Proposta de Fluxo de Trabalho para Organização de Repositórios Abertos de Maneira Colaborativa

Leandro Batista de Almeida[1], Luiz Ernesto Merkle[1], Edson Armando da Silva[2]

[1]Departamento Acadêmico de Informática – Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Sete de Setembro, 3165 – 80.230-901 – Curitiba – PR – Brazil

[2]Departamento de História – Universidade Estadual de Ponta Grosa (UEPG)
Av. Gal. Carlos Cavalcanti, 4748 – 84.030-900 – Ponta Grossa – PR - Brazil

{leandro,merkle}@utfpr.edu.br, edasilva@uepg.br

Abstract. *This paper describes a proposal for a workflow that handles digital documents, ranging from their capture, through scanners, or their reception as digital files, until their release in open access repositories and possible retrieval. This workflow encompasses the use of free software applications and systems, that deal with the capture of the files, the post processing of the images and files, such as optical character recognition, and the superposition of the text layer in the image layer of the document, and the description indexing, and cataloging the information, the content, and the availability in the repository. The repository also allows the analysis and the commentary of documents collaboratively.*

Resumo. *O presente trabalho descreve uma proposta para um fluxo de trabalho que trata documentos digitais, abrangendo desde a sua captação, por meio de scanners ou recepção de arquivos digitais, até a sua disponibilização em repositórios abertos. Dentro desse fluxo de trabalho são incluídas ferramentas baseadas em softwares livres que tratam da captura dos arquivos, do pós processamento da imagem, da realização de reconhecimento ótico de caracteres, a superposição da camada de texto na camada de imagem do documento, a catalogação do texto e do seu conteúdo e a disponibilização em repositórios. Também trata de um fluxo que permite a análise e comentários dos documentos de maneira colaborativa.*

1. Contexto de pesquisa e uso de repositórios em ciências humanas

Desde o sempre, o ser humano mantém sua cultura por diversas mediações, de maneira a permitir a continuidade que tornasse viável a resolução de novos problemas por meio de novos conceitos, ideias e ações. Ao longo de seu desenvolvimento, mudaram-se radicalmente os suportes a informação, passando por fibras, argila, couro, papel, diapositivos, fotografias, e muitos outros meios, e mais recentemente os arquivos digitais, mas que também diferem em codificações, suportes, licenças, tempo de vida e muitos outros quesitos. Embora difira-se o suporte material, se percebeu que não é simplesmente acessório, pois de certa forma eles são concebidos e desenvolvidos

justamente para dar continuidade aos afazeres humanos, em suas múltiplas dimensões. Novos suportes tecnológicos também abrem possibilidades não antevistas, pois embora enderecem problemas anteriores, trazem outros em sua própria concepção, que precisarão ser igualmente resolvidos, dando continuidade ao fazer-se humano ao longo da história, em diferentes contextos e circunstâncias.

O trabalho aqui delineado se dá no contexto do trabalho desenvolvido no contexto da área de arquivística e de história. Essa relação de uso das tecnologias computacionais livres e de desenvolvimento de soluções novas ou adaptadas, faz com que o próprio projeto dessas tecnologias de repositórios possa ser moldado tanto por fornecedores das tecnologias como artefatos, quanto pelo trabalho que aqueles devem mediar, de interesse da área de aplicação e uso.

Desta forma, os procedimentos, as técnicas e os processos de preservação documental, assim como os interesses e as implicações, em parte mudam e moldam as próprias tecnologias e práticas associadas. Como um paralelo aos monges copistas da idade média que reproduziam os originais que estavam se deteriorando, as técnicas atuais procuram preservar os documentos digitais, de forma a poderem ser tratados e manipulados por diferentes grupos de maneira padronizada. Entretanto, a mudança de suporte, de formato, de modo de acesso, torna necessária e coetânea as eventuais mudanças de suporte e organização do digital, tal e qual ao trabalho dos copistas medievos, também introduz mudanças e transformações na própria informação.

Hoje, os conceitos de objeto digital confiável [Conarc, 2006] e de repositórios digitais [Unesco, 2012] passam então a ser preponderantes no cotidiano da pesquisa em diversas áreas, em particular às ciências humanas, que se acostumaram a recorrer à bibliotecas e acervos, arquivos e museus como elos de importância chave em suas práticas.

Um objeto digital é um arquivo que guarda informações que devam ser mantidas por um longo período de tempo. A questão central aqui é que um arquivo pode ser visto como um simples conjunto de bits que só é corretamente interpretado pelo software que o gerou. Mas, a transformação das técnicas entretanto é um dos fatores que coloca em risco a preservação dos próprios objetos digitais ,pois quando não concebidos para serem acessados, compreendidos, descritos, indexados, arquivados, recuperados, alterados e preservados, podem simplesmente se transformarem em enigmas.

É nesse contexto que postula-se aqui a escolha de padrões abertos, descritos em normas que possam ser mantidas de maneira independente dessa ou daquela empresa de software, para preservação dos objetos digitais.

Em relação a repositórios de preservação de objetos digitais, recentemente algumas preocupações surgiram, basicamente decorrentes do fato de que os documentos digitais passaram a ocupar o lugar de outros meios de armazenamento que já contavam com a confiança generalizada da comunidade de usuários, e possuíam uma longevidade já comprovada. Exemplo disso são arquivos baseados em papel ou microfilmes, onde – guardadas as proporções históricas e cuidados necessários – a preservação da informação pode ter o curso de duração de séculos, senão milênios. O armazenamento em meio digital deve seguir esse caminho, a despeito dos seus problemas de conflitos de formatos e correntes tecnológicas. A noção, recentemente apreendida, de que um

arquivo digital gerado por uma determinada aplicação tem o tempo de vida dessa mesma aplicação, corre contra a corrente de uma maneira de preservação que possui a obrigação de durar séculos. Além dos formatos, o próprio meio de armazenamento pode ser um risco em relação à durabilidade.

Os repositórios digitais abertos são uma resposta viável a esses questionamentos, juntamente com os formatos abertos e padronizados para se armazenar um objeto digital confiável. Podemos afirmar com segurança, que se uma informação precisa circular, ser recuperada e pesquisada, e seu formato ou meio de armazenamento não o permite, ela efetivamente não existe.

Outro aspecto é o modo pelo qual a pesquisa em repositórios acontecia, preponderantemente em ciências humanas. Os pesquisadores geralmente realizavam seu trabalho de maneira individual, agregando conclusões e reflexões uma vez que um conjunto de conceitos já se encontrava estabelecido e então era lançado para a comunidade. Apesar de compartilharem suas conclusões, no nível de discussão dos documentos, o compartilhamento da base documental era muito mais restrito, por condições de acesso físico mesmo - porque dificilmente existiam trabalhos conjuntos sobre o mesmo documento.

As presentes tecnologias abrem novas perspectivas para esse modo de trabalho convencional, acrescentando canais de colaboração e criação cooperativa. Nessa divisão social do trabalho intelectual, baseada em ferramentas e padrões livres, se criam grupos que compartilham documentos, visões sobre os documentos e as próprias anotações, fazendo surgir uma sinergia em torno do arquivo em si. O pesquisador então passa a ter uma interação muito mais próxima aos documentos e repositórios, que podem ser múltiplos, desde que se sigam alguns preceitos de objetos digitais confiáveis e repositórios abertos. As possibilidades abertas por esta forma de produção em rede, entretanto, não rompe com os princípios gerais da arquivística no trabalho de limpar, digitalizar, catalogar e tratar os documentos, mas em parte a expande.

Nesse contexto, os repositórios abertos possuem um papel de destaque, sendo o meio de armazenamento intermediário para o fluxo de trabalho e o depósito de destino dos documentos já organizados. Ainda que organizados, a informação relacionada a eles não se configura como estática, já que uma vez depositado no repositório, o documento passa a atender uma comunidade ainda maior de pesquisadores e usuários, permitindo que o trabalho coletivo desses ainda aumente a massa de conceitos e enriqueça a memória coletiva a esse respeito.

2. Teoria Crítica da Tecnologia e Repositórios Abertos

A preservação da memória exige desenvolvimentos em tecnologia, lembrando que informática e computação ainda são disciplinas preponderantemente formais, onde a relação com o mundo onde se atua só recentemente passou a tomar foco mais aprofundado. O computar em história, em patrimônio, exige virá-la do avesso, realçando o contexto, a transformação, a atividade humana. Só aí aparece a crítica e o desenvolvimento desejados.

Não é suficiente criticar o efeito de uma técnica sobre a memória, quando esta mesma técnica não a levou em consideração em seu desenvolvimento. Efeitos adversos

e negativos sempre ocorrerão em tecnologia, o que se conclui disso é a realização da pesquisa. Por outro lado, não se deseja somente criticar o efeito de uma técnica sobre a memória, é necessário o reconhecimento de que a mudança da tecnologia envolve um conjunto de mudanças em outros procedimentos, mesmo que já sejam procedimentos considerados clássicos em algumas áreas, como a arquivística e o conceito de repositórios.

Os “lugares da memória” sofreram a influência da informatização. Criamos instituições especializadas para a manutenção da memória, como museus, bibliotecas e outros. O uso e o desenvolvimento da tecnologia mudou e transformou esses lugares, e essa mudança – novamente – abre possibilidades e limites. A relação desses lugares com documentos de diferentes origens e a interação com ambientes computacionais remete a discussão de Feenberg [Feenberg, 2002] sobre a construção social da tecnologia e de usos não projetados para sistemas tecnológicos.

Os atuais repositórios são fortemente baseados na diversidade e facilidade dos protocolos da Internet, considerada uma quebra de paradigma em muitos aspectos da nossa sociedade. Diversas tecnologias são tratadas como transformadoras de sociedades, como pontos de quebra de paradigmas. Além de alterarem nosso mundo, tecnologias são frequentemente entendidas como criadoras ou catalisadoras de uma nova sociedade, uma nova fase da história, entre outras. Afirmações dessa natureza já foram feitas sobre o motor a vapor, o transporte ferroviário e rodoviário, a bomba atômica, e mais recentemente, sobre a televisão e a Internet. Na obra de Raymond Williams [Williams, 2003], encontramos diversas conclusões a respeito da televisão como tecnologia transformadora, e podemos estender esse raciocínio para uma porção significativa da transformação operada pela Internet nas duas últimas décadas da história da nossa sociedade, concentrando essas transformações a partir de 1995, com a popularização das aplicações web. Tecnologias normalmente estabilizam depois de um período inicial onde várias configurações diferentes competem entre si. Uma vez estabilizadas, suas implicações sociais e políticas finalmente se tornam claras. Mas a despeito de décadas de desenvolvimento, a Internet permanece em fluxo a medida que usos inovadores continuam a aparecer [Feenberg, 2011].

Perspectivas que podem ser caracterizadas como determinismo tecnológico sustentam que tecnologias são independentes do meio em que se aplicam, neutras e imunes a interesses, a valores. Diversas correntes têm mostrado com sucesso que fatores relativos a sociedade, e aos grupos sociais que se utilizam da tecnologia em questão, podem ser preponderantes na aplicação de uma tecnologia, senão no seu projeto por completo. Tecnologias como a televisão e a Internet conseguiram, por um lado, se massificar globalmente de forma rápida e padronizada, e por outro, serem influenciadas também de forma global, refletindo a influência de determinados grupos sociais em outros completamente diversos, permitindo que culturas outrora isoladas pudessem oferecer sua contribuição a outras comunidades. Até recentemente, o enorme número de atividades humanas que acontecem em grupos pequenos não eram tecnicamente mediadas e por isso podiam apenas acontecer em configurações face a face. A Internet permite essa comunicação recíproca entre pequenos grupos. Esse é um avanço importante que tendemos a considerar como natural após trinta anos de comunicações online [Feenberg, 2012].

Assim, entendemos os conceitos de objetos digitais e repositórios como sendo tecnologias socialmente influenciadas, trazendo consigo – como a Internet – a história do seu desenvolvimento, e a influência dos grupos que mais os utilizam, e não só os grupos que os projetaram e desenvolveram em primeiro lugar. E é exatamente essa visão que pretendemos estender aos pesquisadores que utilizam os repositórios, permitindo que se faça um desenvolver e pensar de maneira coletiva e colaborativa.

3. Nossa Proposta de Fluxo de trabalho

A presente proposta tem como principal objetivo a preservação e manutenção da longevidade de objetos digitais, bem como tornar acessível a pesquisa e catalogação dos mesmos objetos. Realiza isso através da definição de um fluxo de trabalho que permita a um grupo de pesquisadores capturar imagens de documentos físicos ou o próprio arquivo digital, armazenar isso em formatos abertos, realizar reconhecimento ótico de caracteres, catalogar e referenciar o documento, e finalmente armazená-lo em um repositório aberto, de maneira a facilitar sua pesquisa e recuperação. Um segundo fluxo de trabalho trata da catalogação de comentários e anotações do documento, realizado por um grupo de usuários, e agregar essas informações ao repositório.

As tarefas rapidamente listadas aqui serão realizadas por softwares livres, utilizando formatos abertos e padronizados, e com hardware de baixo custo, permitindo que grupos de pesquisa com recursos limitados possam utilizar a estrutura sem arcar com o custo de soluções comerciais.

O resultado final deste fluxo é um repositório aberto com objetos digitais confiáveis e duráveis, em formato aberto, que possa ser acessado de maneira padronizada e aberta, criando um ambiente de colaboração onde pesquisadores de grupos e áreas diversas podem acrescentar conteúdo aos objetos digitais e criar relacionamentos entre eles, aumentando a massa crítica de conhecimento no repositório. A colaboração entre grupos é um instrumento chave para permitir isso, e a escolha dos softwares foi baseada nessa necessidade.

A integração dos softwares selecionados se dará por interfaces construídas pela equipe, em sistemas adequados a cada ambiente. Nas fases de captura e organização do documento, um aplicativo desktop será utilizado para agregar as funcionalidades dos softwares e bibliotecas usados, e após a disponibilização em repositórios, uma aplicação web, possivelmente usando recursos de nuvem, será utilizada para permitir que os grupos possam colaborar no aperfeiçoamento das meta informações dos objetos digitais e do reconhecimento ótico de caracteres.

3.1. Definição dos fluxos de trabalho

De maneira geral, os passos do fluxo de trabalho principal são os seguintes:

1. Digitalização do documento físico (através de scanners convencionais ou documentais) ou obtenção do documento em meio digital que possa ser manipulado, isto é, que permita reconhecimento de caracteres e conversão para formatos abertos (notadamente imagens, PDFs, documentos de pacotes de escritório).

2. Realização de reconhecimento ótico de caracteres, de forma a obter o máximo de informações do conteúdo do documento. O reconhecimento pode não ser exato, ainda mais quando os documentos físicos forem muito antigos (fontes não padronizadas) e manuscritos, mas qualquer massa de dados buscáveis pode significar uma redução significativa nas pesquisas realizadas posteriormente.
3. Se considerado necessário (se o documento possuir relevância suficiente), será realizada uma etapa de correção do reconhecimento de caracteres, usando uma infra-estrutura de software a ser construída (pela agregação de bibliotecas existentes), onde um grupo de usuários será responsável pela verificação manual do texto reconhecido e realizará as correções necessárias, permitindo que o software de reconhecimento melhore seu acerto, ou corrigindo diretamente o texto encontrado.
4. Criação do objeto digital confiável, usando os padrões da NOBRADE [Conarc, 2006], em formato DjVu [DjVu.org, 2012] ou PDF, que são abertos, multiplataforma, e permitem a implantação de critérios arquivísticos no próprio objeto digital. Nessa etapa também são catalogadas as informações de referência do documento, segundo a NOBRADE, informações essas que são também inseridas dentro do objeto digital.
5. Disponibilização do objeto digital em um repositório aberto, baseado em software livre, que permite a pesquisa e o acesso aos objetos através de interface web, bem como interfaces programáticas, como webdav e web-services.

O fluxo de trabalho de catalogação de anotações, comentários e destaque do documento é mostrado a seguir. Este segundo fluxo de trabalho depende do documento já estar disponibilizado no repositório.

1. Acesso ao documento, no formato de objeto digital confiável, no repositório. Relacionar o documento através de software de organização de acervo, em nuvem, apontando para o repositório.
2. Destacar porções do documento, criar anotações e citações, armazenando isso em software de organização de acervo, e software de leitura de documentos, que permita marcações e exportação dessas marcações.
3. Extrair as marcações e agregar estas informações no repositório.

Este segundo fluxo é realizado em grupos de interesse por assunto e/ou documentos, agregando esses comentários e anotações, e permitindo uma maior sinergia entre as equipes que vão tanto realizar quanto usar essas informações como fichamentos e citações.

3.2. Seleção dos softwares para o fluxo de trabalho

No processo de seleção de softwares para compor o fluxo de trabalho, foi seguida a lógica de se buscar softwares livres já disponíveis para a execução das tarefas, com a integração e sequencia do seu uso. Neste sentido, o fluxo de trabalho humano associado

a preservação e disponibilização documental está fortemente imbricado na escolha e no uso de uma cadeia de ferramentas que mediam estas atividades. Nos pontos onde não existirem softwares disponíveis ou que sejam necessárias integrações mais sofisticadas, para evitar aumentar a complexidade do uso do fluxo, então serão desenvolvidos módulos de software que serão disponibilizados como software livre para a comunidade. Seguindo ainda a lógica de uso de softwares disponíveis, esses novos módulos desenvolvidos, quando possível, serão criados como plugins ou melhorias a softwares, e a contribuição se dará na linha principal de desenvolvimento de cada um.

No processo de digitalização, serão utilizados scanners convencionais ou documentais, incluindo um modelo desenvolvido por um pesquisador da equipe, mostrado na Figura 1, que se propõe a realizar a captura de imagens de documentos em formato de livros, revistas ou jornais, que ou sejam muito grandes para scanners documentais, ou estejam em um estado tal de fragilidade que sua manipulação deva ser mais cuidadosa. Este equipamento é relativamente barato, se aproveitando do rápido desenvolvimento dos processadores incorporados em máquinas fotográficas digitais de baixo custo. Permite que a captura de documentos encadernados e frágeis seja feita em alta velocidade e sem danificá-los.

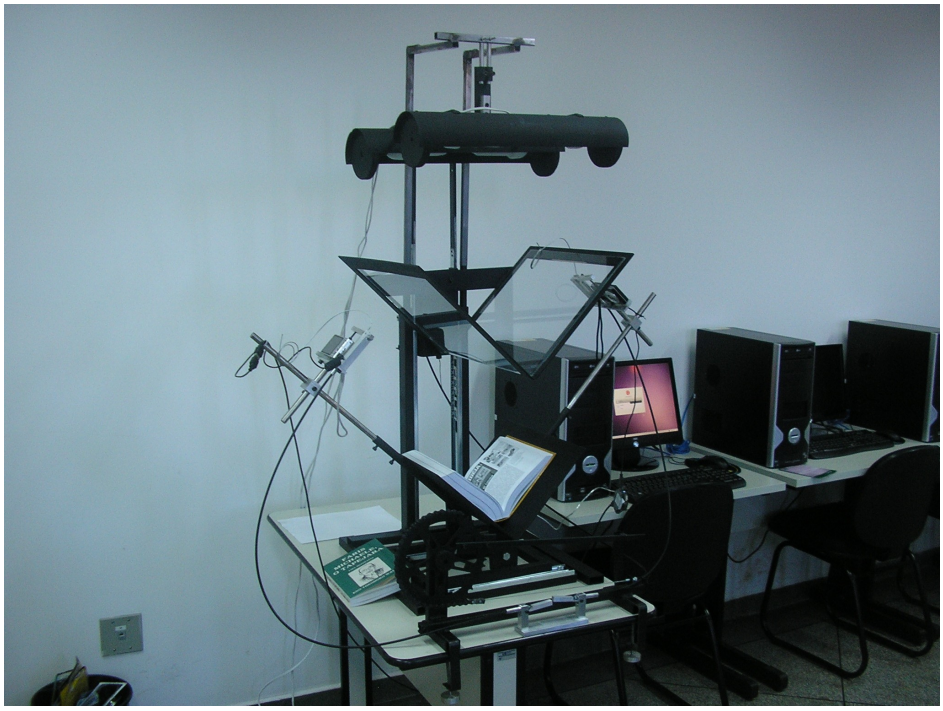

Figura 1: scanner documental desenvolvido pela equipe de pesquisa.

O reconhecimento ótico de caracteres será executado pelo software Ocrupus [Ocrupus, 2012] ou o Tesseract [Tesseract-OCR 2012], que também é utilizado pela Google em seu site Books, e a saída do processo de reconhecimento será reunida com as imagens em arquivos em formato DjVu, permitindo que o arquivo resultante seja buscável por qualquer biblioteca disponível, como a Apache Lucene [Apache Lucene, 2012], entre outras. Esse arquivo terá suas meta-informações agregadas e será gerado como um objeto digital confiável, nas normas da NOBRADE.

O repositório escolhido para o depósito dos objetos digitais é o Omeka, desenvolvido pelo Center for History and New Media da George Mason University

[Omeka, 2012], que conta com recursos para criação de coleções e narrativas, usando padrões como o Dublin Core (mas flexível o suficiente para a implantação da Norma Brasileira de Descrição Arquivística - NOBRADE), e permitindo a adição de novos padrões através de plugins. O software Ica-Atom [Ica-Atom, 2012] também foi considerado como repositório, por seguir normas internacionais e ser um padrão de fato na área, além de ser extensível com plugins. A escolha recaiu sobre o Omeka por possuir uma fácil integração com outras ferramentas e plataformas.

Do mesmo grupo da George Mason University é o software escolhido para organizar as coleções e citações no segundo fluxo, o Zotero [Zotero, 2012], que pode funcionar como um plugin para Mozilla Firefox ou ser executado como aplicação em sistemas operacionais Linux, Windows e Mac. O Zotero permite que grupos compartilhem acesso a coleções, e já possui um plugin para integração com o Omeka, facilitando a integração com o repositório.

O protocolo de funcionamento do Zotero é também aberto, o que deu origem a aplicativos de terceiros que acessam os repositórios, incluindo aplicativos para tablets e smartphones, permitindo que o pesquisador tenha uma ferramenta portável e prática para realizar a organização das suas coleções e a anotação e marcação em documentos.

3.3. Limitações das ferramentas e trabalhos propostos

Pela natureza colaborativa do fluxo de trabalho proposto, nem sempre encontram-se ferramentas disponíveis com todas as características necessárias para essa integração. Parte deste trabalho é justamente transpor essas barreiras com a agregação de plugins ou outros softwares, ou com o desenvolvimento de módulos e plugins que possam suprir as necessidades.

Algumas necessidades já foram identificadas, como a tradução do software Omeka, e a implementação das normas de arquivística brasileiras no mesmo software. As duas necessidades possuem solução plausível (a criação de um “tema” internacionalizado e a definição das normas em XML), e fazem parte do cronograma do grupo.

A integração dos softwares de reconhecimento ótico de caracteres, a geração do objeto digital e a correção dos possíveis erros de reconhecimento serão executadas por um módulo desenvolvido pelo grupo, de forma a trazer maior flexibilidade e facilidade de integração.

Na organização dos destaques e anotações, um módulo será desenvolvido para extrair dos arquivos digitais as marcações e anotações e inserir no arquivo residente no repositório.

4. Conclusões e continuidade do trabalho

O uso de repositórios abertos para manter objetos digitais utilizados em pesquisas acadêmicas por si só já é um avanço em relação a situação atual do campo, e também abre fronteiras para comunidades externas usarem os mesmos documentos. Com um fluxo de trabalho para apoiar a construção colaborativa desse repositório, e a agregação de informações ao objeto digital, podemos proporcionar a comunidade uma ferramenta

eficiente e de fácil acesso, que permitirá que custos sejam reduzidos, e pesquisas sejam realizadas em menor tempo, e com maior abrangência.

A já citada nova divisão social do trabalho do pesquisador, que em grupo compartilham sua visão e anotações dos documentos, pode ser um avanço que possa ampliar as possibilidades de cruzamentos no processo criativo de interpretar os documentos.

O presente fluxo de trabalho ainda se configura como uma proposta, mas se pretende colocá-lo em funcionamento em duas universidades, de maneira a analisar a efetividade do processo, e o reflexo disso no dia a dia do pesquisador e da qualidade do material produzido.

Referências

Apache Lucene (2012), Apache Lucene Core – search engine library, http://lucene.apache.org, visitado em setembro de 2012

Conarq (2006), NOBRADE – Norma brasileira de descrição arquivística, Conselho Nacional de Arquivos, Arquivo Nacional.

DjVu.org (2012), http://www.djvu.org, visitado em setembro de 2012.

Feenberg, Andrew (2002), Transforming Technology: A Critical Theory Revisited, Oxford University Press.

Feenberg, Andrew (2011), (Re)Inventing the Internet: Critical Case Studies, Sense Publishers.

ICA-AtoM (2012), ICA-AtoM web-based archival description software, http://www.ica-atom.org, visitado em setembro de 2012

Ocrupus (2012), Google Code Ocrupus – open source document analysis and OCR system, http://code.google.com/p/ocrupus, visitado em setembro de 2012

Omeka (2012), Omeka web-publishing platform, http://www.omeka.org, visitado em setembro de 2012

Unesco/Commonwealth of Learning, (2012). Taking OER beyond the OER Community. http://oerworkshop.weebly.com, visitado em setembro de 2012.

Tesseract-ocr (2012), Google Code Tesseract OCR engine, http://code.google.com/p/tesseract-ocr, visitado em setembro de 2012

Williams, Raymond (2003), Television, Routledge.

Zotero (2012), Zotero research tool, http://www.zotero.org, visitado em setembro de 2012