a e 1b.

Gráfico 4 – Espaço de armazenamento da obra 01 com os níveis 1a e 1b.



Fonte: Dados da pesquisa (2012)

Para a obra 01 capturada em densidade de 100dpi, o nível 1a ficou 70% maior em relação ao formato PDF/A de nível 1b. A diferença caiu progressivamente conforme se aumentou o nível de densidade. Para densidade de 240dpi, o nível 1a ficou 52% maior que o nível 1b. Já com as densidades em 300dpi, 400dpi e 600dpi, as diferenças ficaram em 50%, 44% e 30% maiores do nível 1a para o nível 1b. Em média, o nível 1a ficou 49,2% maior que o nível 1b.

Quando comparado os mesmos dois níveis do PDF/A-1 na conversão da obra 02, a captura em cores gerou diferenças menores. No quadro 10 visualiza-se que independente da densidade utilizada, a diferença é pequena e proporcional ao tamanho total do arquivo.

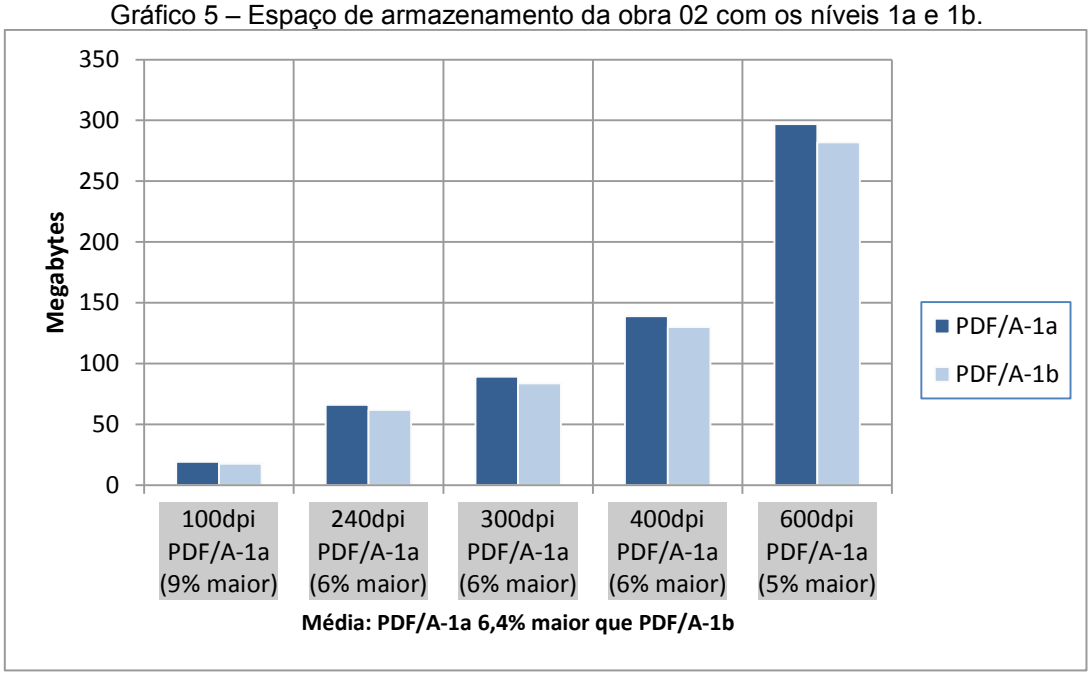
Quadro 10 – Espaço de armazenamento da obra 02 nos níveis 1a e 1b.

Obra 02 (A4 Colorido) – PDFA-1a / PDFA-1b		
Qualidade da captura	**Tamanho em PDF/A-1a**	**Tamanho em PDF/A-1b**
Densidade: 100dpi Tonalidade: colorido (24 bit) Compressão: JPEG	19,3 MB	17,6 MB
Densidade: 240dpi Tonalidade: colorido (24 bit) Compressão: JPEG	66,1 MB	62 MB
Densidade: 300dpi Tonalidade: colorido (24 bit) Compressão: JPEG	89,3 MB	83,8 MB
Densidade: 400dpi Tonalidade: colorido (24 bit) Compressão: JPEG	139 MB	130 MB
Densidade: 600dpi Tonalidade: colorido (24 bit) Compactação: JPEG	297 MB	282 MB

Fonte: Dados da pesquisa (2012)

Esta diferença reduzida fica mais clara de ser percebida ao olhar-se para o gráfico 5, onde houve uma elevação da diferença entre os arquivos conforme se aumentou a densidade da captura do documento.

Gráfico 5 – Espaço de armazenamento da obra 02 com os níveis 1a e 1b.



Fonte: Dados da pesquisa (2012)

Calcularam-se estas diferenças e foi verificado que a obra 02 capturada em densidade de 100dpi, o PDF/A de nível 1a ficou apenas 9% maior em relação ao de nível 1b. Com a densidade em 240dpi, a diferença caiu para 6%, a mesma porcentagem ocorreu para 300dpi e 400dpi. Capturada em 600dpi, a obra 02 ficou 5% maior em nível 1a em relação ao nível 1b. Em média, o nível 1a ficou 6,4% maior que o nível 1b.

A obra 03 não foi incluída no processo comparativo entre os níveis 1a e 1b por não ter sido submetida ao reconhecimento de OCR, devido a não ter conter áreas com marcação, o que conferiria a qualificação de nível 1a.

5 CONCLUSÃO

O proposito desta pesquisa foi o de criar um estudo sobre a utilização do formato PDF/A como meio de produzir representações digitais das obras digitalizadas na Biblioteca Central da Universidade Federal de Santa Catarina. Foram examinados os conceitos sobre o tema da preservação digital, que são a base do desenvolvimento e criação do formato PDF/A. Para atingir os objetivos propostos, também foi comparado o uso dos formatos TIFF e PDF/A, como meio de armazenar as obras em meio eletrônico.

As bibliotecas universitárias são unidades de informações que precisam melhorar a acessibilidade aos seus acervos, e, além disso, garantir a longevidade dos arquivos digitais. Desta forma, o profissional da informação tem papel relevante neste contexto. Pois, é recomendável que esteja ciente sobre cada aspecto que envolve a digitalização, como custos e questões técnicas.

As informações obtidas nesta pesquisa podem contribuir no auxilio ao profissional da informação no desafio de desenvolver estratégias de digitalização de acervo.

De uma maneira geral, os resultados mostraram que há diferenças no tamanho dos arquivos gerados pelos formatos analisados. Estas diferenças sinalizam que o profissional a cargo de planejar a digitalização deve analisar o acervo previamente. Cada decisão sobre os aspectos técnicos irá alterar o tamanho dos arquivos digitais produzidos, influenciando os custos da toda operação.

Entretanto, o PDF/A possui as características necessárias para oferecer acesso aos documentos em longo prazo. Ações como a adoção do formato PDF/A visam amenizar o impacto que as mudanças tecnológicas causam na acessibilidade aos documentos eletrônicos.

Em comparação direta com o formato TIFF, o PDF/A destacou-se por suas características pertinentes a preservação digital. A utilização do formato TIFF como meio de armazenamento eletrônico de documentos digitalizados não preenche as demandas da preservação digital.

Muito se tem ainda a pesquisa sobre preservação digital, uma vez que as tecnologias de documentos eletrônicos estão sempre se alterando, trazendo inovações sim, porém com um preço, que é a obsolescência constante dos formatos de arquivos eletrônicos.

REFERÊNCIAS

ADOBE SYSTEMS INCORPORATED. **Adobe and industry standards**. 2012. Disponível em: <http://www.adobe.com/enterprise/standards/index.html>. Acesso em: 25 ago. 2012.

_____. **Making the case for PDF/A and Adobe Acrobat**. Adobe Systems Incorporated, 2010. Disponível em: <http://www.adobe.com/enterprise/pdfs/pdfaforAcrobat.pdf>. Acesso em: 10 ago. 2012.

_____. **TIFF 6.0 Specification**. 1992. Disponível em: <http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf>. Acesso em: 10 set. 2012.

AMERICAN LIBRARY ASSOCIATION. **Definitions of Digital Preservation**. 2008. Disponível em: <http://www.ala.org/alcts/resources/preserv/defdigpres0408>. Acesso em: 22 ago. 2012.

AQUINO, M.. Metamorfoses da cultura: do impresso ao digital, criando novos formatos e papéis em ambiente de informação. **Ciência da Informação**, Brasília, v. 33, n. 2, p. 7-14, maio/ago. 2004. Disponível em: <http://www.scielo.br/pdf/ci/v33n2/a01v33n2.pdf>. Acesso em: 19 set. 2012.

BESSER, H. Longevidade digital. **Revista Acervo**, Rio de Janeiro, v. 23, n. 2, 2010. Disponível em: <http://revistaacervo.an.gov.br/seer/index.php/info/article/view/11/9>. Acesso em: 19 abr. 2012.

BLATTMANN, Úrsula; FACHIN, Gleisy R. B.; RADOS, Gregório J. V. O bibliotecário na posição do arquiteto da informação. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 11, 2000, Florianópolis. **Anais**... Florianópolis: BUUFSC, 2000. CD-ROM. Disponível em: <http://www.ced.ufsc.br/~ursula/papers/arquinfo.html>. Acesso em: 28 maio 2012.

CAVALCANTE, Lídia Eugênia. Patrimônio digital e informação: política, cultura e diversidade. **Encontros Bibli**, Florianópolis, n.23, p.152-170, 1º. Sem. 2007, Disponível em: <http://www.periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2007v12n23p152/401>. Acesso em: 21 set. 2012.

CONARQ. Conselho Nacional de Arquivos (Brasil). **Recomendações para digitalização de documentos arquivísticos permanentes**. 2010. Disponível em: <http://www.conarq.arquivonacional.gov.br/media/publicacoes/recomenda/recomendaes_para_digitalizao.pdf>. Acesso em: 22 ago. 2012.

____. **e-ARQ Brasil**: modelo de requisitos para sistemas informatizados de gestão arquivística de documentos. Câmara Técnica de Documentos Eletrônicos - CTDE, 2011. Disponível em: <http://www.documentoseletronicos.arquivonacional.gov.br/media/e-arq-brasil-2011-corrigido.pdf>. Acesso em: 02 ago. 2012.

CONWAY, Paul. **Preservação no universo digital**. Coord. Ingrid Beck; trad. José Luiz Pedersoli Júnior e Luiz Antônio Cruz Souza. Rio de Janeiro: Projeto Conservação Preventiva em Bibliotecas e Arquivos: Arquivo Nacional, 2001. 32 p. Disponível em: < http://www.arqsp.org.br/cpba/pdf_cadtec/52.pdf >. Acesso em: 12 ago. 2012.

CUNHA, Murilo Bastos da. Das bibliotecas convencionais às digitais: diferenças e convergências. **Perspect. ciênc. inf.**, Belo Horizonte, v. 13, n. 1, abr. 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362008000100002&lng=en&nrm=iso>. Acesso em: 12 set. 2012.

DRUMMER, Olaf; OETTLER, Alexandra; VON SEGGERN, Dietrich. **PDF/A in a Nutshell**: Long-Term Archiving with PDF. Berlin: Association for Digital Document Standards ADDS – PDF/A Competence Center, 2007. Disponível em: <http://www.pdfa.org/download/pdfa-in-a-nutshell/>. Acesso em: 12 ago 2012.

FERREIRA, Miguel. **Introdução à preservação digital:** Conceitos, estratégias e actuais consensos. Guimarães, Portugal: Escola de Engenharia da Universidade do Minho, 2006. Disponível em: <https://repositorium.sdum.uminho.pt/bitstream/1822/5820/1/livro.pdf>. Acesso em: 12 ago. 2012.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002. 176 p.

GODOY, Arilda Schmidt. Introdução à pesquisa qualitativa e suas possibilidades. In: **Revista de Administração de Empresas - RAE**, v.35, n.2, mar./abr., 1995, p.57-63. Disponível em: <http://rae.fgv.br/sites/rae.fgv.br/files/artigos/10.1590_S0034-75901995000200008.pdf>. Acesso em: 22 abr. 2012.

HAZEN, Dan. Desenvolvimento, gerenciamento e preservação de coleções. In: Planejamento de preservação e gerenciamento de programas. 2.ed. Rio de Janeiro: Projeto Conservação Preventiva em Bibliotecas e Arquivos. Arquivo Nacional, 2001. Disponível em:

<http://www.abracor.com.br/novosite/txt_tecnicos/txt_tecnicos.htm>. Acesso em: 30 out. 2012.

KENNEY, Anne; CHAPMAM, Stephen. **Requisitos de resolução digital para textos**: métodos para o estabelecimento de critérios de qualidade de imagem. Rio de Janeiro: Projeto Conservação Preventiva em Bibliotecas e Arquivos. Arquivo Nacional, 2001. Disponível em: <http://www.arqsp.org.br/cpba/pdf_cadtec/51.pdf>. Acesso em: 12 ago. 2012.

KUNY, Terry. **A digital dark ages?** Challenges in the preservation of electronic information. 63rd IFLA Council and General Conference. 1997.

LACOMBE, C., SILVA, M.. Padrões para Garantir a Preservação e o Acesso aos Documentos Digitais. **Revista Acervo**, Rio de Janeiro, v. 20, nº 1-2, p.113-124, jan/dez. 2007. Disponível em: <http://revistaacervo.an.gov.br/seer/index.php/info/article/view/142>. Acesso em: 12 out. 2012.

LAKATOS, Eva M.; MARCONI, Marina de A. **Metodologia do trabalho científico**. 4. ed. São Paulo: Atlas, 1992. 214 p.

LÉVY, Pierre. **Cibercultura.** (Trad. Carlos Irineu da Costa). São Paulo: Editora 34, 1999.

LIBRARY OF CONGRESS. **Sustainability of Digital Formats**: Planning for Library of Congress Collections. 2011. Disponível em: <http://www.digitalpreservation.gov/formats/index.shtml>. Acesso em: 02 set. 2012.

LOPATIN, Laurie. Library digitization projects, issues and guidelines: a survey of the literature. **Library Hi Tech**, Vol. 24. Issue 2. p. 273-289. Disponível em: <http://www.emeraldinsight.com/journals.htm?articleid=1558880&show=abstract>. Acesso em: 02 set. 2012.

MARCONDES, Carlos Henrique. Linguagem e documento: fundamentos evolutivos e culturais da Ciência da Informação. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 15, n. 2, p. 2-21, 2010. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/viewFile/1019/729>. Acesso em: 10 ago. 2012.

MARCONDES, Carlos Henrique; KURAMOTO, Hélio; TOUTAIN, Lídia Brandão; SAYÃO, Luís Fernando (Org.). **Bibliotecas digitais**: saberes e práticas. 2.ed. Salvador: UFBA; Brasília: IBICT, 2006.

MÁRDERO ARELLANO, M. A.. Preservação de Documentos Digitais. **Ciência da Informação**, Brasília, DF, Brasil, 33, dez. 2004. Disponível em: <http://www.scielo.br/pdf/ci/v33n2/a02v33n2.pdf>. Acesso em: 08 ago. 2012.

MÁRDERO ARELLANO, Miguel Ángel; ANDRADE, Ricardo Sodré. Preservação digital e os profissionais da informação. **DataGramaZero**, v. 7, n. 5, out. 2006. Disponível em: <http://www.dgz.org.br/out06/Art_05.htm>. Acesso em: 24 ago. 2012.

MELLO, José Barboza. **Síntese histórica do livro.** Rio de Janeiro: Editora Ibrasa. 1979, 334 p.

MILES, Doug. **PDF/A makes slow but steady progress**. AIIM, 2010. Disponível em: <http://aiim.typepad.com/ecmbynumbers/2010/09/pdfa-makes-slow-but-steady-progress.html>. Acesso em: 28 out 2012.

NATIONAL LIBRARY OF AUSTRALIA (NLA). **Guidelines for the preservation of digital heritage**. Paris: UNESCO, 2003. 177p. Disponível em: <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>. Acesso em: 25 ago. 2012.

ORTEGA, Cristina Dotta; LARA, Marilda Lopes Ginez de. A noção de documento: de Otlet aos dias de hoje. **DataGramaZero**, v. 11, n. 2, abr. 2010. Disponível em: <http://www.dgz.org.br/abr10/Art_03.htm>. Acesso em: 24 abr. 2012.

PAES, Marilena Leite. **Arquivo**: teoria e prática. Rio de Janeiro: Editoria FGV. 2007, 225 p.

PEARSON, David. Digitization: Do We Have a Strategy? January 2002, **Ariadne**, Issue 30. Disponível em: <http://www.ariadne.ac.uk/issue30/digilib/>. Acesso em: 28 out. 2012.

PROENÇA, Ana Luísa Morão Raposo Martins; LOPES, Sandra Guerra. **Digital Preservation**. Covilhã, Portugal: Universidade da Beira Interior, [200?]. Disponível em: <http://www.di.ubi.pt/~api/digital_preservation.pdf>. Acesso em: 30 maio 2012.

PUGLIA, Steven. Technical primer. In: SITTS, Maxine K. (Ed.). **Handbook for digital projects**: a management tool for preservation and access. Andover (MA): Northeast Document Conservation Center, 2000. Disponível em: <http://www.nedcc.org/resources/digitalhandbook/dighome.htm>. Acesso em: 25 set. 2012.

RIKOWSKI, Ruth (Ed.). **Digitisation perspectives**. Rotterdam: Sense Publishers. 2011, 303 p.

RONDINELLI, Rosely Cury. **Gerenciamento arquivístico de documentos eletrônicos**: uma abordagem teórica da diplomática arquivística contemporânea. Rio de Janeiro: FGV, 2005.

ROSENHTAL, David S.H. Format obsolescence: assessing the threat and the defenses. **Library Hi Tech**, Vol. 28, nº 2, p. 195-210, 2010. Disponível em: <http://www.emeraldinsight.com/journals.htm?articleid=1864748>. Acesso em: 02 set. 2012.

SANTOS, Maria José Veloso da Costa. A representação da informação em arquivos: viabilidade de uso dos padrões utilizados na biblioteconomia. **Revista Acervo**, Rio de Janeiro, v.20, nº 1-2, p.57-66, jan/dez, 2007. Disponível em: <http://revistaacervo.an.gov.br/seer/index.php/info/article/view/138>. Acesso em: 24 maio. 2012.

SILVA, Edna Lucia da; MENEZES, Estera Muszkat. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. rev. atual. Florianópolis: Ed. da UFSC, 2005. Disponível em: <http://projetos.inf.ufsc.br/arquivos/Metodologia_de_pesquisa_e_elaboracao_de_teses_e_dissertacoes_4ed.pdf>. Acesso em: 02 ago. 2012.

SIQUEIRA, Jéssica Câmara. Biblioteconomia, documentação e ciência da informação: história, sociedade, tecnologia e pós-modernidade. **Perspect. ciênc. inf.** [online]. 2010, vol.15, n.3, pp. 52-66. ISSN 1413-9936. Disponível em: <http://www.scielo.br/pdf/pci/v15n3/04.pdf>. Acesso em: 08 ago. 2012.

SPELLMAN, Rosemary; HOLLEY, Robert P. An Overview of the Google Books Project and Other Digitization Initiatives: Implications for Libraries. **Journal of Library and Information Science**. Vol. 37, Issue 1, 2011. Disponível em: <http://jlis.glis.ntnu.edu.tw/ojs/index.php/jlis/article/view/548>. Acesso em: 20 out. 2012.

TAMMARO, Anna Maria; SALARELLI, Alberto. **A biblioteca digital**. Brasília: Briquet de Lemos, 2008. 377p.

UNESCO. **Guidelines for the preservation of digital heritage**. 2003. 170 p. Disponível em: <http://unesdoc.unesco.org/ulis/cgi-bin/ulis.pl?catno=130071>. Acesso em: 19 ago. 2012.

WATERS, Donald J. **Do microfilme a imagem digital**: como executar um projeto para estudo dos meios, custos e benefícios de conversão para imagens

digitais de grandes quantidades de documentos preservados em microfilme. 2.ed. Rio de Janeiro: Projeto Conservação Preventiva em Bibliotecas e Arquivos. Arquivo Nacional, 2001. Disponível em: <http://www.abracor.com.br/novosite/txt_tecnicos/txt_tecnicos.htm>. Acesso em: 28 out. 2012.

APENDICE A

- Manual de utilização do ABBYY FineReader 11 –

Edição e conversão de páginas para o formato PDF/A

Elaborado por Alexandre Pedron Martins para utilização no trabalho de digitalização das teses e dissertações da Biblioteca Central da Universidade Federal de Santa Catarina, como produto da disciplina de estágio obrigatório do Curso de Biblioteconomia da UFSC.

Este manual lida com os recursos do software **ABBYY FineReader** para a criação de documentos PDF/A a partir de páginas escaneadas. Trata-se de um aplicativo de OCR (reconhecimento óptico de caracteres) que converte documentos impressos, documentos em PDF e imagens de documentos em arquivos para edição em computador.

Atualmente há diversos softwares com a mesma função do FineReader, entre eles Expervision TypeReader, OmniPage e AnyDoc. Porém o FineReader é o primeiro software de OCR que realmente "vê" o documento como um todo. Este software contem o sistema ADRTTM *(Tecnologia Adaptável de Reconhecimento de Documento)*, que analisa o documento como entidade única e compreende todos os elementos de sua estrutura, tais como: texto principal, colunas, tabelas, cabeçalhos, rodapés, notas de rodapé e numeração de páginas.

Software: ABBYY FineReader 11 Corporate Edition (Windows)

Versão: 11.0.102.519

Conteúdo

1 – Configurações

Antes de iniciar a utilizar o ***ABBYY FineReader***, verifique as configurações do programa conforme orientações abaixo. Para verificar o painel de configuração, acesse no menu superior a opção *Ferramentas* e em seguida *Opções*.

Na primeira aba de opções indicada como *Documentos*, verifique se a configuração esta como a demonstrada na imagem 1.

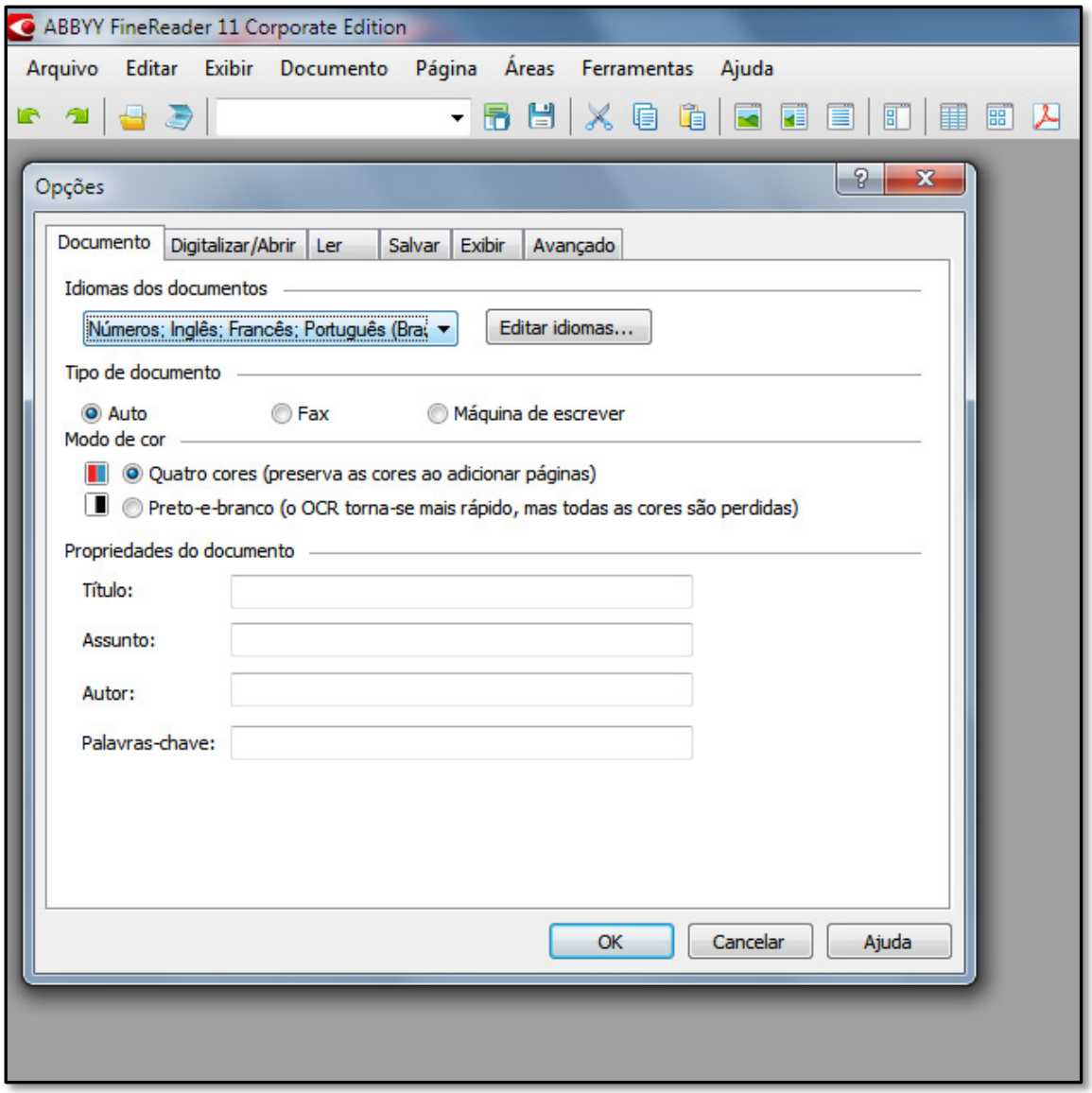


Imagem 1 – Opções de documento

No campo referente aos idiomas dos documentos, deixe selecionadas as opções: Números, Inglês, Francês, Português (Brasil), Português e Espanhol.

A próxima aba de configurações trata do processo de digitalizar e ler os documentos carregados. Conforme imagem 2, certifique-se que esteja marcado somente a opção para não ler automaticamente as imagens.

As funções de pré-processamento não serão utilizadas pois normalmente as páginas precisam ser editadas para somente depois serem processadas.

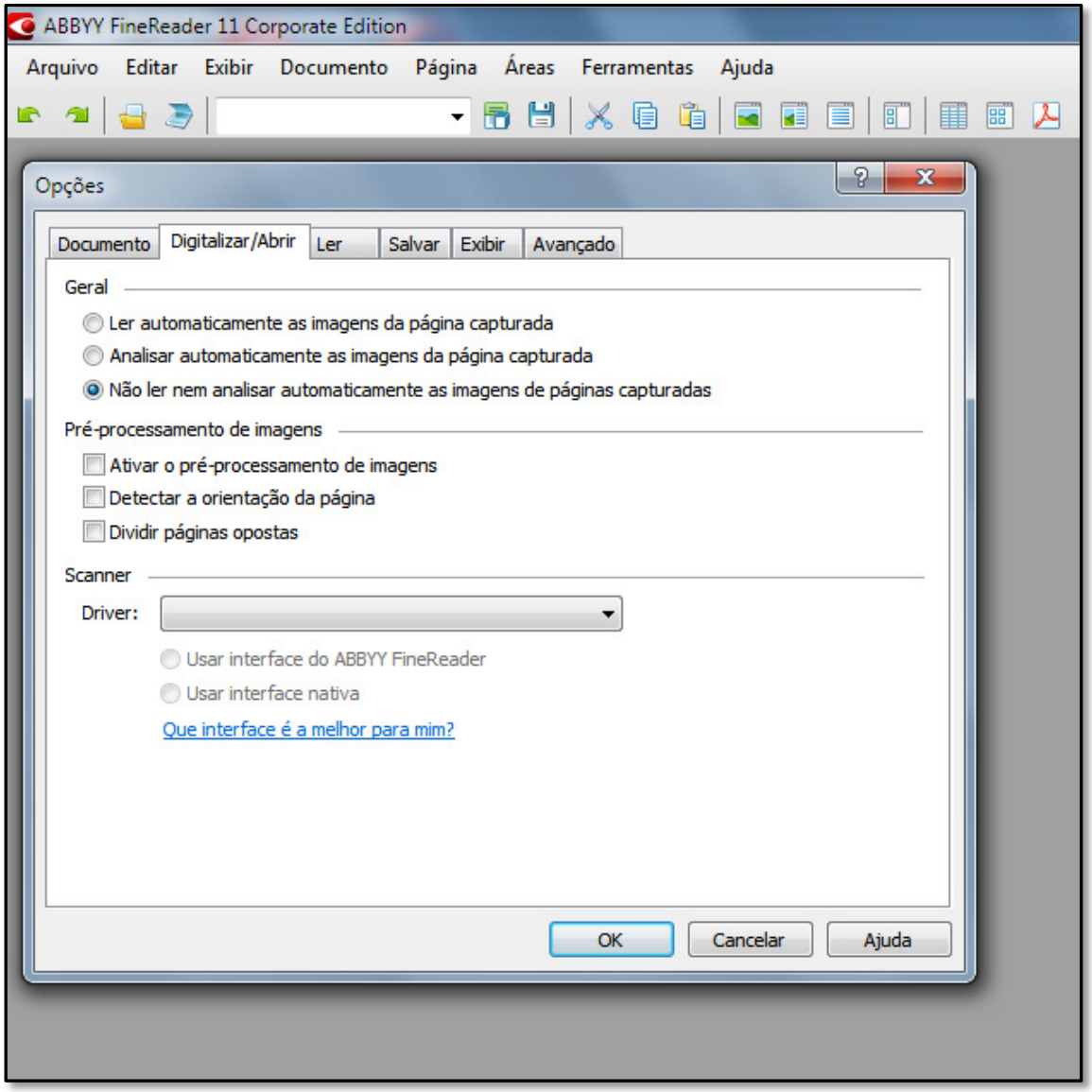


Imagem 2 – Opções para Digitalizar/Abrir documentos

A aba das opções de leitura das imagens, identificada como *Ler*, deve ser mantido conforme a imagem 3. Deixe marcadas as opções *Leitura completa* e *Utilize apenas padrões internos*.

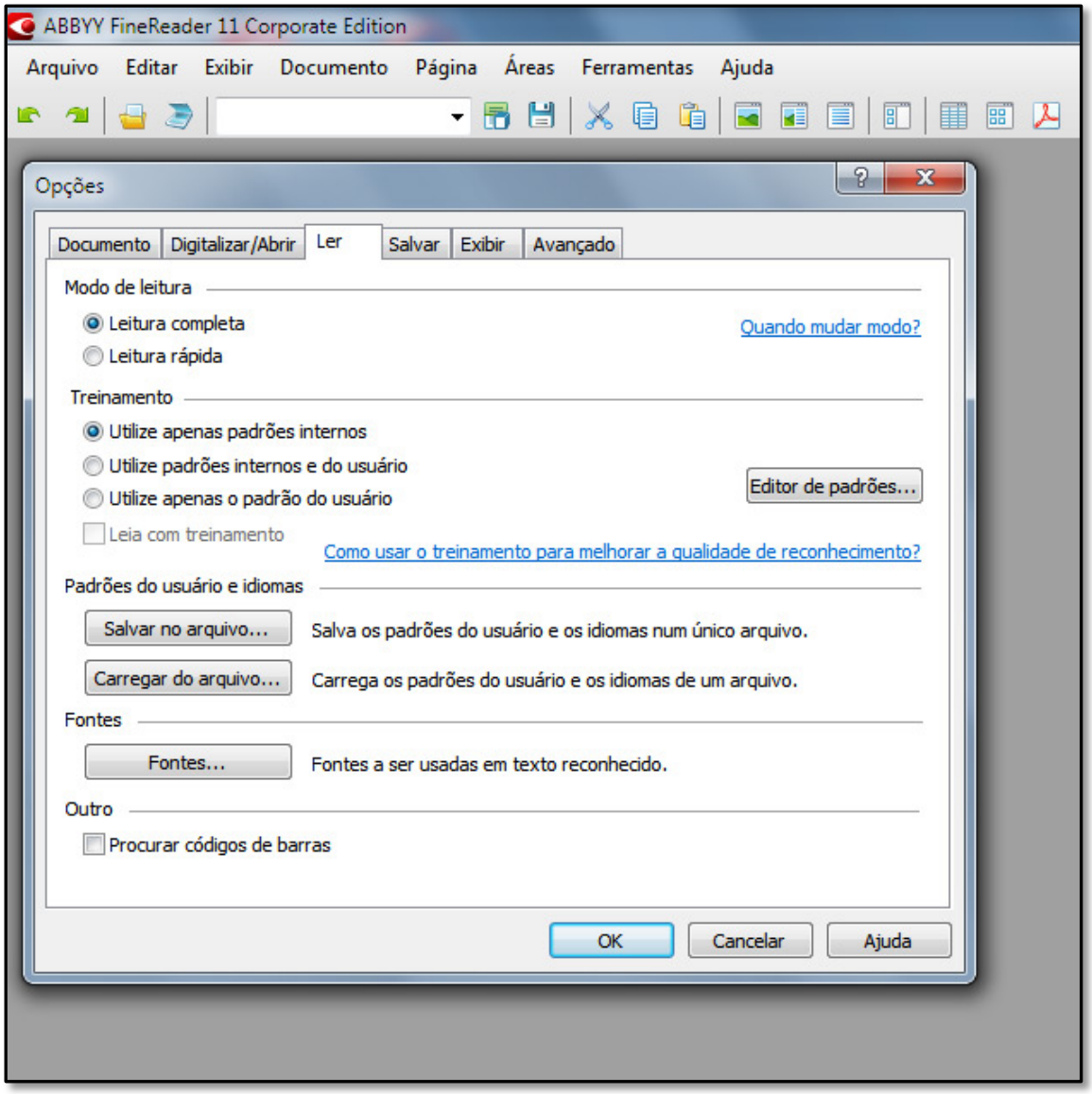


Imagem 3 – Opções de leitura de documentos

Após selecionar a aba *Salvar*, clique em seguida na opção *PDF/A* conforme a imagem 4. No campo *Tamanho de papel padrão*, deverá ser utilizado a opção *A4* ou *Automático*. Esta escolha será de acordo com o tipo material, utilizando-se a opção *A4* para quando todas as páginas forem do formato A4 e estiverem na disposição retrato (vertical) ou a opção *Automático* para quando tiver páginas em formato A4 na disposição paisagem (horizontal) e/ou contiver mapas e gráficos em tamanhos maiores.

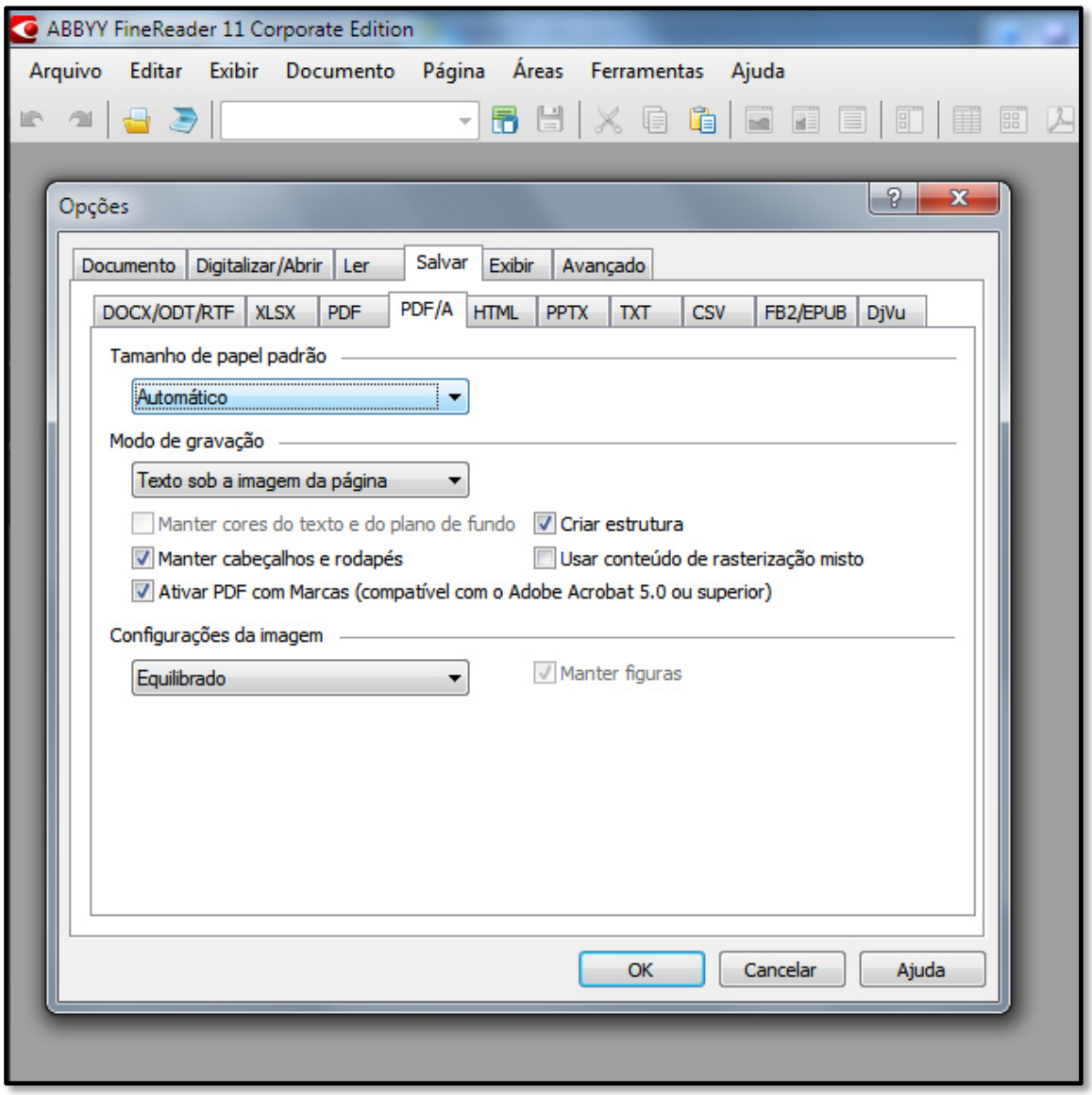


Imagem 4 – Opções de PDF/A

Ainda com relação à opção de PDF/A, há a escolha do modo de gravação. Na primeira caixa de seleção, serão utilizadas duas das opções disponíveis. A primeira opção que utilizaremos é *Texto sob a imagem da página*, que será usada para deixar os caracteres reconhecidos pelo OCR por baixo da imagem das páginas.

A outra opção disponível é *Imagem da página somente*, onde não será realizado o reconhecimento dos caracteres através do OCR. Esta opção deverá ser utilizada somente quando o material compuser de muitas páginas com imagens (desenhos, fotografias, etc), que demandaria muito tempo de processamento.

As demais opções que devem ficar ativas são:

- **Manter cabeçalhos**: mantem os cabeçalhos e rodapés do arquivo final.
- **Ativar PDF com Marcas**: mantem uma ordenação correta do texto e conteúdos (figuras, tabelas, ...).

- **Criar estrutura**: monta um índice no arquivo através do uso dos cabeçalhos detectados.
- **Usar conteúdo de rasterização misto:** realiza uma divisão do documento em camadas, realizando uma compressão diferenciada para cada uma. Reduz o tamanho do documento final, porém é mais efetiva em documentos coloridos.

O item **Configuração de imagem** deve ser mantido no modo *Equilibrado*.

Na aba *Exibir*, deixe as opções padrões conforme imagem 5.

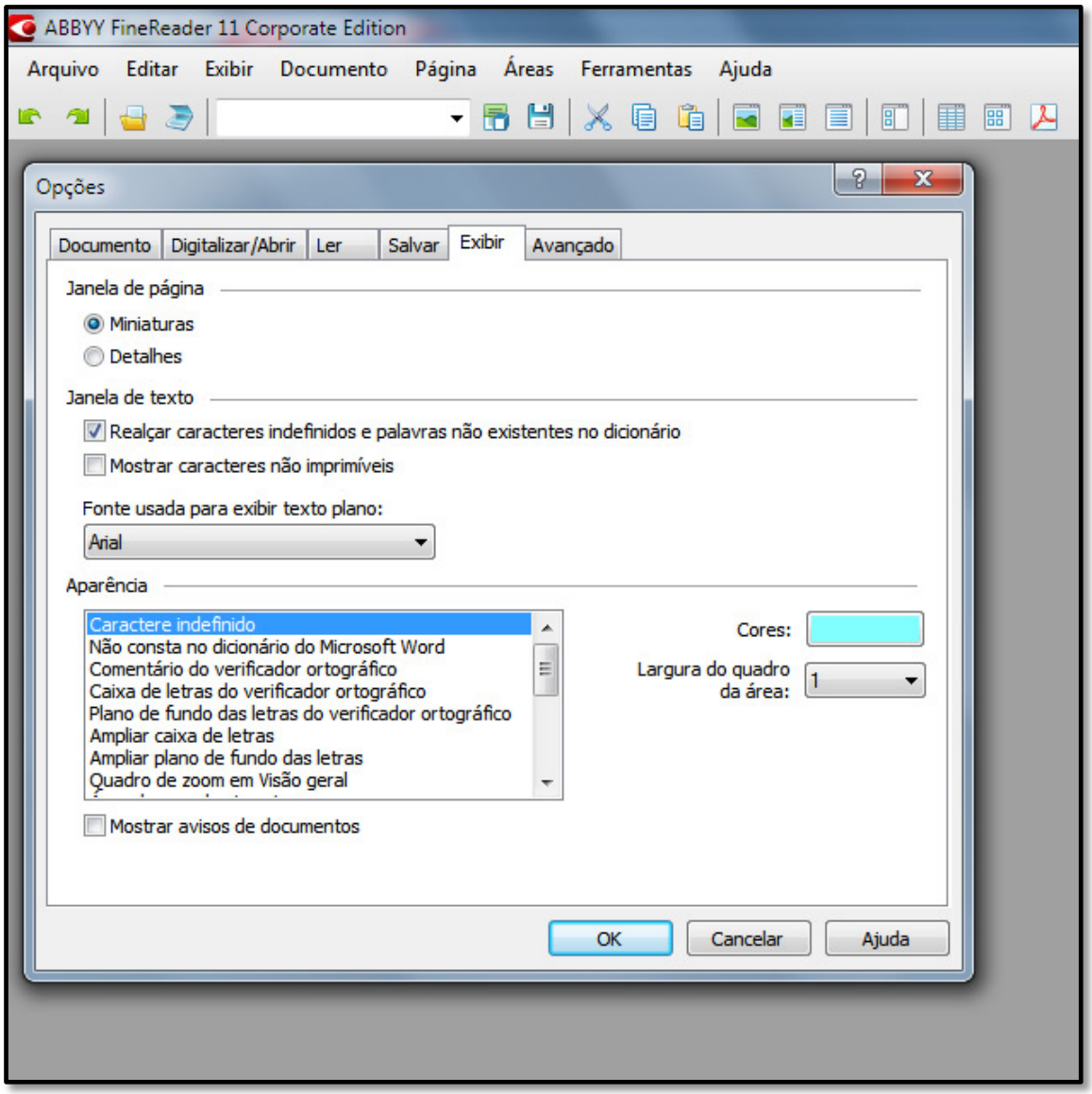


Imagem 5 – Opções de exibição

Para a última aba, intitulada *Avançado*, deixe marcado as opções de acordo com a imagem 6.

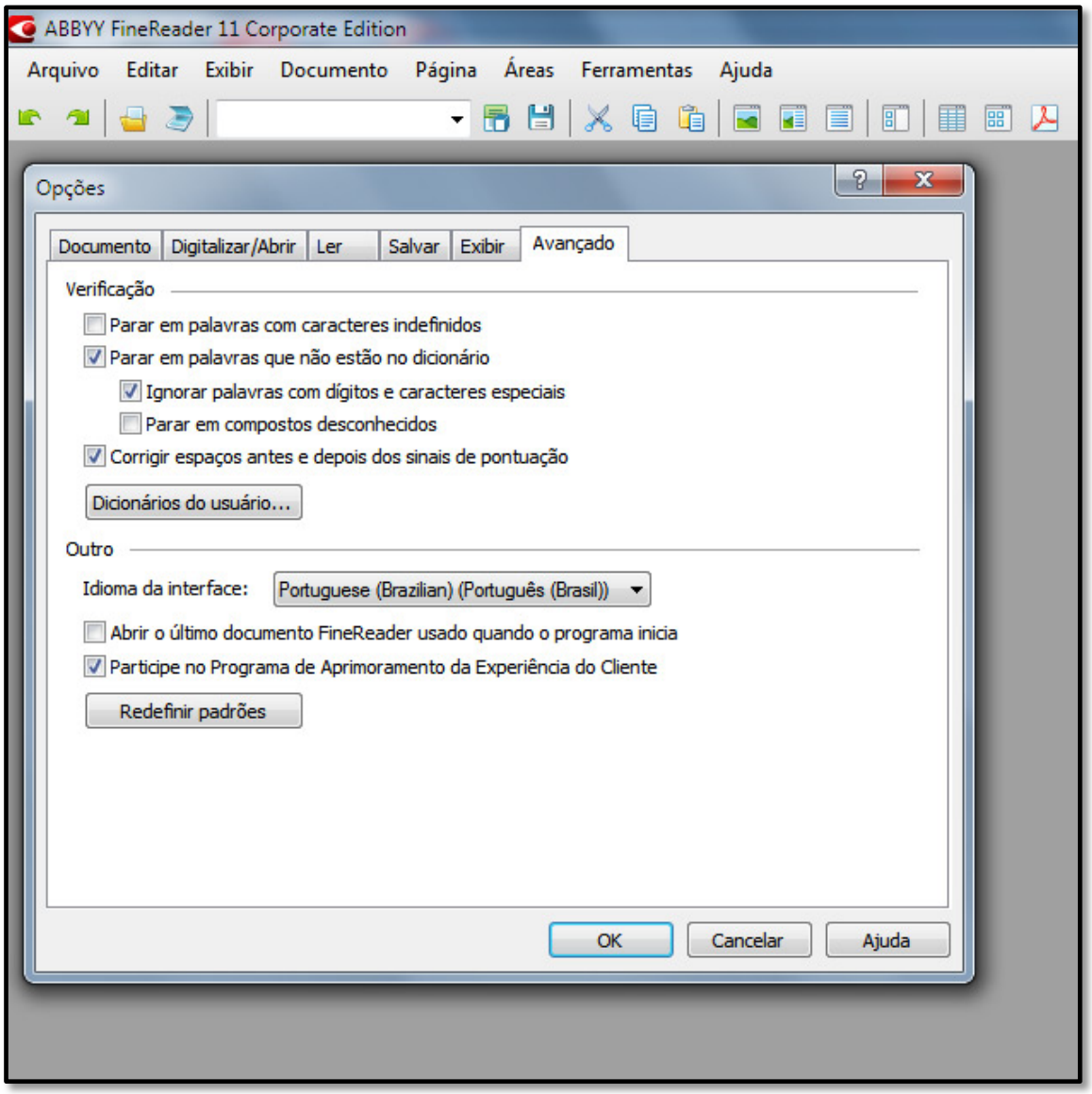


Imagem 6 – Opções avançadas

2 – Carregando as imagens

A próxima etapa do processo consiste em carregar as imagens para serem editadas e convertidas. Na interface do FineReader, utilize a opção de abrir arquivo através do menu *Arquivo* e em seguida *Abrir arquivo PDF/imagem* ou então com o comando **CRTL**+**O**.

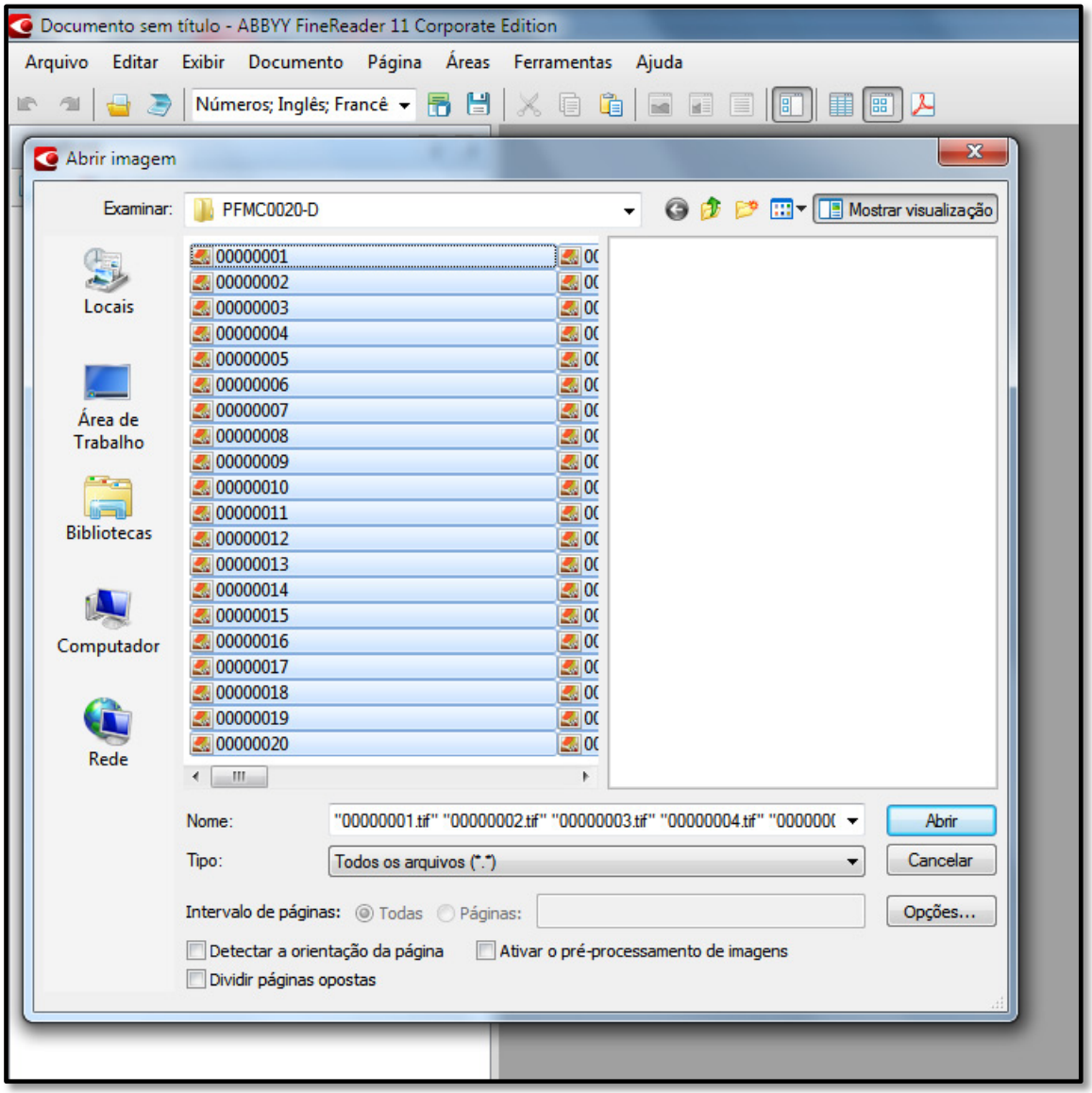


Imagem 7 – Abrindo imagens

Conforme mostra a imagem 7, navegue até o local onde se encontram as imagens escaneadas. Selecione todas elas com o comando **CRTL**+**A** e deixe desmarcadas as opções visíveis na parte inferior da janela.

3 – Editando as imagens

Após as imagens terem sido carregadas no FineReader, é preciso editar através do botão *Editar* ou com o comando **CRTL+SHIFT+C**, onde então ficará disponível na interface o menu com os recursos.

3.1 Removendo sombras e manchas

Quando as imagens contiverem sombras ou manchas decorrentes do processo de escaneamento, estas precisam ser removidas com os ajustes de brilho e contraste. De acordo com o exemplo mostrado na imagem 8, utilize a opção *Brilho & Contraste* no menu lateral.

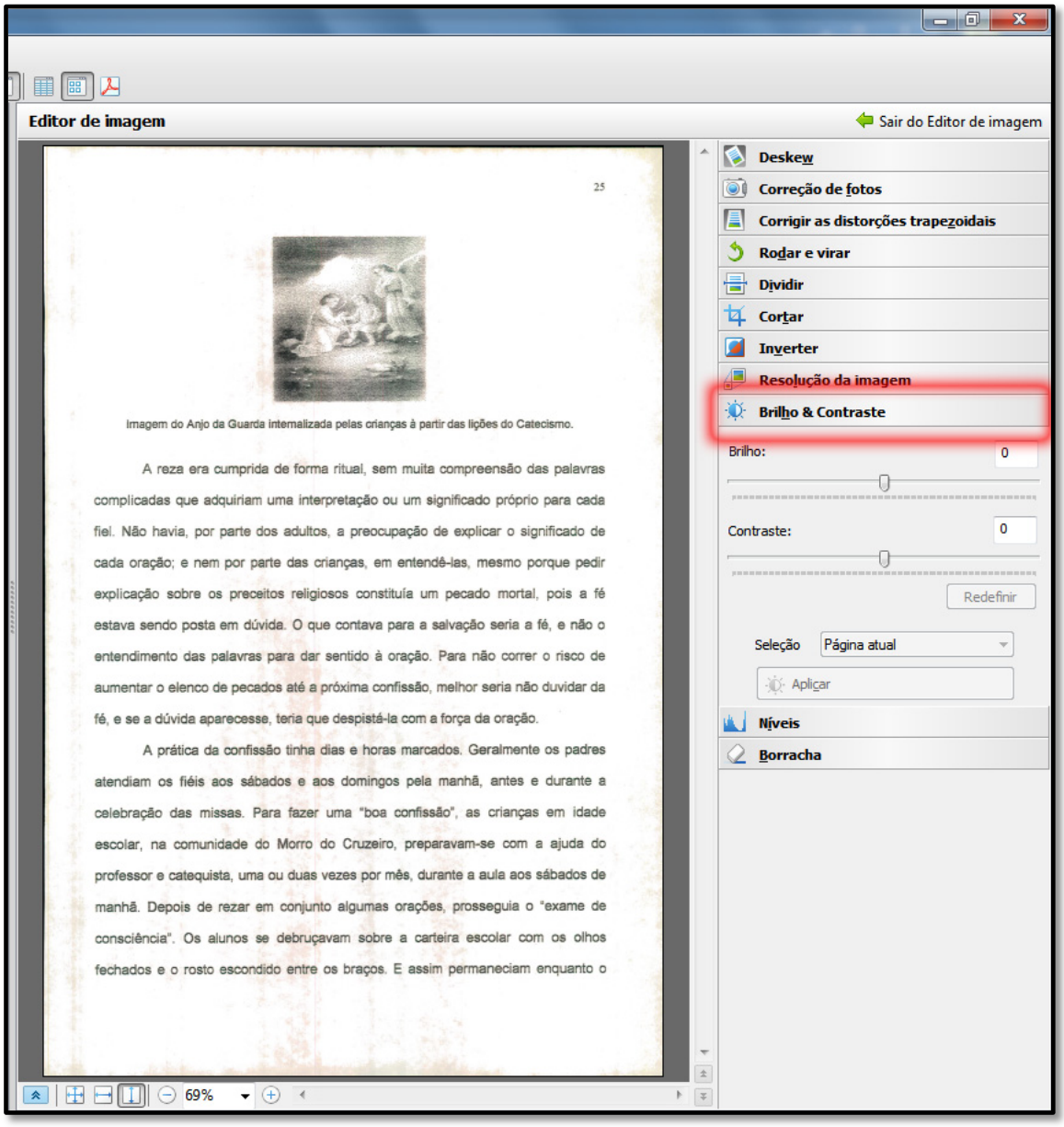


Imagem 8 – Ajuste de brilho e contraste

Na imagem 9 com o exemplo mostrado, foi utilizado um ajuste de brilho em 8 e de contraste em 20. Após inserir os valores, deve-se clicar no botão Aplicar. Estes valores podem variar conforme cada caso, sendo preciso tomar cuidado para não clarear demasiadamente as imagens e perder informação.

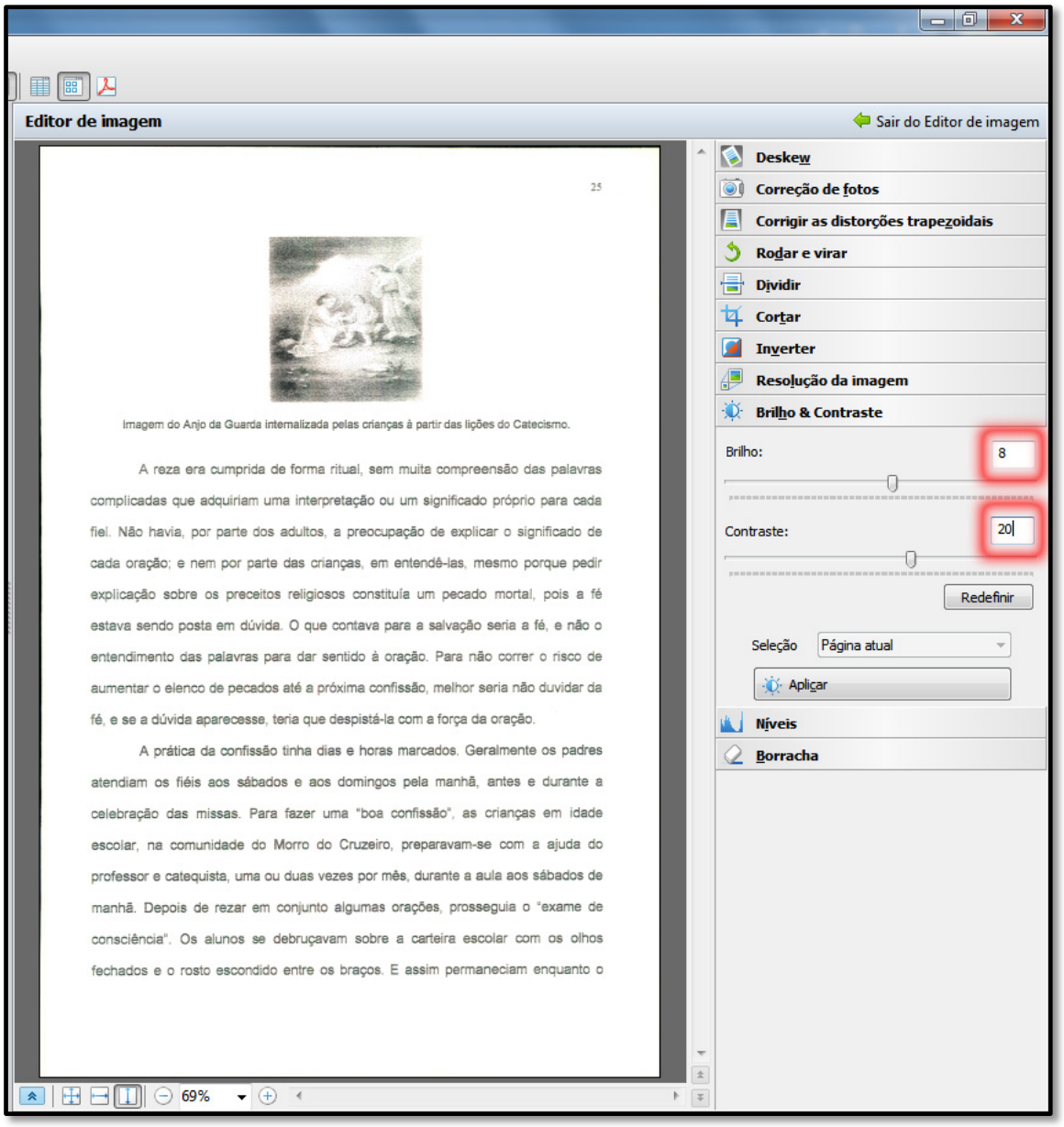
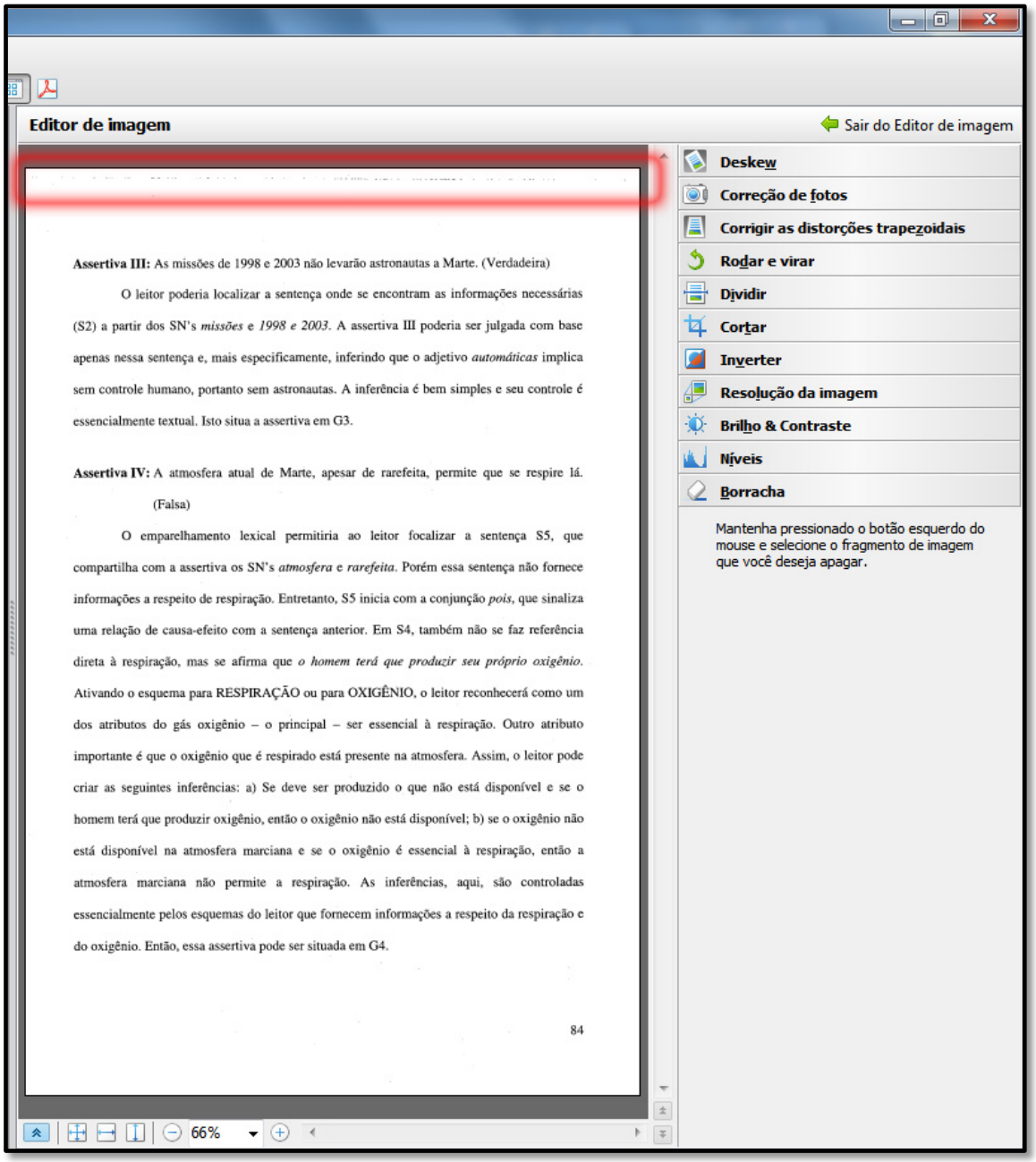


Imagem 9 – Ajustando brilho e contraste

3.2 Cortando as páginas

Algumas páginas escaneadas poderão conter também marcas e manchas em suas bordas, assim como estarem com margens fora de padrão. Para estes casos é utilizado a ferramenta *Cortar*, como mostrado na imagem 10.



Assertiva III: As missões de 1998 e 2003 não levarão astronautas a Marte. (Verdadeira)

O leitor poderia localizar a sentença onde se encontram as informações necessárias (S2) a partir dos SN's *missões* e *1998 e 2003*. A assertiva III poderia ser julgada com base apenas nessa sentença e, mais especificamente, inferindo que o adjetivo *automáticas* implica sem controle humano, portanto sem astronautas. A inferência é bem simples e seu controle é essencialmente textual. Isto situa a assertiva em G3.

Assertiva IV: A atmosfera atual de Marte, apesar de rarefeita, permite que se respire lá. (Falsa)

O emparelhamento lexical permitiria ao leitor focalizar a sentença S5, que compartilha com a assertiva os SN's *atmosfera* e *rarefeita*. Porém essa sentença não fornece informações a respeito de respiração. Entretanto, S5 inicia com a conjunção *pois*, que sinaliza uma relação de causa-efeito com a sentença anterior. Em S4, também não se faz referência direta à respiração, mas se afirma que *o homem terá que produzir seu próprio oxigênio*. Ativando o esquema para RESPIRAÇÃO ou para OXIGÊNIO, o leitor reconhecerá como um dos atributos do gás oxigênio – o principal – ser essencial à respiração. Outro atributo importante é que o oxigênio que é respirado está presente na atmosfera. Assim, o leitor pode criar as seguintes inferências: a) Se deve ser produzido o que não está disponível e se o homem terá que produzir oxigênio, então o oxigênio não está disponível; b) se o oxigênio não está disponível na atmosfera marciana e se o oxigênio é essencial à respiração, então a atmosfera marciana não permite a respiração. As inferências, aqui, são controladas essencialmente pelos esquemas do leitor que fornecem informações a respeito da respiração e do oxigênio. Então, essa assertiva pode ser situada em G4.

84

Imagem 10 – Selecionando área de corte

Conforme imagem 11, após selecionar a ferramenta de corte, será mostrada sobre a imagem da página a grade para demarcar a área de corte.

Antes de ajustar a área de corte, é necessário verificar no painel da esquerda se não há alguma página que tenha conteúdo muito próximo às bordas, caso contrário será perdido informação importante no documento final. Páginas que esteja no formato de paisagem (horizontal) devem ser removidas da seleção, para não serem cortadas de forma indevida.

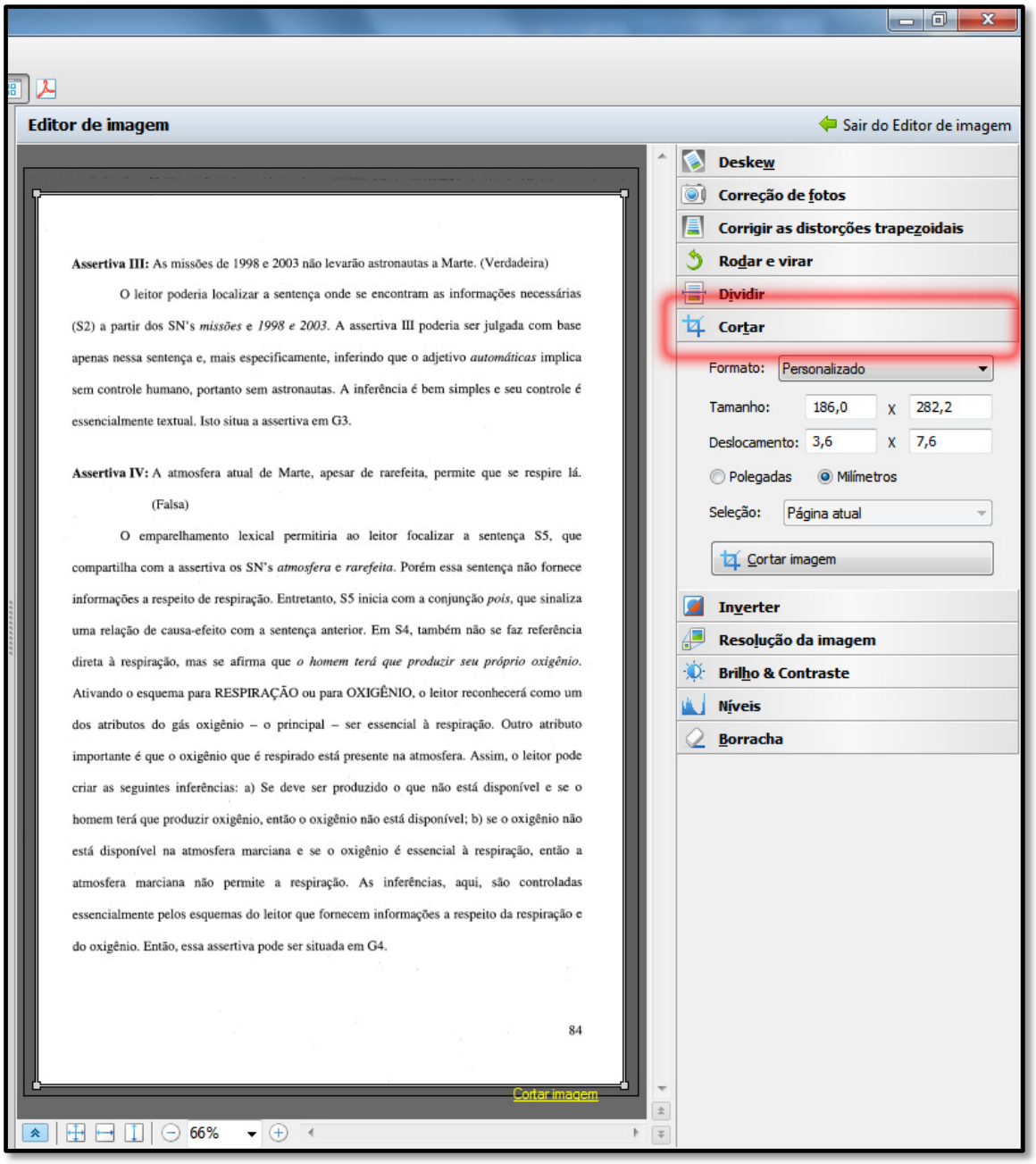
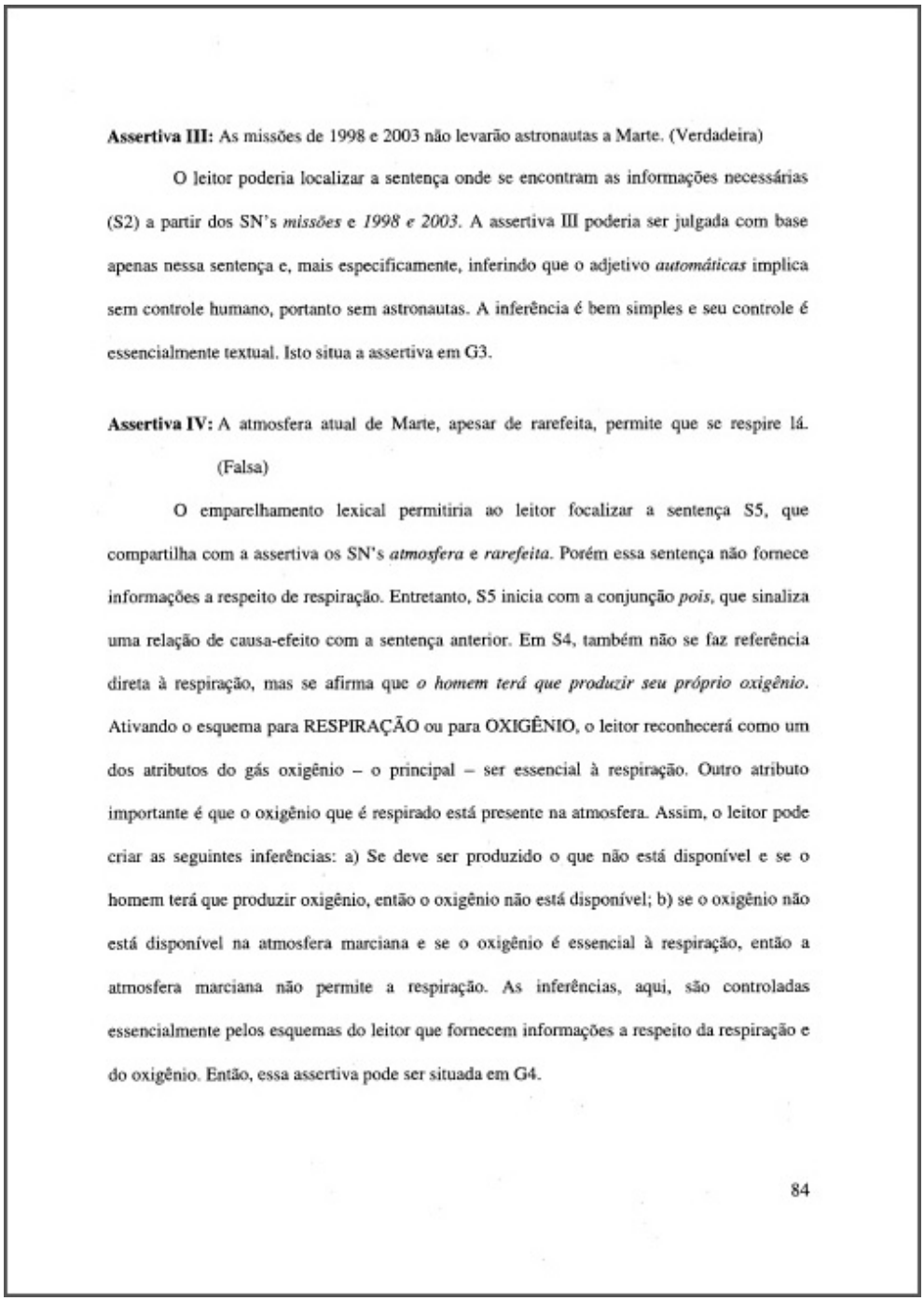


Imagem 11 – Cortando documento

Percebe-se na imagem 11 que após o corte a margem do texto ficará pequena, porém isto pode ser ignorado no caso de criação de PDF no formato A4, pois o FineReader irá centralizar a imagem numa nova página em branco. Isto pode ser confirmado na imagem 12, onde a imagem cortada fica centralizada na página final.



Assertiva III: As missões de 1998 e 2003 não levarão astronautas a Marte. (Verdadeira)

O leitor poderia localizar a sentença onde se encontram as informações necessárias (S2) a partir dos SN's *missões* e *1998 e 2003*. A assertiva III poderia ser julgada com base apenas nessa sentença e, mais especificamente, inferindo que o adjetivo *automáticas* implica sem controle humano, portanto sem astronautas. A inferência é bem simples e seu controle é essencialmente textual. Isto situa a assertiva em G3.

Assertiva IV: A atmosfera atual de Marte, apesar de rarefeita, permite que se respire lá. (Falsa)

O emparelhamento lexical permitiria ao leitor focalizar a sentença S5, que compartilha com a assertiva os SN's *atmosfera* e *rarefeita*. Porém essa sentença não fornece informações a respeito de respiração. Entretanto, S5 inicia com a conjunção *pois*, que sinaliza uma relação de causa-efeito com a sentença anterior. Em S4, também não se faz referência direta à respiração, mas se afirma que *o homem terá que produzir seu próprio oxigênio*. Ativando o esquema para RESPIRAÇÃO ou para OXIGÊNIO, o leitor reconhecerá como um dos atributos do gás oxigênio – o principal – ser essencial à respiração. Outro atributo importante é que o oxigênio que é respirado está presente na atmosfera. Assim, o leitor pode criar as seguintes inferências: a) Se deve ser produzido o que não está disponível e se o homem terá que produzir oxigênio, então o oxigênio não está disponível; b) se o oxigênio não está disponível na atmosfera marciana e se o oxigênio é essencial à respiração, então a atmosfera marciana não permite a respiração. As inferências, aqui, são controladas essencialmente pelos esquemas do leitor que fornecem informações a respeito da respiração e do oxigênio. Então, essa assertiva pode ser situada em G4.

84

Imagem 12 – Alinhando no formato A4

3.3 Enquadramento das páginas

Algumas páginas também podem vir escaneadas inclinadas ou quando o próprio material original tiver defeito, conforme se observa na imagem 13. Nestes casos devemos realizar o enquadramento destas páginas com a ferramenta *Deskew*.

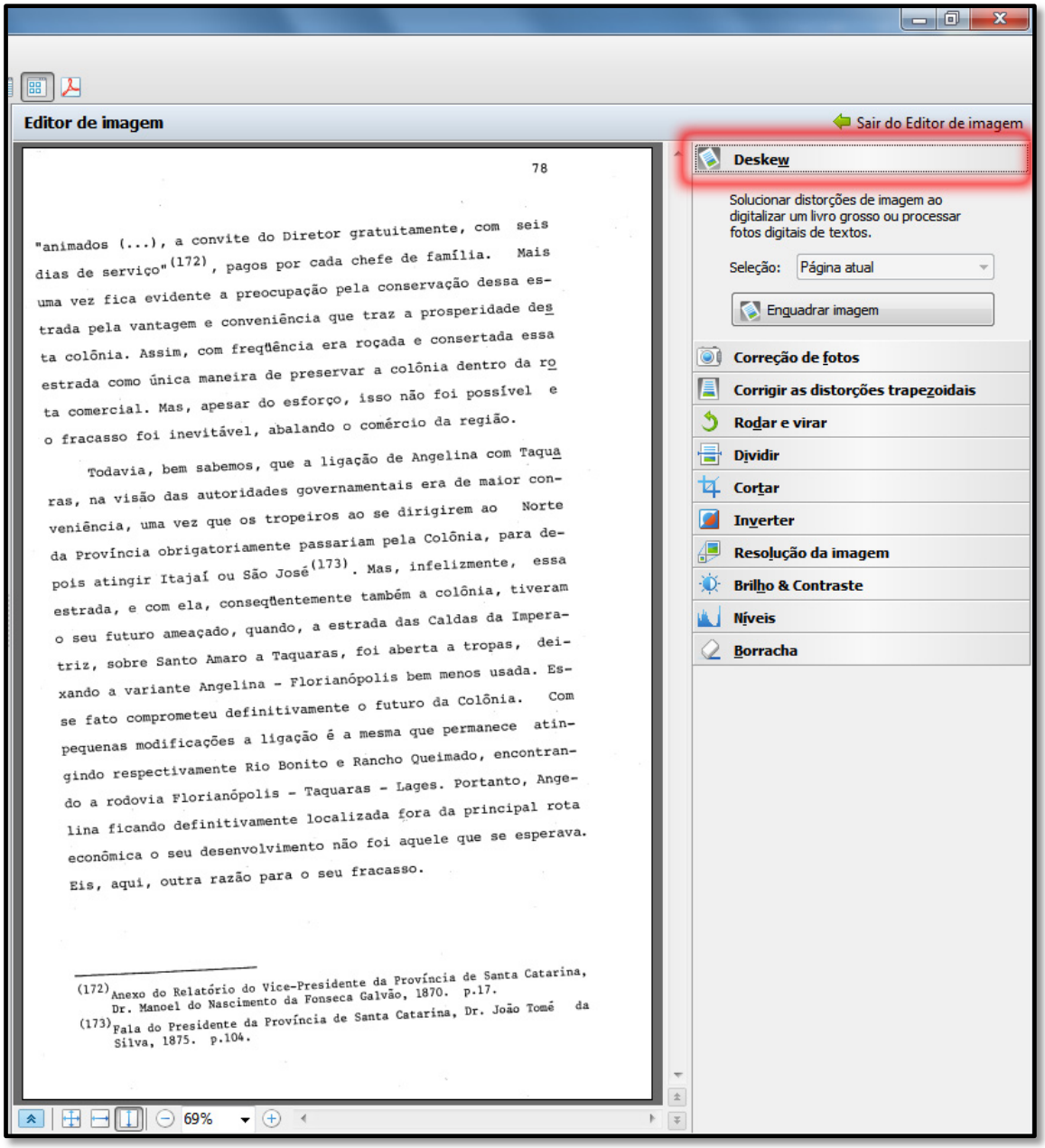


Imagem 13 - Deskew

No painel de visualização das páginas, selecione quais deseja corrigir ou então selecione todas as imagens com o comando **CRTL+A**. Em seguida clique na opção *Enquadrar imagem*, conforme a imagem 14.

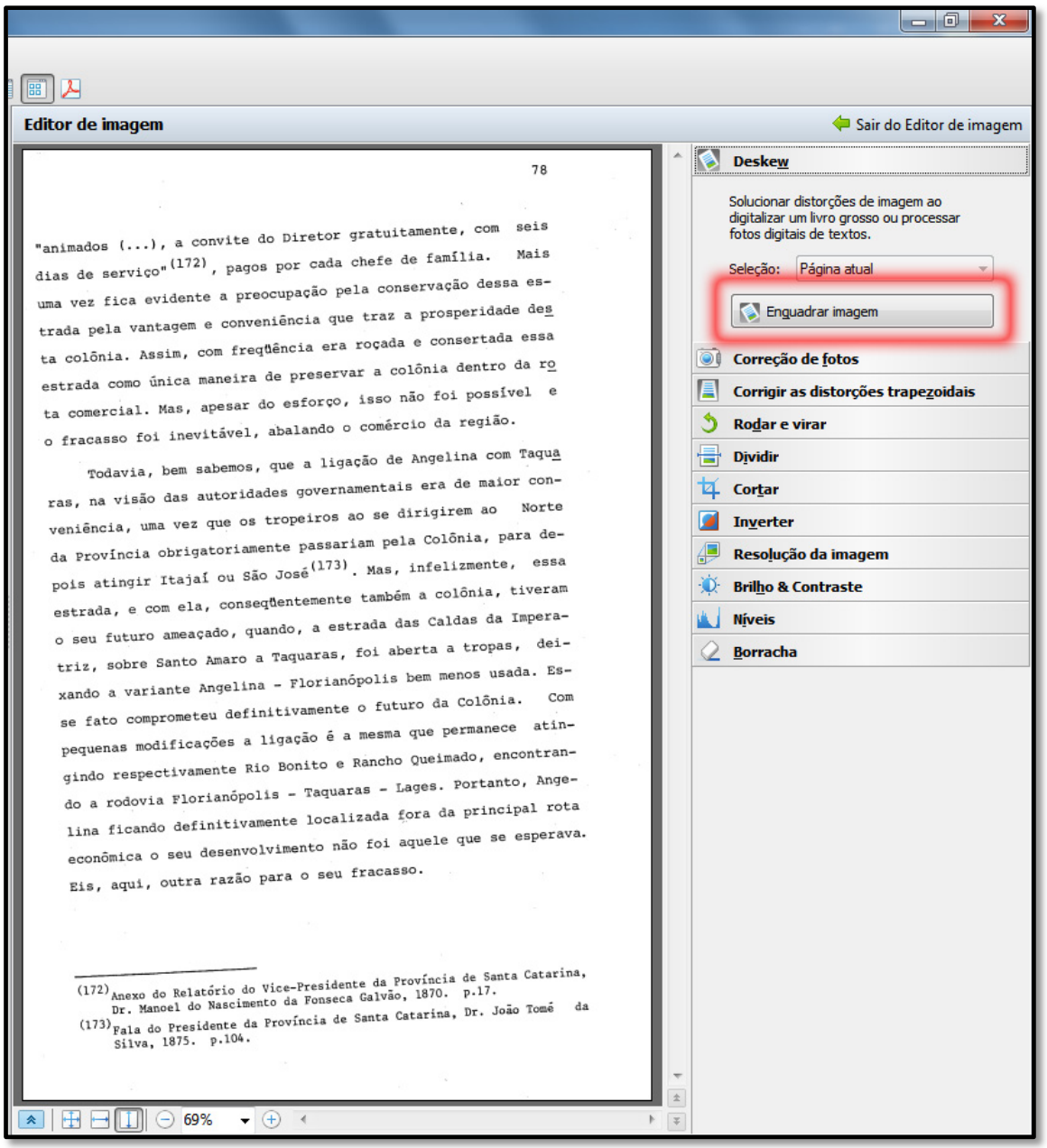
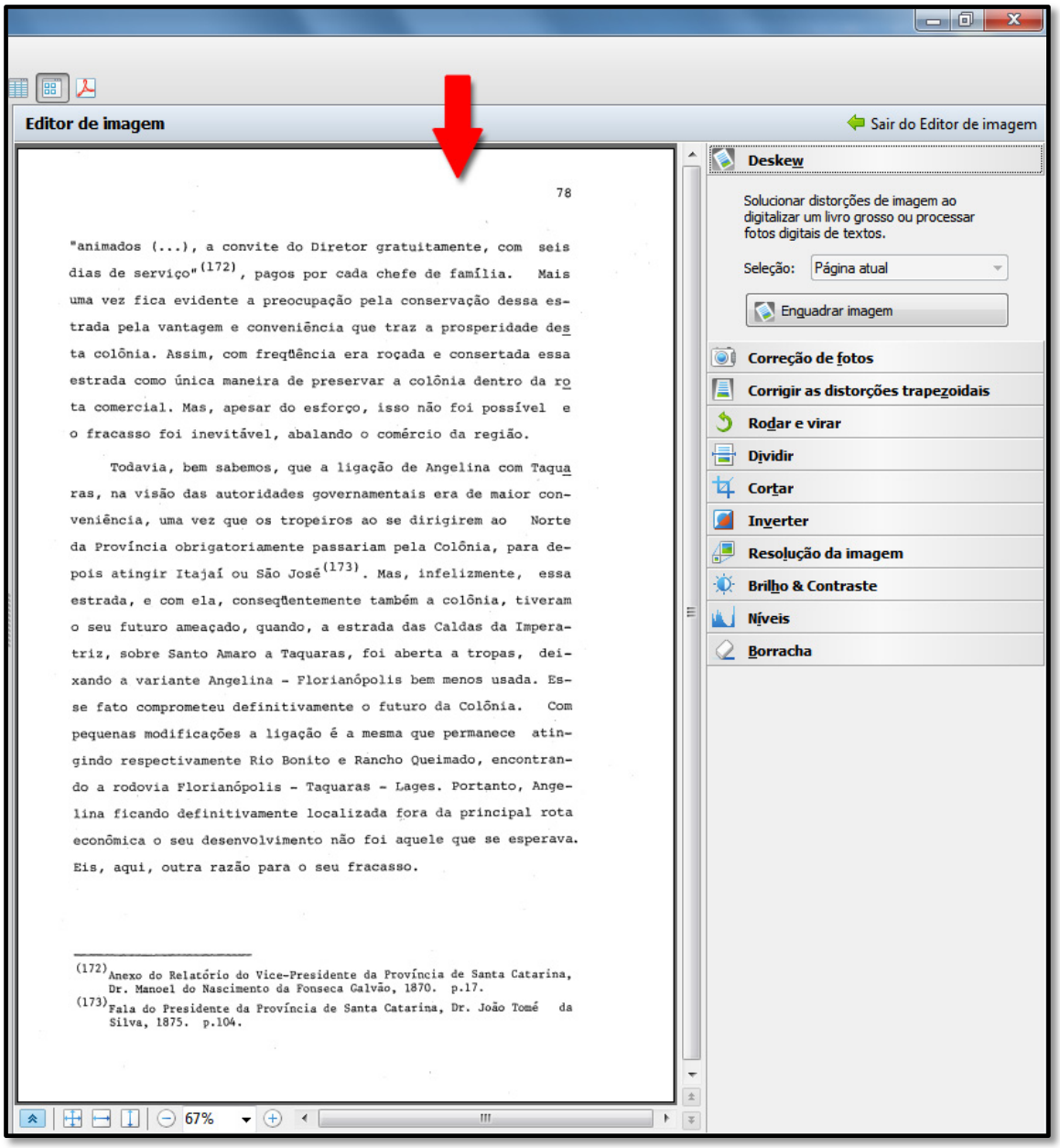


Imagem 14 – Enquando imagem

O resultado do processo de enquadramento pode ser visto na imagem 15.



78

"animados (...), a convite do Diretor gratuitamente, com seis dias de serviço"(172), pagos por cada chefe de família. Mais uma vez fica evidente a preocupação pela conservação dessa estrada pela vantagem e conveniência que traz a prosperidade desta colônia. Assim, com freqüência era roçada e consertada essa estrada como única maneira de preservar a colônia dentro da rota comercial. Mas, apesar do esforço, isso não foi possível e o fracasso foi inevitável, abalando o comércio da região.

Todavia, bem sabemos, que a ligação de Angelina com Taquaras, na visão das autoridades governamentais era de maior conveniência, uma vez que os tropeiros ao se dirigirem ao Norte da Província obrigatoriamente passariam pela Colônia, para depois atingir Itajaí ou São José(173). Mas, infelizmente, essa estrada, e com ela, conseqüentemente também a colônia, tiveram o seu futuro ameaçado, quando, a estrada das Caldas da Imperatriz, sobre Santo Amaro a Taquaras, foi aberta a tropas, deixando a variante Angelina - Florianópolis bem menos usada. Esse fato comprometeu definitivamente o futuro da Colônia. Com pequenas modificações a ligação é a mesma que permanece atingindo respectivamente Rio Bonito e Rancho Queimado, encontrando a rodovia Florianópolis - Taquaras - Lages. Portanto, Angelina ficando definitivamente localizada fora da principal rota econômica o seu desenvolvimento não foi aquele que se esperava. Eis, aqui, outra razão para o seu fracasso.

(172) Anexo do Relatório do Vice-Presidente da Província de Santa Catarina, Dr. Manoel do Nascimento da Fonseca Galvão, 1870. p.17.
(173) Fala do Presidente da Província de Santa Catarina, Dr. João Tomé da Silva, 1875. p.104.

Imagem 15 – Resultado do enquadramento

3.4 Endireitando linhas de texto

Um defeito possível de ser encontrado também são linhas de texto desalinhadas, como observado na imagem 16. Estas linhas de texto desalinhadas permanecem mesmo após a utilização do enquadramento de imagem visto anteriormente. Para estes situações é preciso utilizar o recurso *Endireitar Linhas de Texto*, acessível pelo menu na opção *Correção de fotos*.

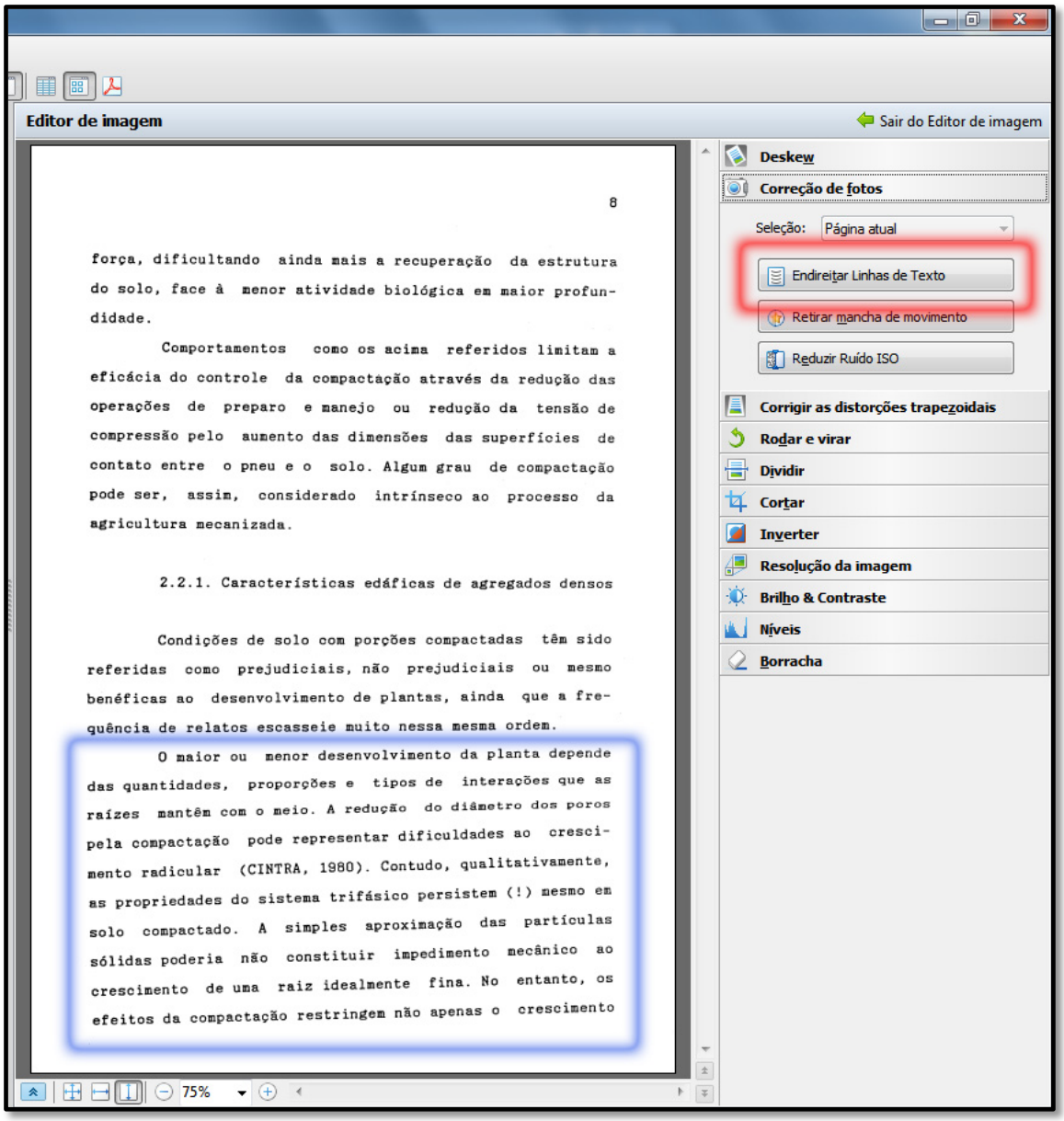


Imagem 16 – Endireitando texto

As linhas de texto já alinhadas podem ser vistas na imagem 17.

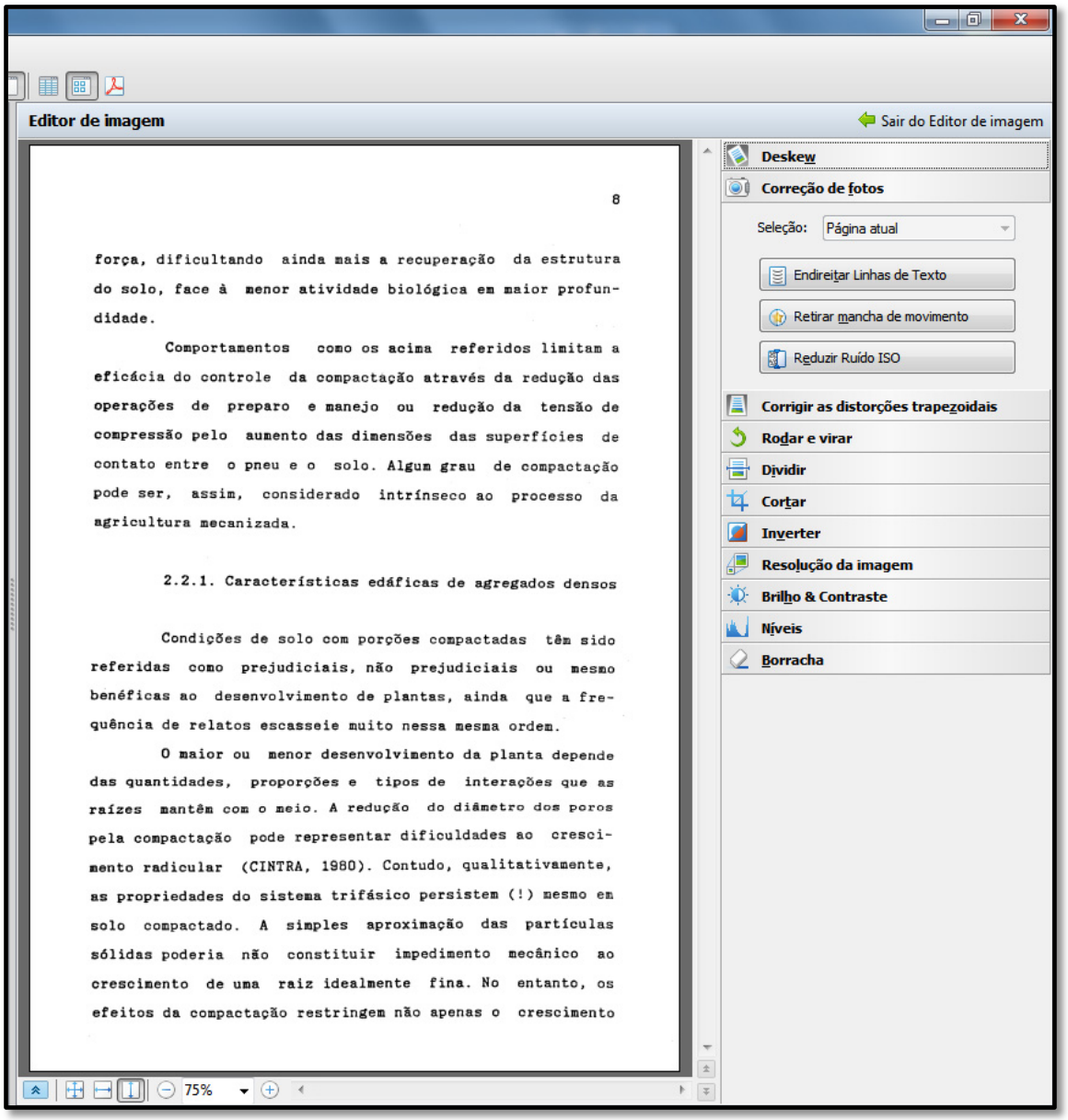


Imagem 17 – Texto endireitado

3.5 Removendo manchas e marcações

Mesmo após o processo de ajuste de brilho e contraste, podem permanecer nas páginas digitalizadas manchas de sombreamento. Estas manchas podem ser removidas, assim como riscos, marcações, rasgos e outros sinais indesejáveis nas páginas. Para esta função é utilizada a opção *Editar imagem* na interface, como vista na imagem 18.

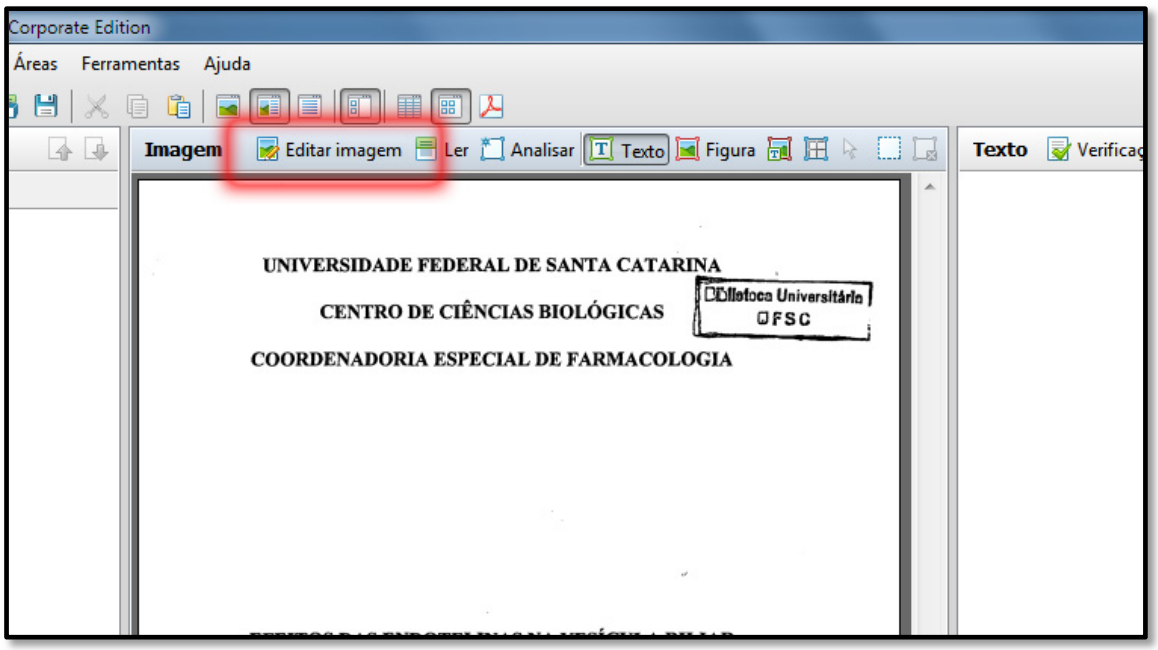


Imagem 18 – Editando imagem

No meu de edição de documentos, selecione a opção *Borracha*, como mostrado na imagem 19.

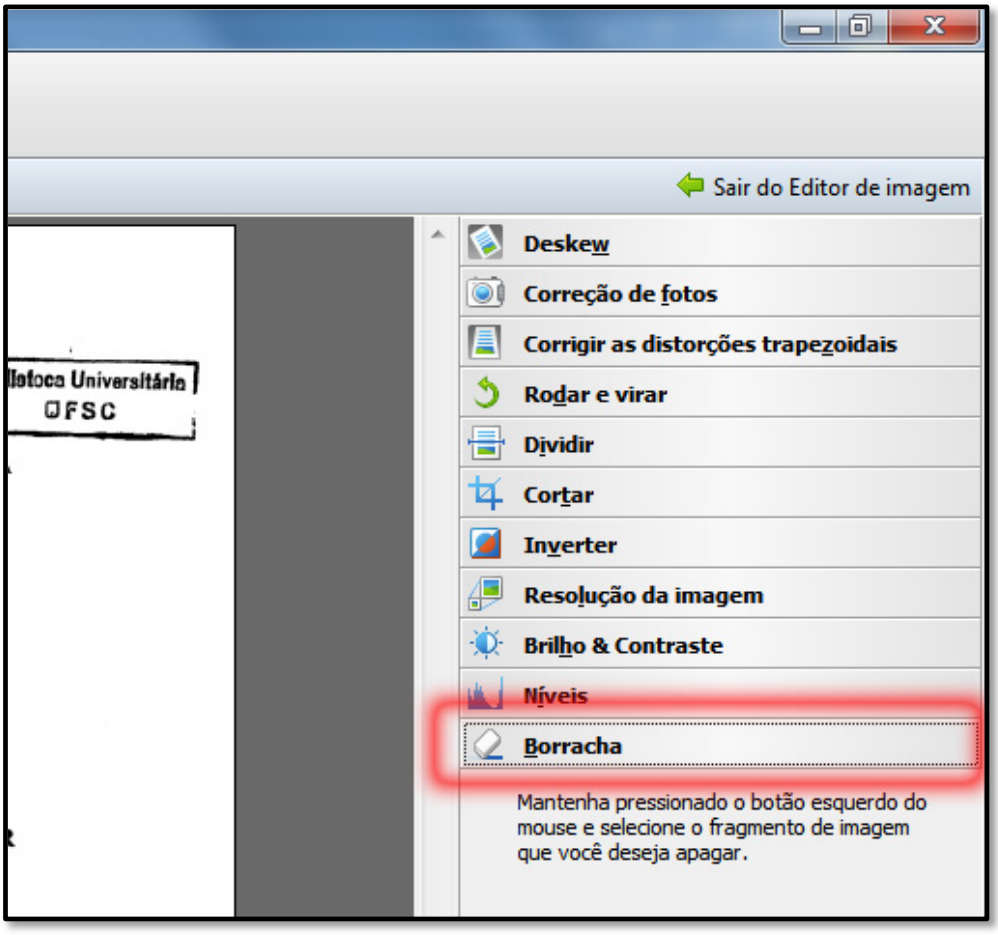


Imagem 19 - Borracha

No exemplo demonstrado a seguir, temos um carimbo que necessita ser removida da imagem da página. Para esta ação basta selecionar com o mouse a área que será apagada (visto na imagem 20).

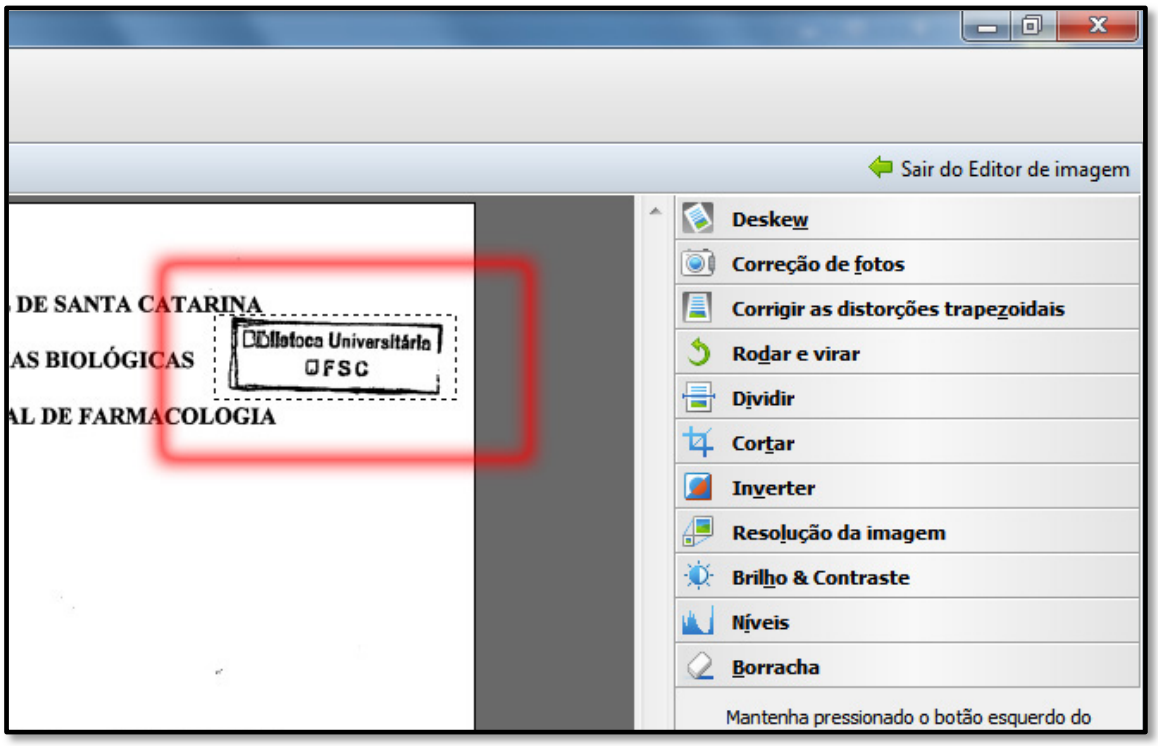


Imagem 20 – Selecionada área a ser apagada

O resultado da ação pode ser verificado na imagem 21.

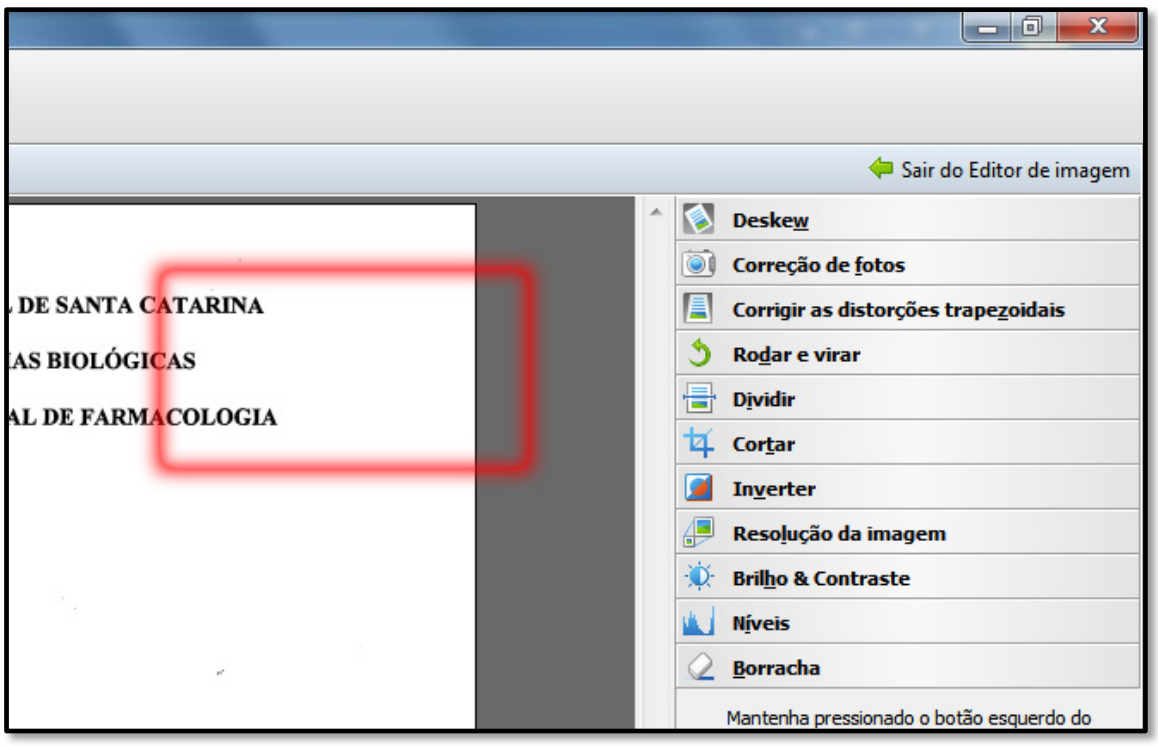


Imagem 21 – Área apagada

4 – Reconhecimento de caracteres

O reconhecimento de caracteres nas imagens trabalhadas no FineReader é feito através da tecnologia de OCR *(Optical Character Recognition)*. Estes caracteres reconhecidos ficam armazenados no próprio documento, podendo ou não, serem visualizados sobre ou sob as imagens originais escaneadas (de acordo com as configurações do software).

Riscos, marcas e sujeiras que ficaram sobre a imagem podem interferir no processo de reconhecimento de caracteres, sendo recomendo a sua remoção prévia com o uso da ferramenta de borracha.

4.1 Textos na vertical

Alguns textos contido nas páginas podem estar alinhados na vertical e neste caso o FineReader poderá não reconhecer corretamente os caracteres, como aparece na imagem 22.

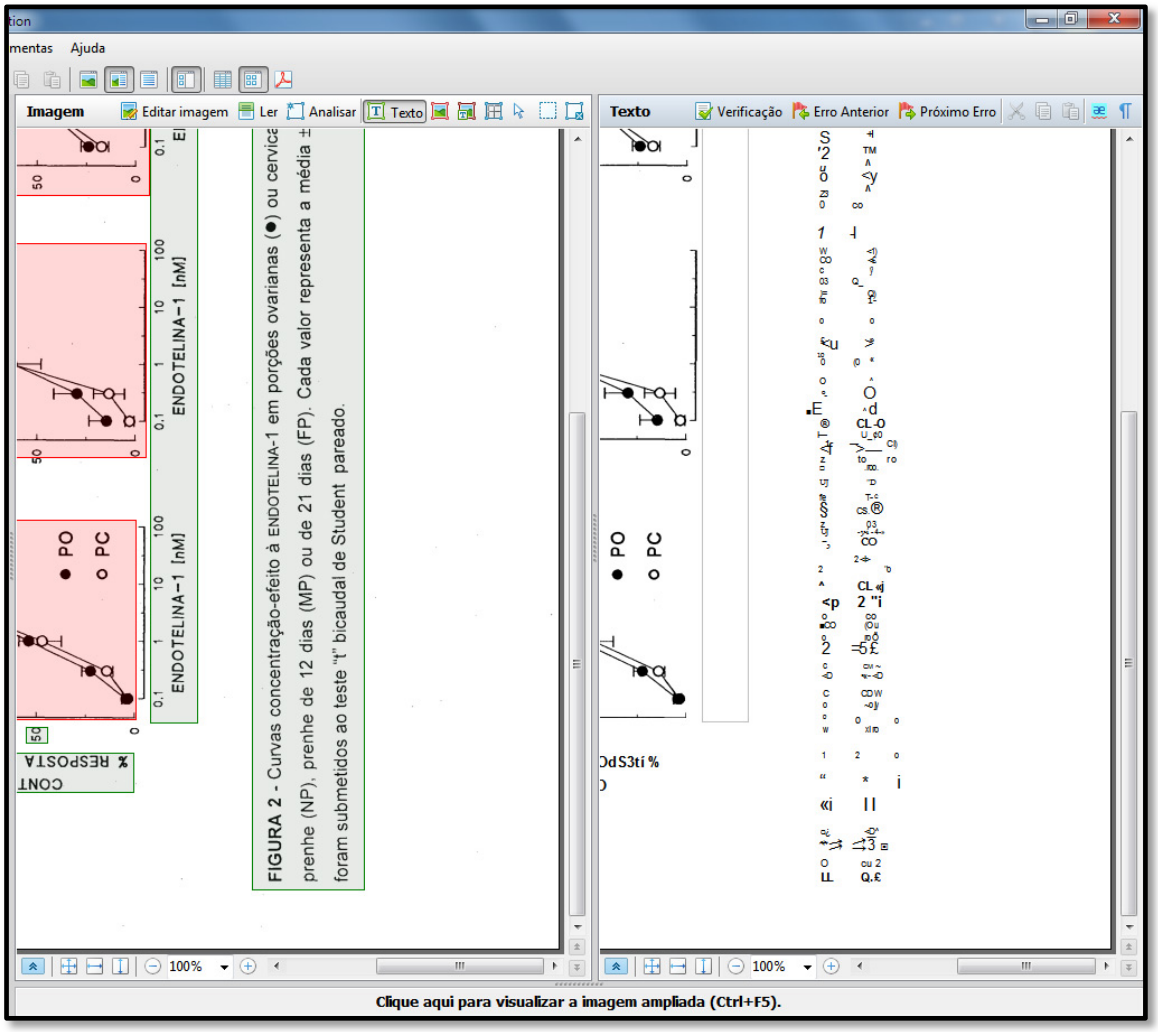


Imagem 22 – Texto reconhecido incorretamente

Para corrigir este situação é preciso clicar com o botão da direita do mouse n a área do texto delimitada e em seguida, apontar para a opção *Orientação do texto.* Além da opção Normal que é a padrão do software, pode se utilizar as opções para girar a esquerda, direita e também no sentido inverso (de cabeça para baixo).

No caso abaixo (imagem 23) devido ao texto original estar na posição vertical, o sistema de OCR não obteve êxito no reconhecimento dos caracteres. Desta forma será utilizado a opção *Girar para a esquerda*, para que o texto seja corretamente reconhecido.

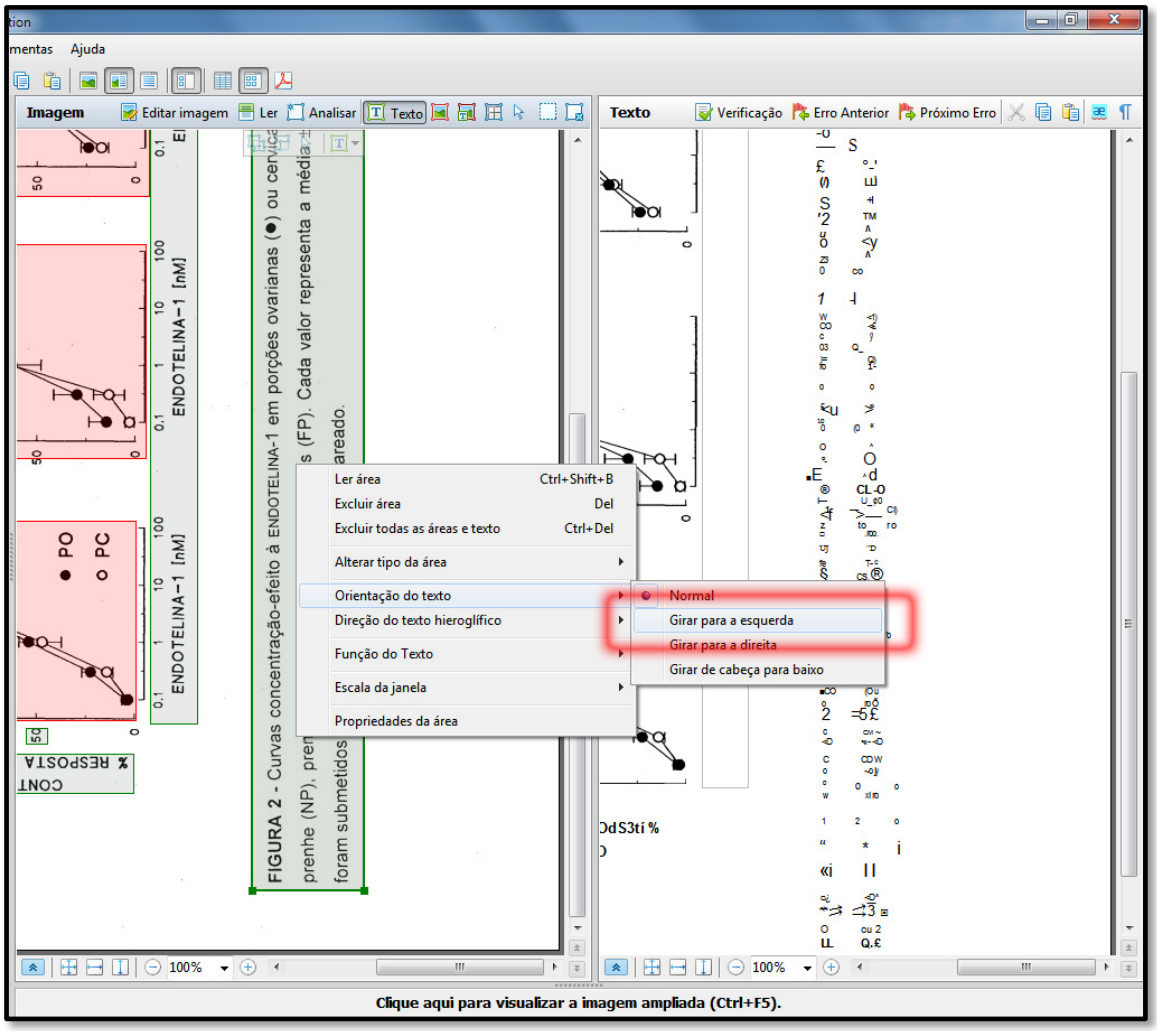


Imagem 23 – Menu de ajuste do OCR

Como resultado a ação, podemos ver o correto reconhecimento na imagem 24.

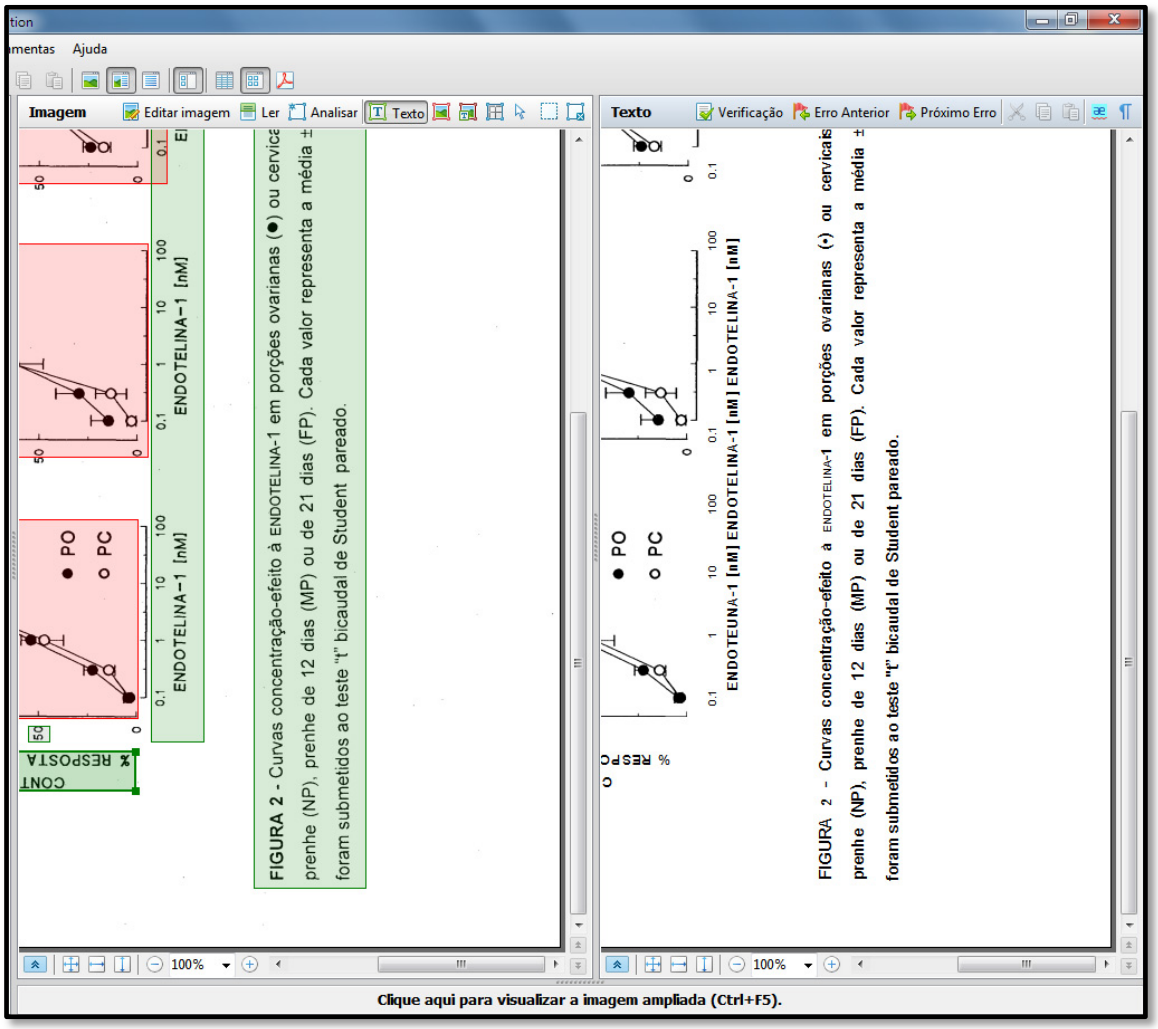


Imagem 24 – Resultado do ajuste vertical

5 – Dicas

5.1 Área de remoção

Ao utilizar a ferramenta de remoção (Borracha), é preciso dar atenção à área selecionada da imagem. Esta ferramenta realiza cálculos sobre a área selecionada, criando um padrão de cor a partir do resultado para utilizar no preenchimento. Quando a imagem for do tipo bitonal (preto e branco), o resultado sempre será a cor branca ou preta. Para imagens em tons de cinza e imagens coloridas, será utilizada uma média de todas as cores que estiverem na área selecionada.

Usaremos como exemplo uma página escaneada que possui uma área de sombra escura à esquerda (imagem 25).

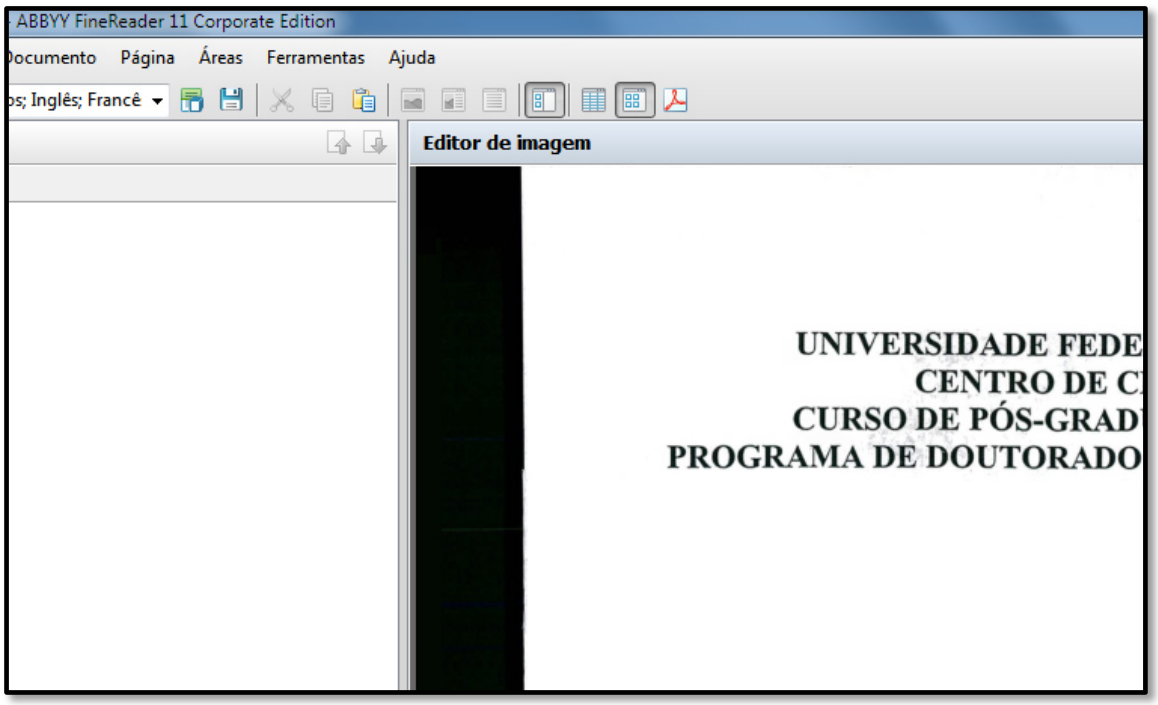


Imagem 25 – Página com área de sombra.

Neste caso ao selecionar-se uma área escura maior que a área clara, a ferramenta calcula que a cor para preenchimento deva ser a escura (imagem 26).

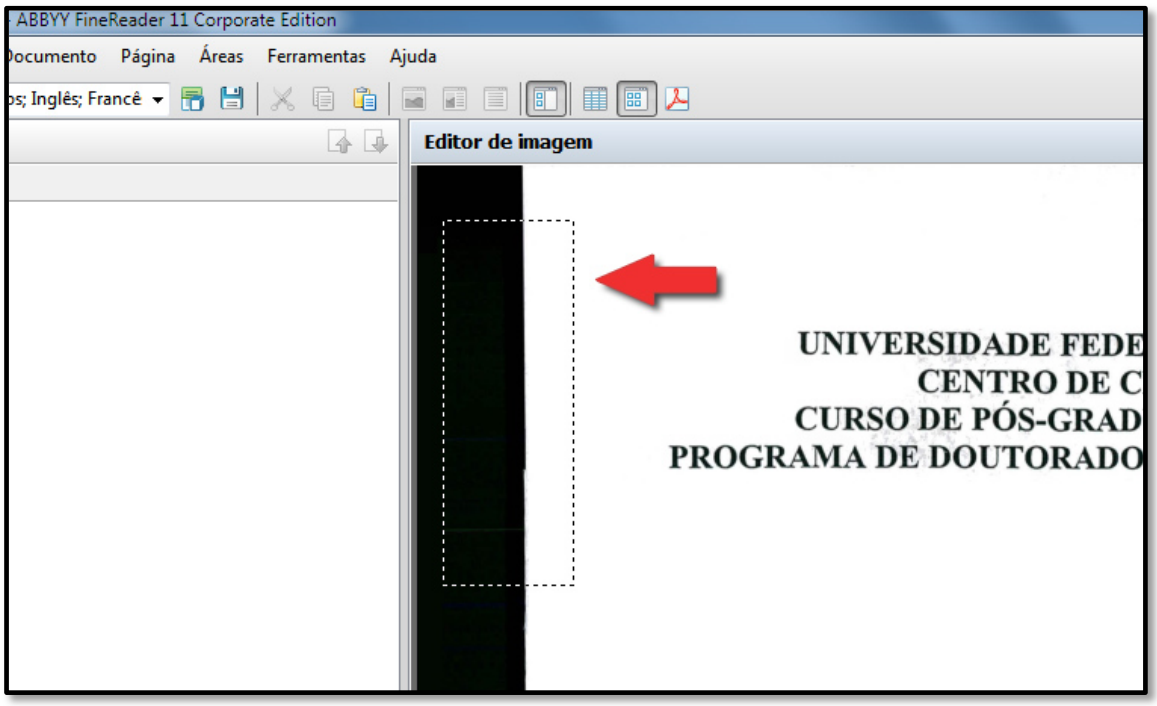


Imagem 26 – Selecionando a área para correção.

Como a área preta compreende a maior parte da imagem selecionada, a ferramenta realizou o preenchimento com esta cor (imagem 27). Para evitar este erro, basta sempre selecionar a maior área possível da cor que for o fundo da imagem.

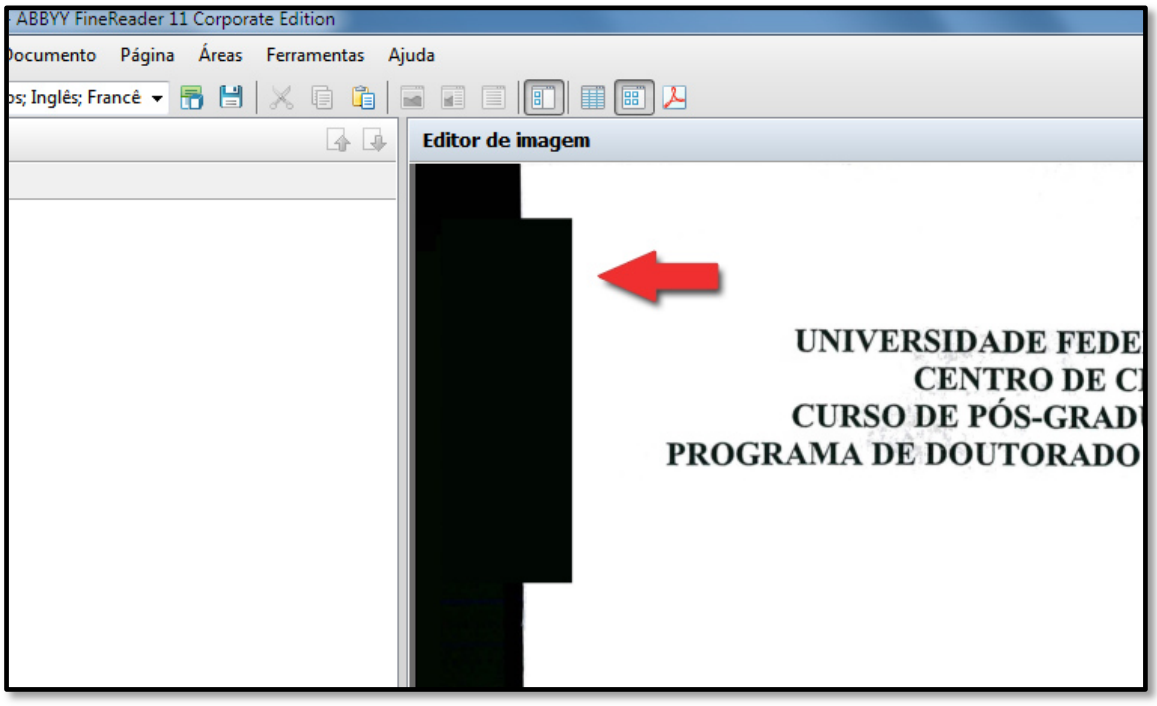


Imagem 27 – Preenchimento incorreto.

Na imagem 28 podemos ver um exemplo utilizando uma página escaneada no modo colorido, onde é maior a possibilidade gerar preenchimentos de cores que diferente do fundo da imagem.

Note que na esquerda foi selecionada uma pequena área para ser apagada, gerando um tom cinza para preenchimento. Já na direita, ao selecionar-se uma área mais ampla, o resultado foi um tom branco para preenchimento.

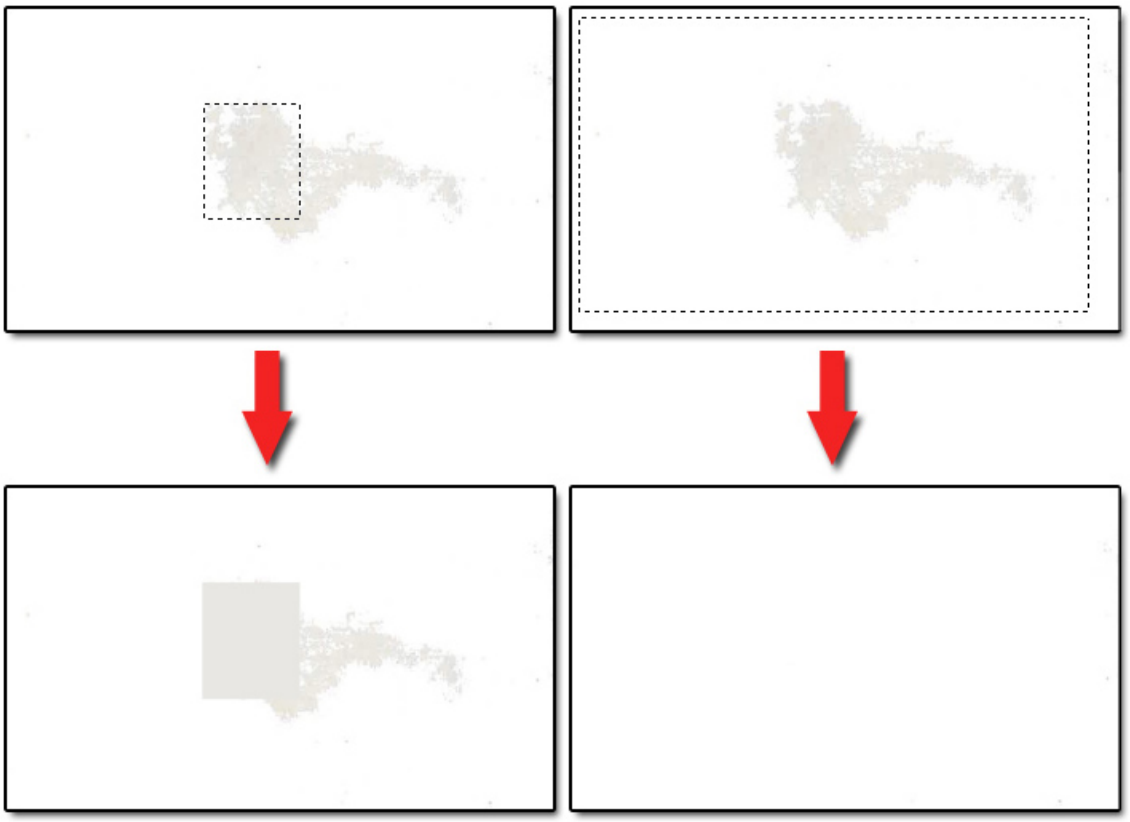


Imagem 28 – Exemplo de tamanho de área para selecionar.

5.2 Fontes decorativas

Algumas páginas escaneadas poderão contar fontes decorativas, que dificilmente são reconhecidas pelo sistema de OCR do software, gerando informações imprecisas quanto aos caracteres.

Na imagem 29 percebe-se que a fonte decorativa do texto não consegue ser reconhecida corretamente pelo OCR. O resultado exibido à direita deixa claro o erro.

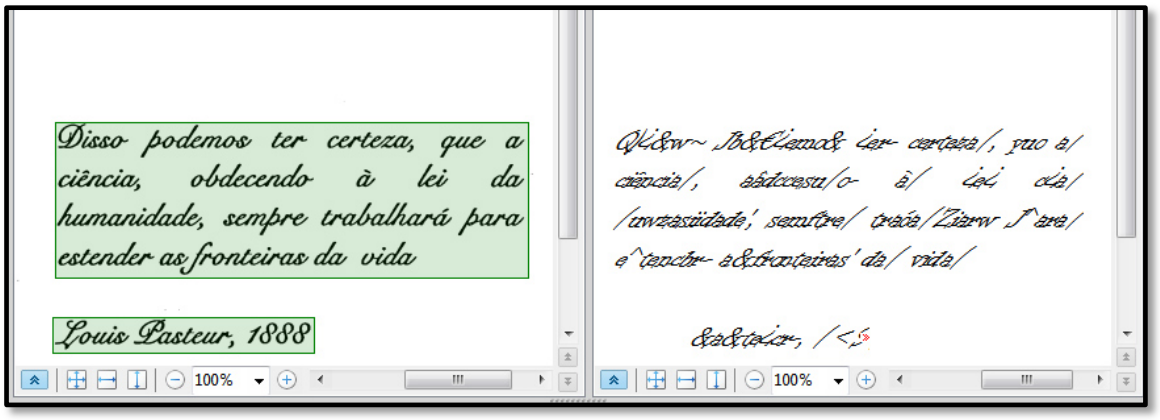


Imagem 29 – Texto contando fontes decorativas.

Para estas situações faz-se necessário o preenchimento manual do texto que ficará na camada de caracteres. Como visualizado na imagem 30, basta selecionar com o curso a área do texto e digitar os caracteres um a um.

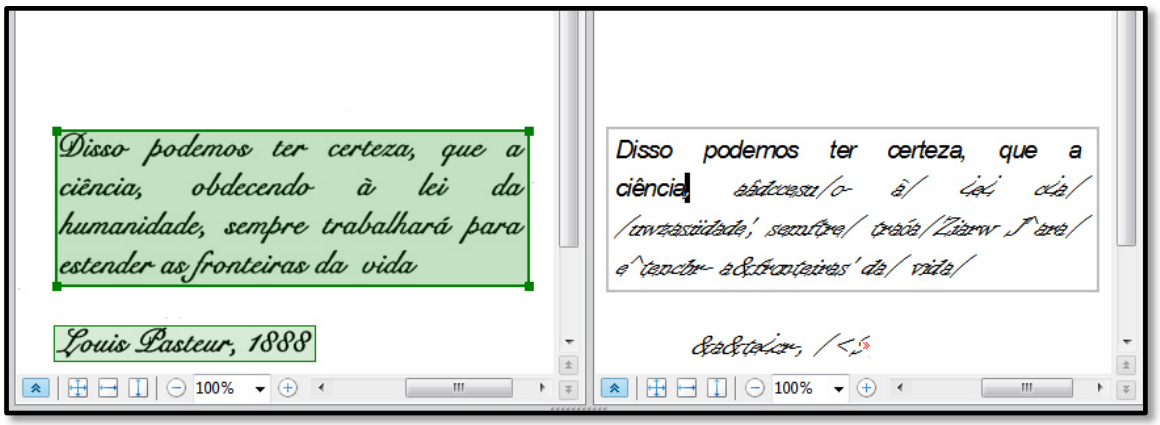


Imagem 30 – Preenchimento manual da camada de texto.

5.3 Elementos gráficos

Alguns materiais poderão conter elementos gráficos como organogramas, diagramas e fluxogramas que podem não ser corretamente identificados como texto ou imagem. Para estas situações é possível realizar um melhor reconhecimento de OCR escolhendo-se a área como uma tabela.

Na indicação mostrada na imagem 31 percebe-se que o FineReader mesclou diferentes percepções como figuras, textos e tabelas. O reconhecimento de caracteres não foi prejudicado como um todo, porém a organização das informações ficará prejudicada visualmente.

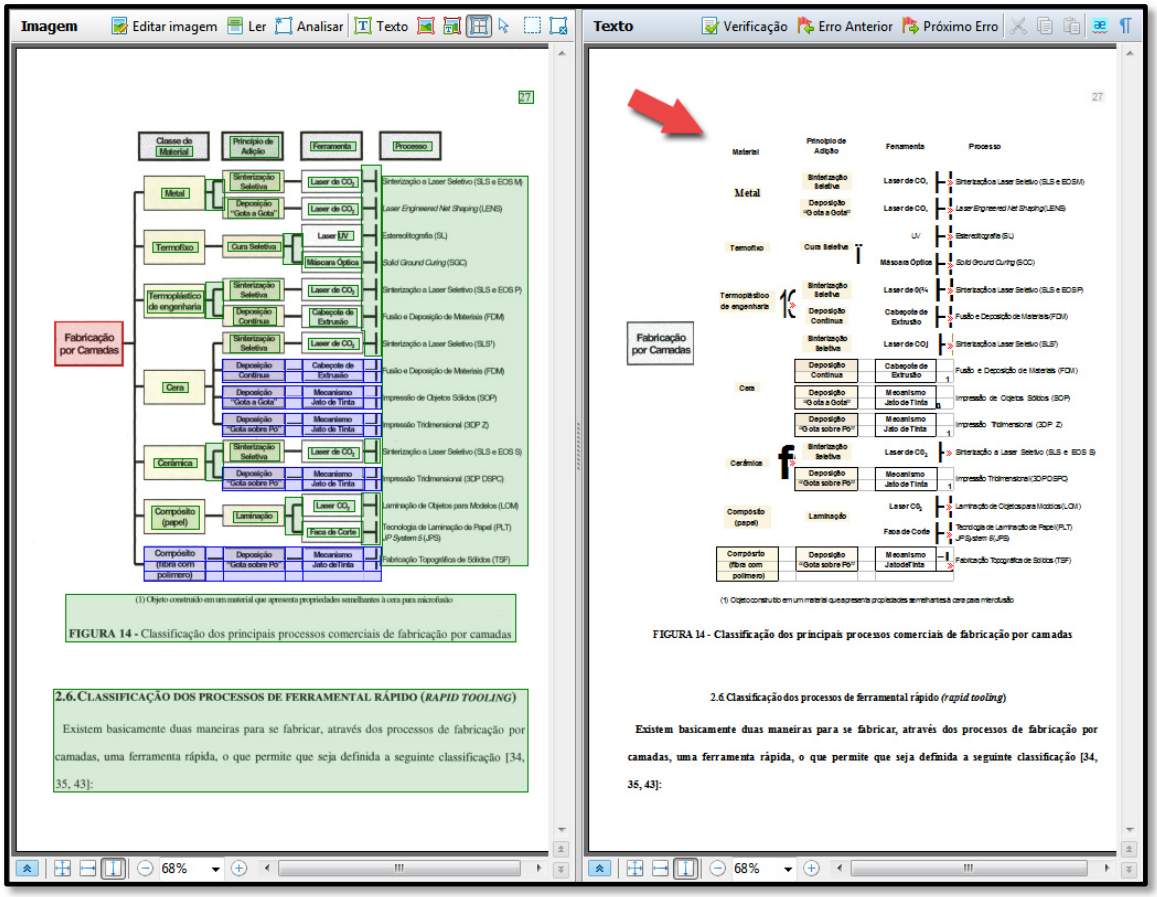


Imagem 31 – Resultado da tabela após OCR padrão.

Para contar esta situação podemos forçar o reconhecimento desta área como sendo uma tabela.

Como demonstrado na imagem 32, basta clicar no ícone de seleção de área de tabela (destacado em vermelho) e em seguida selecionar a área correspondente. Após a seleção, clique com o botão direito sobre a área (na cor azul) e selecione no menu a opção *Alterar tipo de área* e em seguida a opção *Tabela*.

Também é possível ativar este recurso com o atalho **CRTL**+**3**.

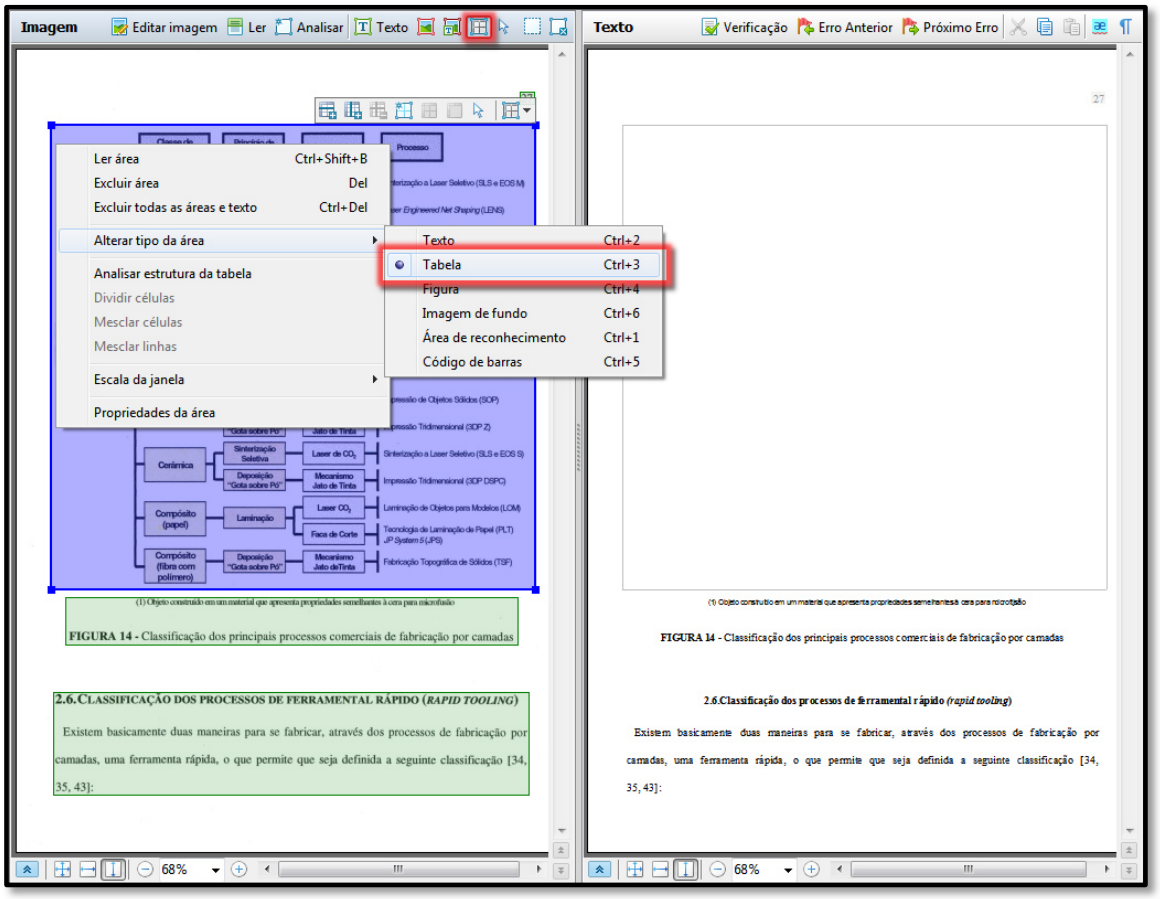


Imagem 32 – Menu de opção de tipo de área.

Caso o resultado da ação não seja satisfatório, tente escolher uma das outras opções do tipo de área, até que fique visível da melhor maneira possível.

Após selecionar o tipo de área, clique na opção Ler para que o FineReader inicie o reconhecimento.

O resultado deste processo pode ser visualizado na imagem 33.

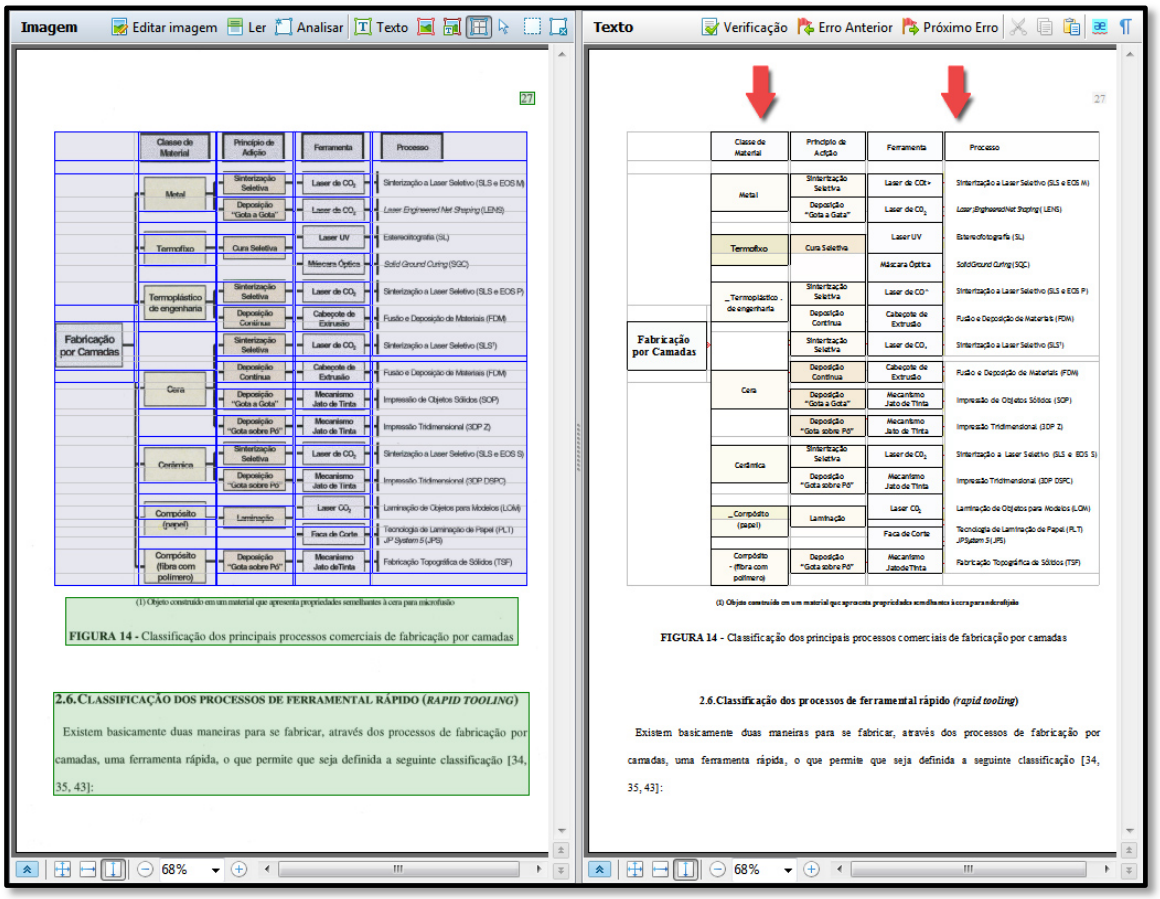
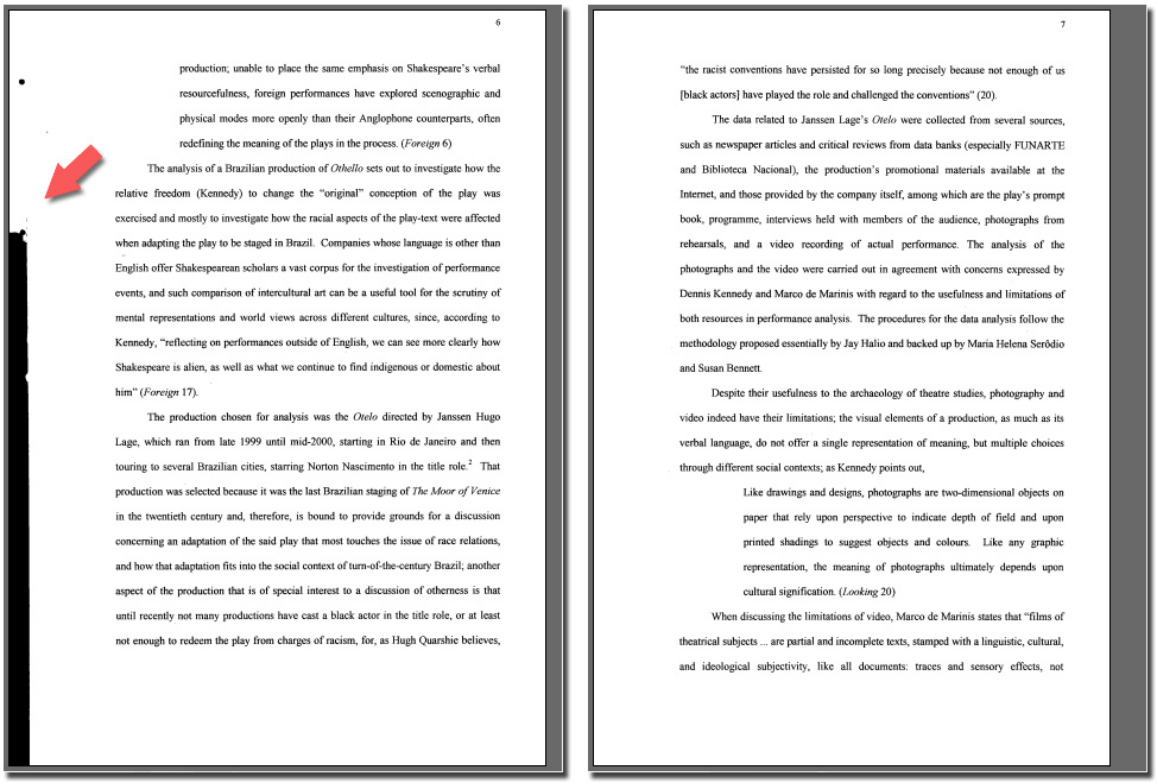


Imagem 33 – Resultado após selecionar a opção de tabela.

5.4 Área de corte das páginas

Como resultados do corte das páginas (quando utilizado guilhotina) ou da própria impressão incorreta do material, algumas páginas escaneadas podem ficar com proporções diferentes. Este tipo de diferença pode ser verificado no exemplo mostrado na imagem 34, onde a página da esquerda ficou com uma parte sobrando em relação as demais páginas da obra, visto a direita na mesma imagem.



6

production; unable to place the same emphasis on Shakespeare's verbal resourcefulness, foreign performances have explored scenographic and physical modes more openly than their Anglophone counterparts, often redefining the meaning of the plays in the process. (*Foreign* 6)

The analysis of a Brazilian production of *Othello* sets out to investigate how the relative freedom (Kennedy) to change the "original" conception of the play was exercised and mostly to investigate how the racial aspects of the play-text were affected when adapting the play to be staged in Brazil. Companies whose language is other than English offer Shakespearean scholars a vast corpus for the investigation of performance events, and such comparison of intercultural art can be a useful tool for the scrutiny of mental representations and world views across different cultures, since, according to Kennedy, "reflecting on performances outside of English, we can see more clearly how Shakespeare is alien, as well as what we continue to find indigenous or domestic about him" (*Foreign* 17).

The production chosen for analysis was the *Otelo* directed by Janssen Hugo Lage, which ran from late 1999 until mid-2000, starting in Rio de Janeiro and then touring to several Brazilian cities, starring Norton Nascimento in the title role.[2] That production was selected because it was the last Brazilian staging of *The Moor of Venice* in the twentieth century and, therefore, is bound to provide grounds for a discussion concerning an adaptation of the said play that most touches the issue of race relations, and how that adaptation fits into the social context of turn-of-the-century Brazil; another aspect of the production that is of special interest to a discussion of otherness is that until recently not many productions have cast a black actor in the title role, or at least not enough to redeem the play from charges of racism, for, as Hugh Quarshie believes,

7

"the racist conventions have persisted for so long precisely because not enough of us [black actors] have played the role and challenged the conventions" (20).

The data related to Janssen Lage's *Otelo* were collected from several sources, such as newspaper articles and critical reviews from data banks (especially FUNARTE and Biblioteca Nacional), the production's promotional materials available at the Internet, and those provided by the company itself, among which are the play's prompt book, programme, interviews held with members of the audience, photographs from rehearsals, and a video recording of actual performance. The analysis of the photographs and the video were carried out in agreement with concerns expressed by Dennis Kennedy and Marco de Marinis with regard to the usefulness and limitations of both resources in performance analysis. The procedures for the data analysis follow the methodology proposed essentially by Jay Halio and backed up by Maria Helena Serôdio and Susan Bennett.

Despite their usefulness to the archaeology of theatre studies, photography and video indeed have their limitations; the visual elements of a production, as much as its verbal language, do not offer a single representation of meaning, but multiple choices through different social contexts; as Kennedy points out,

> Like drawings and designs, photographs are two-dimensional objects on paper that rely upon perspective to indicate depth of field and upon printed shadings to suggest objects and colours. Like any graphic representation, the meaning of photographs ultimately depends upon cultural signification. (*Looking* 20)

When discussing the limitations of video, Marco de Marinis states that "films of theatrical subjects ... are partial and incomplete texts, stamped with a linguistic, cultural, and ideological subjectivity, like all documents: traces and sensory effects, not

Imagem 34 – Página com proporção incorreta.

Como meio para corrigir a área destas páginas, deve-se utilizar a ferramenta de corte. Para iniciar o processo, basta selecionar qualquer uma das páginas que estiverem no tamanho correto. A página que for escolhida como molde para a demais páginas, deve ser aquela que se aproximar do tamanho original da obra.

Em seguida, escolha a ferramenta Cortar e posicione os cursores de seleção nas extremidades da página como visto na imagem 35.

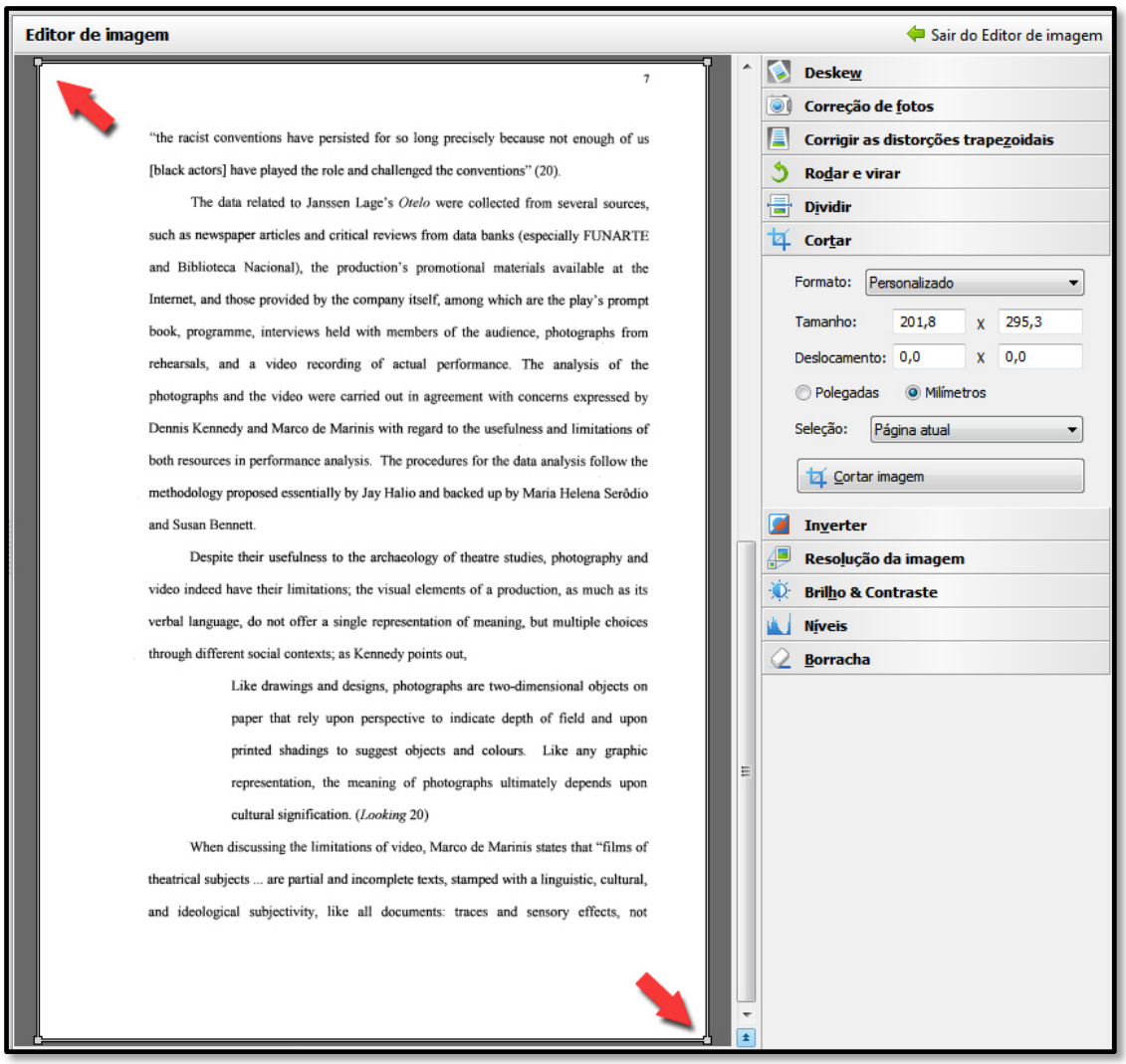


Imagem 35 – Selecionando toda a página correta.

Após selecionar toda a página pelas suas extremidades como visto na imagem 35, deve ser mantido a tela na ferramenta Cortar. Agora selecione a página que estiver com o tamanho incorreto e mova a área de seleção até a posição mais adequada.

No caso da página utilizada aqui como exemplo aqui, foi necessário mover a área até a lateral direita, deixando de fora somente a parte indesejada que será cortada da imagem.

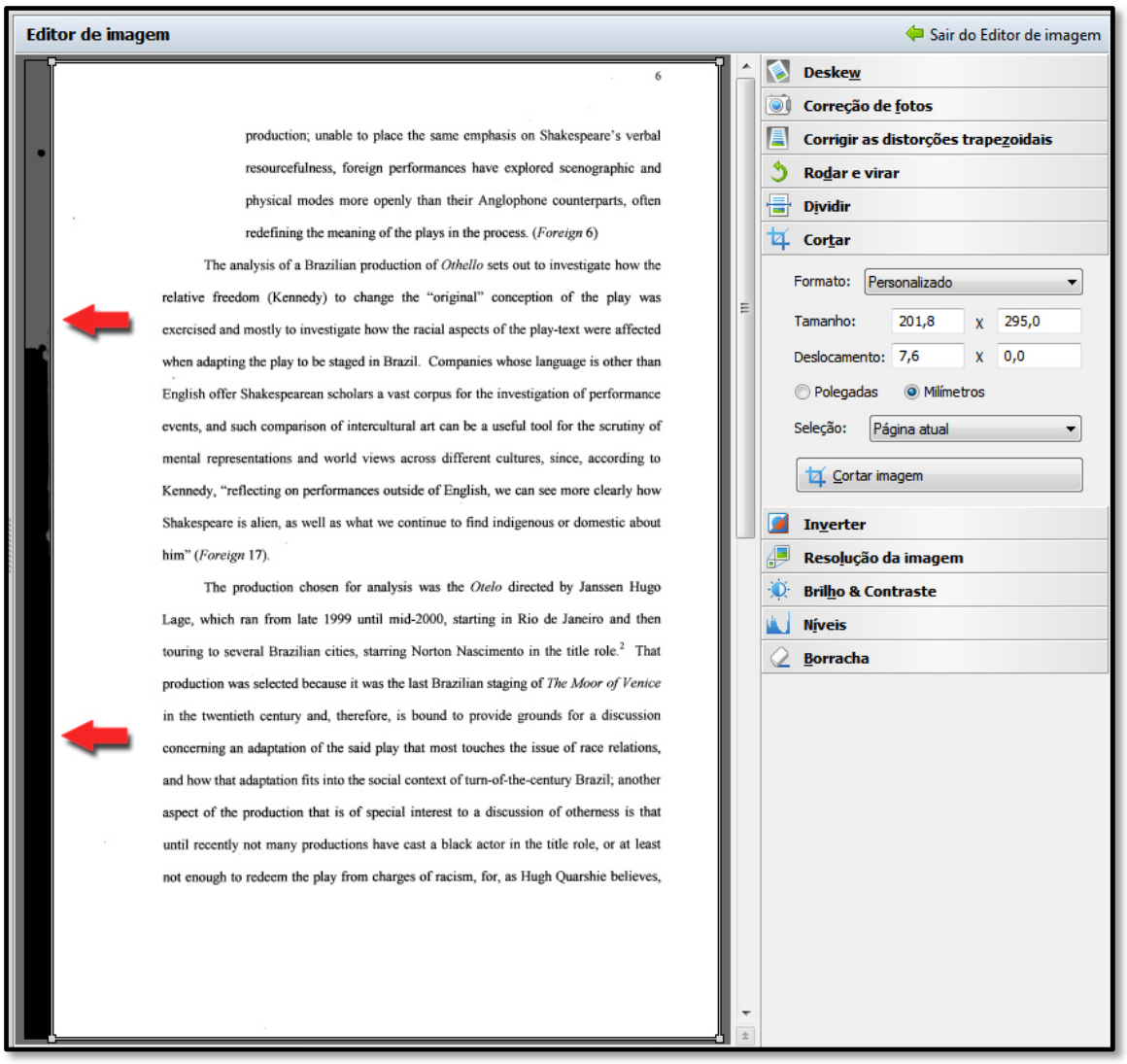


Imagem 36 – Área de corte.

5.5 Substituindo páginas

No processo de edição das páginas pode ocorrer algum tipo de efeito indesejado numa única página, devido ao uso da ferramenta de corte ou borracha. Para não recarregar todas as imagens novamente, é possível carregar somente a página danificada.

Com as páginas ainda carregadas no FineReader, clique no botão *Abrir arquivo PDF/imagem* ou através do comando **CRTL**+**O**. Selecione a imagem original da página que ficou defeituosa e clique em *Abrir*.

No exemplo mostrado na imagem 37, foi recarregada a página de número 10. Porém ela foi posicionada na posição 215 da fila de páginas, após a última página que estava carregada. A página 10 com defeito continua na posição anterior (não há substituição imediata).

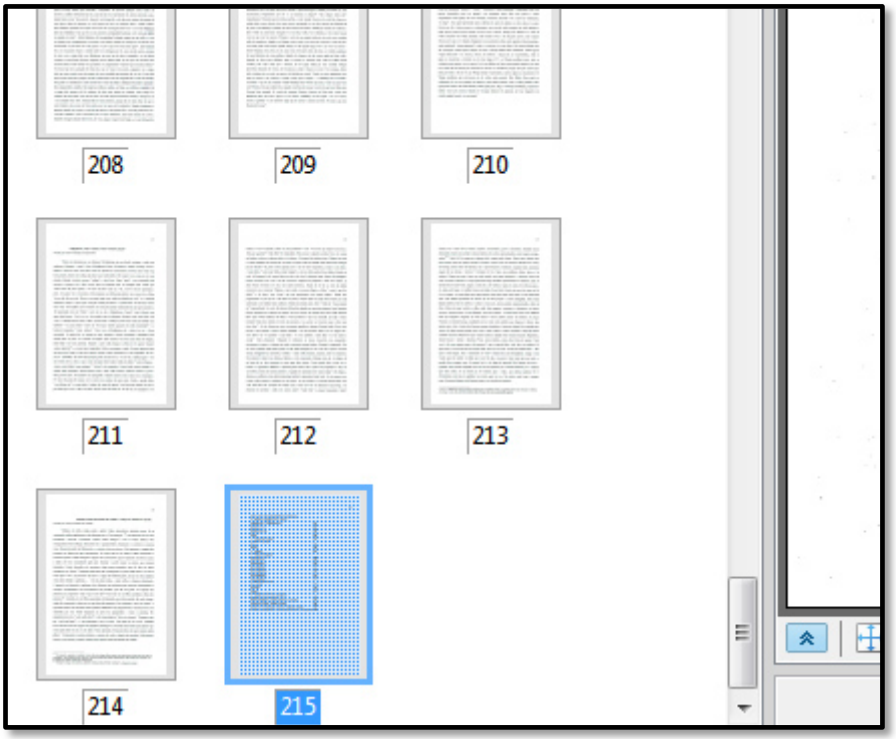


Imagem 37 – Página recarregada

Para substituir corretamente, é necessário selecionar a página recém carregada e clicar com o botão da direita do mouse para abrir o menu de propriedades. Como exibido na imagem 38, dentro das propriedades do arquivo conta a numeração que ele tem com relação a fila de arquivo.

No campo destacado em vermelho, deverá ser inserido o número de página pela qual desejamos substituir (em nosso exemplo, substituindo o número 215 pelo número 10).

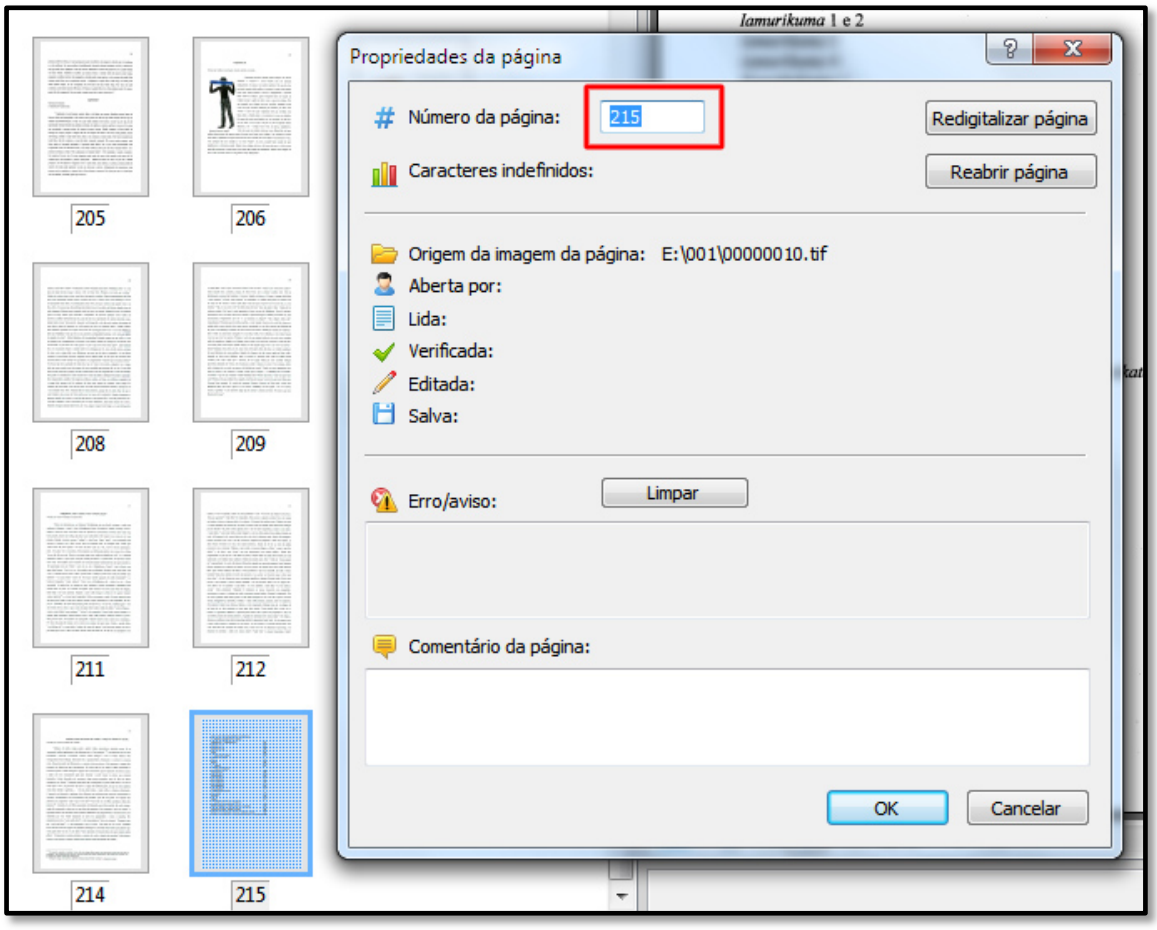


Imagem 38 – Inserindo número correto da página

Mesmo com este procedimento, a página recarregada não será sobrescrita por cima da outra. Em nosso exemplo, ficarão duplicadas as páginas, existindo agora a página 10 e a página 11, conforme pode ser visualizado na imagem 39.

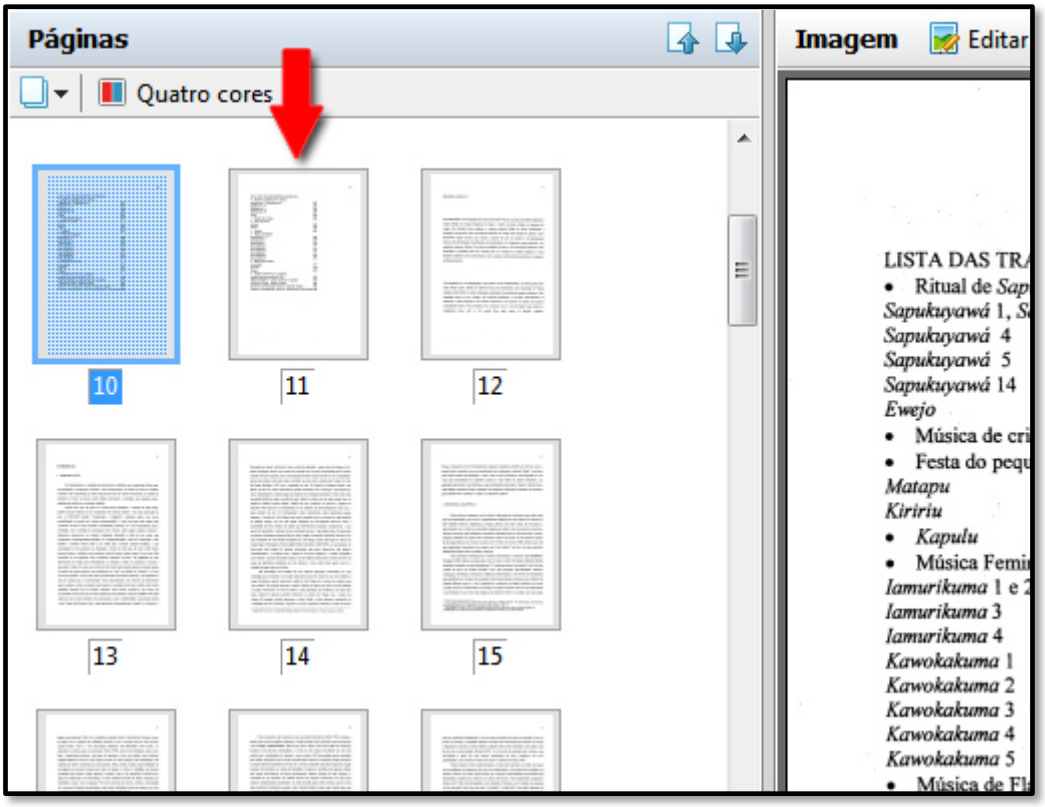


Imagem 39 – Páginas duplicadas

Para proceder, basta selecionar a página danificada (em nosso caso, a página 11) e apagar com o botão **Delete** ou através da opção excluir acessada com o botão direito do mouse.