PROJETO 914 BRA 2015 – IBICT

EDITAL Nº 027/2014

PRODUTO Nº 01 : DOCUMENTO TÉCNICO N°1

Levantamento dos *softwares* livres disponíveis para análise de informação estruturada

Luc QUONIAM

NOME DO (A) CONSULTOR (A)

São Paulo / SP / 08 / 2014

OBS.: **Entregar à Coordenação solicitante 02 (duas) vias impressas (não precisa encadernar), assinar na capa e rubricar as demais páginas e 01 (uma) cópia em CD no formato pdf.ocr.**

Sumario

Introdução

Cumprindo com as exigências do contrato a apresentar softwares de acesso livre, é essencial descrever um número maior de softwares, pois raramente são encontrados softwares livres que satisfaçam várias soluções. Então no mundo do "livre" será necessário procurar um conjunto de soluções, cada uma adequada para uma determinada função.

Serão apresentadas várias soluções, das quais caberá ao IBICT restringir e eleger as mais adequados às suas necessidades de acordo com a lista apresentada. Do(s) software(s) recomendado(s) serão descritos: a tecnologia utilizada, as facilidades ofertadas, as formas de integração com portais *web*, a função a ser coberta(s) por ele(s), e o porquê de serem considerados os mais adequados ao IBICT. **Para cada software(s) serão incluídos *links* de acesso como complemento de informação que devem ser considerados como proposta de capacitação da equipe (parte do manual do curso, a ser complementado quando o IBICT selecionar os softwares mais relevantes.**

Embora mencionado, conforme solicitação, não é coerente incluir as respectivas instituições mantenedoras dos softwares, pois soluções de livre acesso e de código aberto são geralmente mantidas por comunidades, das quais o IBICT pode participar, e se tornar ele mesmo mantenedor.

Cabe também entender que software livre nem sempre será um software sem custo. É conveniente considerá-lo como uma forma diferente de custo, e não uma simples supressão total de custo. Nem sempre as soluções apresentadas vão se adequar 100% às necessidades institucionais, e deverão ser customizadas e complementadas. Também, nem sempre será fácil implantá-los. Neste caso, será necessário entrar em contato com o autor (do software) que, naturalmente, estipulará o seu custo. Geralmente é questionado se "a utilização do software livre é mais vantajosa, mesmo que não seja isenta de custo?". Sim, o custo sempre será menor (terá sempre uma economia), e terá o mínimo da garantia de licença perpétua (por ter o código e poder adaptá-lo para funcionar sempre), o que o(s) software(s) não livre(s) muito raramente oferecem. Terá também a garantia de uma customização perfeita para se adequar aos dados do IBICT. Uma vez implantadas as soluções, o instituto também participará de ações de "responsabilidade social" liberando soluções à disposição de cada cidadão sem limitação de custo.

Objetivos

Portanto, os objetivos desta consultoria são:

1. identificar softwares de acesso livre e código aberto com potencial para serem utilizados na análise de bases estruturadas do IBICT;

2. recomendar os softwares apresentando a sua descrição, a tecnologia utilizada, as facilidades ofertadas e as formas de integração com portais *web* e outros;
3. entrar num consenso com o IBICT, no sentido de escolher os softwares mais adequados às suas finalidades, bem como a(s) base(s) que servirão de protótipo para aplicação.
4. Elaborar a proposta de capacitação da equipe que utilizará os softwares escolhidos, usando a(s) base(s) selecionadas.

Plataforma de tratamento

O tratamento das infomações bibliográficas passa por diferentes fases de trabalho em "*back office*" (http://pt.wikipedia.org/wiki/Back_office) e em "f*ront office*" (http://pt.wikipedia.org/wiki/Front_office) . É necessário que a plataforma possua uma boa interoperabilidade (http://pt.wikipedia.org/wiki/Interoperabilidade) entre estas duas facetas do trabalho de mineração dos dados. No "f*ront office*", caso o sistema operacional seja *Windows*, é preferível ter uma máquina capaz de instalar um *Windows* de 64 *bits* por razão de suporte de memória maior que 4Gb. É recomendável também uma placa gráfica de alto desempenho (marca ATI ou NVidia "gama quadro") ou placa de video externa ligada a uma porta USB3.0 (http://www.startech.com/AV/USB-Video-Adapters/USB-3-to-Displayport-Video-Card-Multi-Monitor-Adapter-2560x1600~USB32DPPRO). Para se obter um *display* de alta definição (2500x1600= 4Mo de *pixels*), que seja capaz de restituir os dados de modo a poder interpretá-los, existe um dispositivo de restituição de alto desempenho sendo desenvolvido com base em tecnologias "*low cos*t" (na faixa de 6000 € contra 70 000€ no comércio), disponível na incubadora de empresas "*Marseille Innovation"* (http://www.marseille-innov.org/), situada na França. Mesmo que esta solução não seja o objeto deste relatório técnico, é mencionada para possíveis evoluções.

Figura 1: Tela de alto desempenho para restituiçao de resultados de análise.

Seria desejável também a disponibilização de um centro de recursos que permita o compartilhamento dos recursos já desenvolvidos, a fim de evitar que novos usuários (funcionários) que queiram fazer uso dos mesmos não tenham dificuldade de buscar soluções à partir do início. É recomendável, portanto, deixar os recursos em acesso livre, e não apenas depositados no *Drive* (sistema utilizado anteriormente), para facilitar o auxílio aos usuários que iniciam a utilização dos recursos.

Conservação do sistema operacional *Windows*

Hoje, qualquer exposição de um computador na *web* provoca o risco de contaminação por parte de inúmeros programas indesejáveis para fim de rastreamento do usuário. Para manter um bom desempenho da maquina, é necessário que o computador sempre esteja "limpo". Atualmente existem algumas ferramentas "indispensáveis" a serem utilizadas, uma vez a cada semana ou a cada duas semanas. Algumas delas são listadas abaixo. São soluções "*freeware*" ou na base de *"freemium*" (http://en.wikipedia.org/wiki/Freemium).

AdwCleaner

http://www.bleepingcomputer.com/download/adwcleaner/

AdwCleaner é um programa que procura e elimina *Adware* (http://en.wikipedia.org/wiki/Adware), barras de ferramentas, programas potencialmente indesejáveis (PUP), e sequestradores de navegadores do seu computador. Usando

AdwCleaner você pode facilmente remover muitos destes programas para uma melhor experiência do usuário em seu computador, e enquanto navega na *web*. Trata-se de um verdadeiro "desisntalador de pragas".

Geek Uninstaller

http://www.geekuninstaller.com/

Alguns destes programas "indesejáveis" precisarão ser desinstalados um por um. O programa *Geek Uninstalle*r ajudará a fazer isso. O programa de remoção padrão do *Windows* deixa muitas sobras em seu PC. *GeekUninstalle*r realiza uma varredura profunda e rápida e remove todas as sobras por meio da utilização da "força" na remoção de programas persistentes e danificados.

Ccleaner

https://www.piriform.com/CCLEANER

CCleaner é o nosso sistema de otimização, e sua particularidade é que trata-se de uma ferramenta de limpeza. Ele remove arquivos não utilizados do seu sistema - permitindo que o *Windows* possa trabalhar mais rápido e liberando espaço no disco rígido. Ele também limpa traços de suas atividades *online*, tais como sua história no uso da Internet. Além disso, ele contém um registro limpo com todos os recursos. Mas o melhor ´é que é rápido (normalmente leva menos de um segundo para ser executado) e não contém nenhum *spywar*e ou *adware*! **Também limpa a base de registros do *Windows* .**

Spybot

http://www.safer-networking.org/

Tem a mesma função do A*dwcleaner*, talvez com um desempenho menor, mas com um módulo instalado no "arranque" do computador que o "protege" em tempo real.

Atualização do(s) software(s)

A atualização do sistema operacional *Windows* é feita de modo automático. Seria conveniente encontrar também uma solução que vasculhe automaticamente a internet para atualizar o conjunto de softwares apresentados neste relato, de forma que seja possível, sempre, trabalhar com as versões mais atualizadas. Devido ao número de aplicativos utilizados, é pouco provável que a solução manual seja eficiente.

Limpeza, preparação dos dados

Editoração/manipulação de arquivos texto

Quando se fala de tratamento da informação para análise, tomada de decisão e de valor agregado, o mais importante, e a etapa mais demorada, é a da estruturação/limpeza que deve ser efetuada antes da análise propriamente dita. O *ratio* de tempo entre preparação e análise é da ordem de 80/20. Então, esta parte, embora possa parecer trabalhosa e pouco atrativa, é fundamental, por ter grande repercussão na qualidade dos resultados (tanto em

termos de significado como de custo de produção). É sempre muito importante ter esta visão para dosar o difícil equilíbro entre a informação perfeita de custo absurdo e a análise sem sentido, por estar baseada em dados sem sentido. Da mesma forma, esta fase se enquadra nas vertentes da "ciência da informação" e da "ciência da computação", sem ser a parte "nobre" de nenhuma delas. o fato é que esta fase necessita de uma boa integração entre cientistas da informação e da computação para que haja sinergia e um bom resultado. O bom entendimento entre as partes neste ambiente de trabalho nesta fase é indispensável para conseguir levá-la a bom termo.

A grande maioria dos formatos de arquivos manipulados no âmbito da análise de informação bibliográfica são arquivos estruturados do tipo "texto". Neste sentido, podem ser "manipulados" com um bom "editor de texto". Existem várias soluções possíveis desde que sejam contemplados os seguintes requisitos:

- Conversão dos fim de linha *UNIX/Windows*
- Conversão Maiúscula/Minúscula (inteira ou primeira letra)
- Procura/substitui avançado (fim de linha, *tab*)
- Procura/substitui na forma de "*Regex*"
- Capacidade de edição de "grande arquivos"
- Conversão de *codage* de caracteres
- Possibilidade de "ordenar"
- Visualização dos caracteres especiais (*tab*, fim de linha)
- Função anula, repita
- Manipulação simultânea de vários arquivos (procura/substitui igualmente)
- Capaz de fazer números "procura substitui" ao mesmo tempo (tratamento dos caracteres diacriticos http://pt.wikipedia.org/wiki/Diacr%C3%ADtico, tirar palavras vazias, opções necessárias para vários tratamentos de análise da informação)
- Trabalhar em modo coluna
- Sistema de macro comandos
- Coloração do texto conforme *template f*ornecido
- Uso de conversão XML (XSL)
- Reconhecimento de XPATH

É possível levantar várias soluções para atender a esses requisitos. A maioria dos editores que atendem tais características são voltados à programação. Portanto, o tratamento da informação necessita de um trabalho de aprendizado da parte do usuário com formação em "ciência da informação". Um editor "razoável" seria:

Notepad ++

http://notepad-plus-plus.org/

Instituição mantenedora

Em código livre e aberto sob GPL licença, notepad++ está na plataforma *sourceforge.net*.

Inconvenientes

Faltam algumas caraterísticas para o Notepad++ ser a solução perfeita:

- o sistema de Macro não e "amigável"
- não há procura/substitui múltiplos, mas esta opção poderia facilmente ser implementada com *plugin,* deixando criar *script python* (http://stackoverflow.com/questions/11389466/multiple-word-search-and-replace-in-notepad) desde que sejam acrescentadas as várias opções de procura/substitui do próprio notepad++

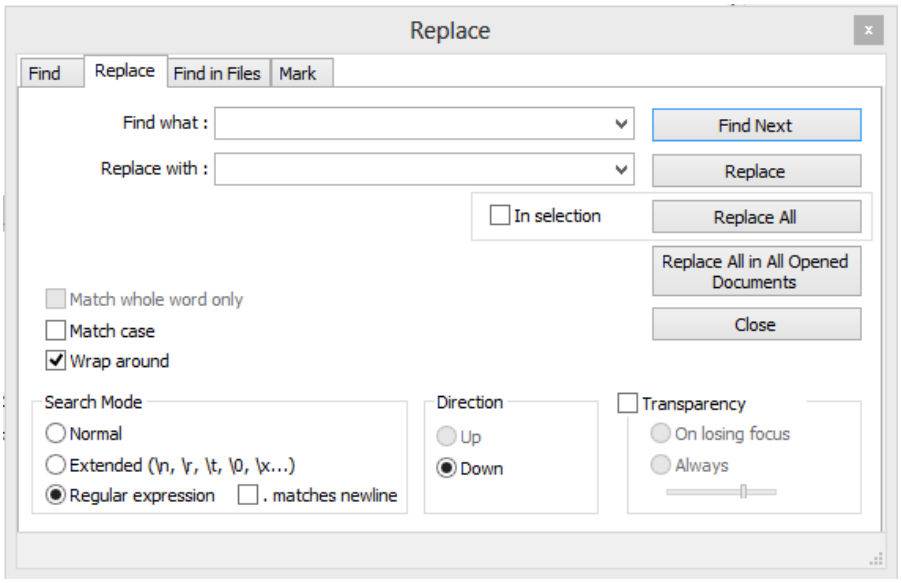

Figura 2: Opções necessárias de procura/substitui.

- Não há versão em português, mas é possível traduzir ou utilizar a versão em Inglês.
- Recomenda-se uma biblioteca de transformação XSLT de padrões, bem como uma pequena documentação para poder automatizar algumas transformações "padrão".

Wreplace

http://www.sharktime.com/us_wReplace.html

Instituição mantenedora

SharkTime Software é uma empresa de desenvolvimento de softwares independente, que se especializou na criação de utilitários de software simples e úteis para PCs com *Windows*. Esta empresa libera de graça um pequeno editor de texto que somente tem como valiosa a possibilidade de multi procura/substitui.

Comentários

Caso esta função seja implementada no Notepad ++ wreplace (não é código aberto), a mesma seria inútil, não fosse a sua utilidade para as listas de procura/substitui que oferece (disponibilizadas em anexo).

Outras listas de procura/sustitui

No caso de tratamento da parte de referências "*full text",* bem como títulos, resumos, descrições ou até o próprio *"full text*", existem vários tratamentos possíveis que serão detalhados mais a frente. O tratamento deste tipo de informação pode ser muito lento e trabalhoso, pois pode ir até a uma análise linguística fina do texto. Serão abordadas somente metodologias mais "*quick and dirty*", que podem parecer simplistas, mas são capazes de fornecer "boas análises", e o melhor, com um bom *ratio* entre custo/qualidade. Uma parte destes tratamentos solicita a remoção de "palavras vazias" ou *"stopwords* (http://en.wikipedia.org/wiki/Stop_words). Estas listas de *"stopwords*" existem em vários idiomas (29 languages stopwords lists: https://code.google.com/p/stop-words/), (disponibilizadas no anexo). Para fins de tratamento de informação multilíngue, existe a opção (para resultados rápidos) de "juntar" vários idiomas em um só arquivo. No caso de tratamento de produção científica brasileira, com os idiomas Português, Inglês, Espanhol, Francês, etc. Seria valioso para tratamentos "mais sofisticados" de texto livre, juntar vários "tipos de listas". Para poder "enriquecer" o tratamento, fornecemos aqui alguns exemplos de listas em Inglês. Porém, recomenda-se que se encontre listas também de outros idiomas mencionados. É conveniente procurar o que "já existe" (legitimidade, referências, otimização do tempo...)

- Listas de "*idioms*" por línguas a fim de tratar estas expressões como "uma palavra só", substituindo o espaço entre palavras por um "-", por exemplo. (http://en.wikipedia.org/wiki/List_of_English-language_idioms, http://www.learn-english-today.com/idioms/idioms_alphalistsA-Z.htm, http://www.eslcafe.com/idioms/id-list.html).
- Lista de termos por significado, exemplo dos verbos de ação (http://career.opcd.wfu.edu/files/2011/05/Action-Verbs-for-Resumes.pdf), afim de realizar tratamento para detectar ações que já foram providenciadas.
- Lista de termos que possibilitariam encontrar os "princípios inovadores", "resolução de problemas", baseados na metodologia TRIZ (http://wbam2244.dns-systems.net//EDB_Welcome.php). Esta parte não será detalhada aqui, mas pode ser promissora.
- Lista de termos encontrados na *Wikipedia*. Detalharemos esta parte posteriormente.

Com uma coleção de listas assim (a serem elaboradas), seria possível implementar facilmente algumas soluções de análise da forma *"keywords in context*" (http://en.wikipedia.org/wiki/Key_Word_in_Context). Estas soluções não serão abordadas detalhadamente aqui.

Outras técnicas a serem conhecidas

Regex

http://pt.wikipedia.org/wiki/Express%C3%A3o_regular

Para pensar em tornar a função procura/substitui interessante e poderosa, além das listas já mencionadas acima, será necessário desenvolver algumas listas "padrão" de procura/ substitui usando *Regex* (padronização de nomes de autores). Detalharemos um recurso possível utilizando a base Lattes que, infelizmente, se aplica somente a autores brasileiros. A ação procura/substitui em *Regex* é indispensável quando se usa o *Notepad* ++ (é necessário instalar *plugins*, disponibilizados em seus manuais em anexo), e no *Gephi* (análise de redes). Uma dificuldade do *Regex* é a variedade de sintaxes (em p*ython, Notepad* ++, Java). Dessa forma, se recomenda uma coleção documentada de "soluções padrão".

Xpath

http://www.w3schools.com/XPath/

XPath é usado para navegar através de elementos e atributos em um documento XML. *XPath* é um elemento importante no padrão XSLT do W3C - e XQuery e XPointer são ambos construídos em expressões *XPath*. É indispensável conhecer *XPath* tanto para poder "manipular" qualquer arquivo organizado hierarquicamente, como para fazer extração de dados de *website*s, mas também junto com a linguagem XSL fazer conversão de formatos de arquivos. Notepad ++ (tem *plugins* que permitem trabalhar com XPath e XSL). A colaboração entre profissionais da computação e da Ciência da informação será de suma importância. Recomenda-se o desenvolvimento de uma "biblioteca" de soluções prontas.

Xpath como *crawler* de dados

Quando uma página *web* não é dividida em várias páginas (todos os documentos encontram-se em uma única página), é possível usar *XPath* para extração de dados (http://pt.wikipedia.org/wiki/Web_crawler)

Por exemplo

Na página seguinte, obtida de um tratamento do ScriptLattes (http://vlab4u.info/uninove/administracao/PR-administracao/PB0-0.html), foi possível extrair com *Xpath*, como texto, para posterior tratamento:

Exemplo 1	Extração dos títulos de trabalho
XPath	html/body/table[*]/tbody/tr[*]/td/b/text ()
Resultado	Conflitos de transparência e confidencialidade na certificação de sistemas de gestão ambiental *Bioaccumulation of potentially toxic trace elements in benthic organisms of Admiralty Bay (King George Island, Antarctica)*

	Revisitando as Tipologias da Teoria de Redes Interorganizacionais Conflitos sócioambientais em Unidades de Conservação em áreas urbanas: O caso do Parque Tizo em São Paulo. As Mídias Sociais sob a Perspectiva do Marketing Esportivo: O Caso São Paulo Futebol Clube Equipes virtuais de projetos, mobilidade do trabalho e o uso de tecnologias de informação móveis: um estudo teórico empírico *The consumer behaviour in different retail formats in Brazil* *The Relationship between Environmental Concern and Declared Retail Purchase of Green Products* ....

Exemplo 2	Extração dos Autores de trabalho
XPath	html/body/table[*]/tbody/tr[*]/td[2]/text()[1]
Resultado (notar neste caso a necessidade tratamento)	AGUIAR, A. O. E. ; CÔRTES, Pedro Luiz. Alessandra Pereira Majer ; PETTI, M. A. ; CORBISIER, T. N. ; RIBEIRO, A. P. ; THEOPHILO, C. S. ; FERREIRA, P. A. L. ; FIGUEIRA, R. C. L.. ALSSABAK, Nawfal Assa Mossa ; SOUZA, Leandro Januário de ; CARNEIRO DA CUNHA, Julio Araujo ; MACAU, Flávio Romero. ARCE, P. ; PENDLOSKI, C. J. S. ; OLIVEIRA, R. B. ; GALLARDO, A. L. C. F. ; RUIZ, M. S.. ASSIS, E. E. ; TOLEDO, L. A. ; PISCOPO, M. R. ; ROSA, C. M.. BELFORT, A. C. ; MARTENS, C. D. P.. BRAGA JR., S.S. ; LOPES, E. L. ; SATOLO, E. G. ; SILVA, Dirceu ; MORETTI, S. L. A.. BRAGA JUNIOR, S. S. ; SATOLO, E. G. ; GABRIEL, M. L. D. S. ; SILVA, D.. ....

Exemplo 3	Extração dos títulos de trabalho por ano
XPath	//html/body/h3/text() |//html/body/table[*]/tbody/tr[*]/td/b/text ()
Resultado	2014 Conflitos de transparência e confidencialidade na certificação de sistemas de gestão ambiental *Bioaccumulation of potentially toxic trace elements in benthic organisms of Admiralty Bay (King George Island, Antarctica)* Revisitando as Tipologias da Teoria de Redes Interorganizacionais Conflitos socioambientais em Unidades de Conservação em áreas urbanas: O caso do Parque Tizo em São Paulo. As Mídias Sociais sob a Perspectiva do Marketing Esportivo: O Caso São Paulo Futebol Clube ....

É bom notar, então, a importância de se pensar a apresentação dos resultados de uma busca de forma que propiciem este tipo de extração.

XPath applicativos

Extensões para os navegadores *web* a fim de fazer *crawling*.

Mozzilla	Firebug	https://www.getfirebug.com/
Mozzilla	FirePath	https://code.google.com/p/firepath/
Mozzilla	FireFinder	http://robertnyman.com/firefinder/
Mozzilla	XPath finder	https://addons.mozilla.org/fr/firefox/addon/xpath-finder/?src=api
Mozzilla	Xpath checker	https://code.google.com/p/xpathchecker/
Chrome	Scraper	https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn
Chrome	data miner	https://chrome.google.com/webstore/detail/dataminer/nndknepjnldbdbepjfgmncbggmopgden

Existem também *crawlers* baseados na linguagem de busca *Regex.*

XSL/XSLT

http://www.w3schools.com/xsl/default.asp

XSL significa *Extensible Stylesheet Language,* e é uma linguagem de folha de estilo para documentos XML. XSLT significa “Transformações XSL”. Arquivos de entrada de software de tratamento/apresentação de informação, quando em XML podem ser convertidos de um formato para outro usando a tecnologia XSL/XSLT

aplicativo	formato entrada	exemplo	comentários
freeplane / mindmap	mm files : http://freeplane.sourceforge.net/wiki/index.php/Current_Freeplane_File_Format	http://freeplane.sourceforge.net/mapsOnline/?map=freeplaneApplications.mm	Muito bom para dados hierárquicos tipo UF / cidade / IES
gephi	gexf files : http://gexf.net/format/	http://vlab4u.info/doencas%20negligenciadas/Dengue/denguecore1/Resultados/Rede/gexf/index.html	Muito bom para dados em rede. o formato não é bem documentado. É necessário ser insistente.

Providência: elaborar alguns XSLT "padrão" para oferecer conversões de um formato para outro, sendo este um passo estratégico.

Tratamento, Análise dos dados

Uma vez que os dados forem limpos e preparados, vem por fim a fase gratificante da análise. Abordaremos várias dimensões de análise: os campos univariados, os campos multivariados, os campos de texto integral e os campos que permitem estabelecer relações.

Tratamento de campos univariados

Entende-se como campos univariados os campos de referência bibliográfica, que somente terão um constituinte em cada nota bibliográfica (por exemplo: data de publicação, nome da revista, UF da IES da defesa, IES da defesa, etc.). O tratamento destes campos se limita à contagem dos elementos constituintes e tabela cruzada de co-aparição destes elementos.
Muito embora softwares como *Excel* ou *libre office* tenham possibilidades sofisticadas para fazer tabelas cruzadas, uma solução fácil de manipular e de compreender para o usuário final e diretamente integrável num *website* é preferível.

PivotTable.js

Instituição mantenedora

Desenvolvido em *java scrip*t por Nicolas Kruchten (http://nicolas.kruchten.com / Staff Software Engineer at Datacratic em Montréal, Québec, Canada, http://datacratic.com/site/), *PivotTable.js* está disponível na plataforma *Github* (https://github.com/nicolaskruchten/pivottable), completamente livre de qualquer direito. Pode ser colocado numa página *web*, bem como ser usado localmente através de um *browser*.

Formatos de entrada

Este aplicativo aceita entradas nos formatos em J*son* (http://json.org/json-pt.html), *csv* (http://pt.wikipedia.org/wiki/Comma-separated_values ou http://en.wikipedia.org/wiki/Comma-separated_values), bem como *array* de php e tabela html (documentadas em https://github.com/nicolaskruchten/pivottable/wiki/Input-Formats).

Melhorias possíveis

Os menu estão em inglês. Seria valioso contar com opções de um arquivo externo, deixando a possibilidade de oferecer outros idiomas (como português). Há uma versão simplificada com possibilidade de *input* de um arquivo *csv* para uso local no *browse*r com arquivo exemplo. A versão *online* deste exemplo está disponível em http://nicolas.kruchten.com/pivottable/examples/local.html.

"O *PivotTable* só possui a versão em inglês. Assim, seria vantajoso ter a possibilidade de internacionalização, com a implementação de arquivo de tradução. Com isso, se poderia obter a opção em várias línguas."

Vantagem

Além de de ser facilmente compreensível, esta solução é independente da plataforma (*Linux, Windows* e *Mac*), e funciona *online* e *offline*. É facilmente alterável, e os resultados são facilmente recuperáveis para redação de documentos finais.

Tratamento de campos multivariados

Os campos como "membros da banca", "palavras-chave" são selecionados por cada referência e vários valores. Embora seja interessante ter contagens do gênero da técnica vista anteriormente *(pivot table*), é difícil fazer tabelas para aplicar a tecnologia de *pivot table*. Seria necessário construir uma tabela em "2 níveis" (tabela de contingência). Seria valioso conseguir uma solução simples e elegante (tabela cruzada tipo *pivot table* acima) para poder representar relações tais como autores-datas ou autores-autores. Aqui vamos focar nas soluções usando software de análise de redes, que podem ser citadas como:

Pajek

http://pajek.imfm.si/doku.php

Instituição mantenedora

Software Esloveno desenvolvido pelo Prof. Dr. Vladimir Batagelj, assumidade mundial em análise de redes (https://www.fmf.uni-lj.si/si/imenik/2909/).

Comentário

Excelente e bem documentado, gratuito e robusto, porém de difícil utilização. Há uma dificuldade em colocar os resultados dinâmicos "*online*". Funciona somente no ambiente *Windows*. Não é recomendável para o IBICT.

Netdraw

http://www.analytictech.com/

Instituição mantenedora

Sofware inglês desenvolvido pelo Prof. Dr.Steve Borgatti, assumidade mundial em análise de redes (http://www.steveborgatti.com/).

Comentário

Excelente e bem documentado, falsamente gratuito (pois, para ter acesso a todas as opções, é necessário comprar o software UCINET). Difícil de colocar os resultados dinâmicos "*online*". Funciona somente no ambiente *Windows*. Não é recomendável para o IBICT.

R statistical package

http://www.r-project.org/

Talvez o melhor, mais amplo, robusto e completo *package* estatístico disponível atualmente. Ampla comunidade ao redor. Funciona em todas as plataformas, Completamente gratuito e de código aberto. Todavia, necessita ser desenvolvido em linguagem "R" para poder ser utilizado de modo eficaz. O que seria pouco provável no mundo da ciência da Informação. Os resultados não são facilmente instaláveis de forma dinâmica num *website*. Não é recomendável para o IBICT por requerer programação avançada, a menos que possam ser encontradas soluções já programadas (veja IRAMUTEQ mais à frente).

Gephi

https://gephi.github.io/

Instituição mantenedora

Desenvolvido em Java na França no INRIA - ciência da computação), sob a coordenação de Mathieu Bastian (engenheiro). Funciona independente da plataforma, desde que tenha possibilidade da instalação do Java. E recomendável ter um bom recurso de visualização, bem como uma plataforma de 64 *bits* para sua utilização de forma relevante. A maior parte do código *Gephi Plataform* está disponível sob a licença dupla CDDL 1.0 e GNU General Public License (GPL) v3.

É necessário entrar em um acordo com um mínimo de vocabulário de descrição de uma rede.

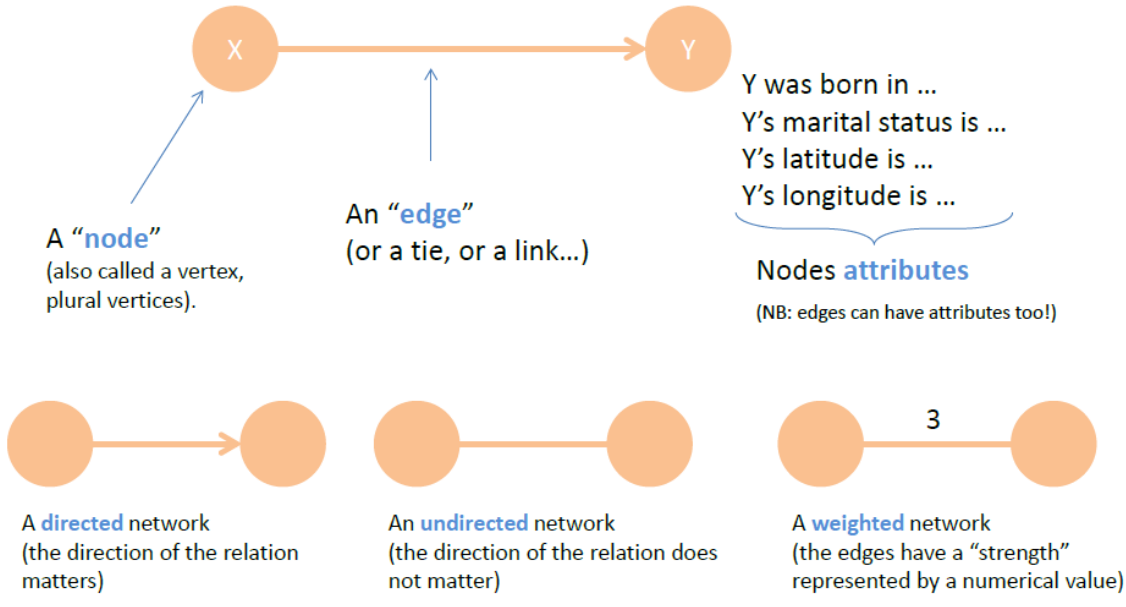

Figura 3: Terminologia de uma rede

Contras do Gephi

Embora Gephi seja o aplicativo de representações das relações contidas em um conjunto de referências bibliográficas, o Gephi apresenta alguns inconvenientes:

- A documentação não é bem feita. Não há documentação "centralizada", sendo encontrada em formato PDF e em vídeos no Youtube criados por usuários. Os arquivos em "pdf" com a documentação mínima se encontram em anexo. Esta limitação torna difícil o início de sua aprendizagem.
- Gephi possui *plugins*, alguns imprescendiveis, mas feitos por terceiros. Alguns deles fazem o programa travar (instabilidade), especialmente quando são solicitados cálculos estatísticos sobre a rede.
- As associações contidas em um conjunto de bibliografias representa rapidamente uma rede de grande tamanho, gerando arquivos consideravelmente pesados (vários megas). É recomendável que o computador seja equipado com uma placa gráfica de alto desempenho e que possua uma larga memória.
- Os gráficos dinâmicos (evolução da rede com o tempo) são difíceis de elaborar, pois estão em *gexf format* (veja mais baixo), nem sempre documentado corretamente.

A Favor do *Gephi*

Arquivos de entrada e de saida

Gephi aceita vários formatos de entrada e de saída, que o tornam também numa plataforma de interface com outras opções de software (*Pajek, Netdraw*). Nem todas as funcionalidades estão disponíveis em todos os formatos.

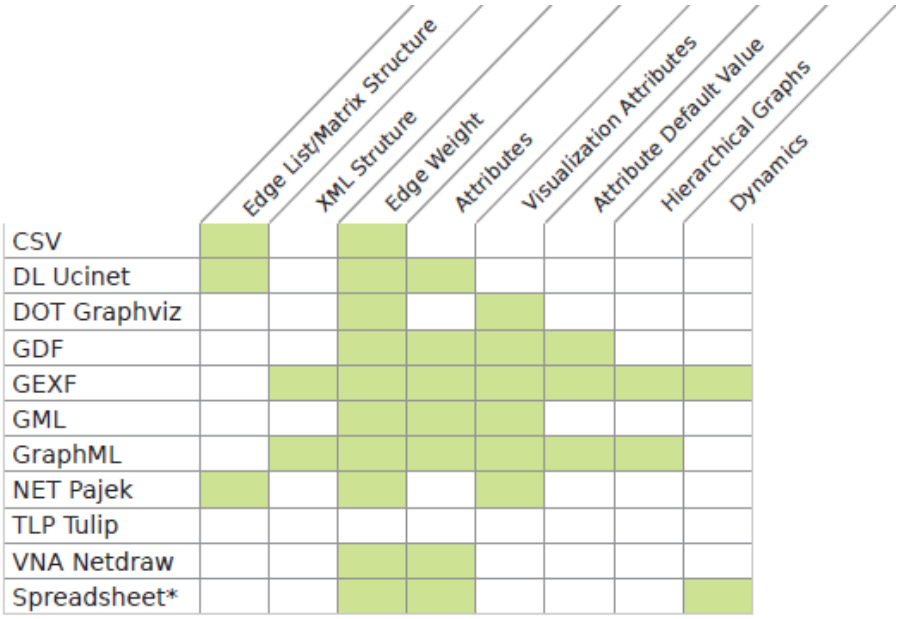

Figura 4: Formato de entrada para o *Gephi* e suas possibilidades.

CSV

O Gephi aceita CSV files (http://pt.wikipedia.org/wiki/Comma-separated_values) como entrada. Descrição do formato *csv* necessário se encontra em http://gephi.github.io/users/supported-graph-formats/csv-format/ e https://gist.github.com/futureperfect/2096812.

Caso venha a importar diretamente o arquivo de *edges (edges.csv)*, com os nomes dos *nodes* dentro, no gephi, ele criará tabelas dos nodes necessários (sendo necessário apenas copiar a coluna ID até o Label no laboratório de dados). É necessária uma atenção especial no cabeçalho (primeira linha) no arquivo CSV. Se usar o plugin Excel/csv converter (https://marketplace.gephi.org/plugin/excel-csv-converter-to-network/) é possível importar simultaneamente 2 campos multivariados (autor e palavra-chave) sendo que, para os demais campos, será necessário importar cada combinação possível de 2 em 2.

GDF

Formato do software GUESS (http://graphexploration.cond.org/). O formato GDF poderia ser chamado CSV generalizado, pois é simples como CSV, porém permite inserir atributos de cada node (UF, IES.). O formato completo está descrito em http://guess.wikispot.org/The_GUESS_.gdf_format e http://gephi.github.io/users/supported-graph-formats/gdf-format/. É este o formato que o Scriptlattes (http://scriptlattes.sourceforge.net/) utiliza para salvar a rede de co-autoria de um conjunto de ID Lattes fornecido nos arquivos de entrada. O ScriptLattes salva como atributo o geoposicionamento (posiçao geográfica baseada no endereço profissional do autor), que

representa a geolocalização com *Gephi* (não está 100%, pois nem sempre se encontra a posição do endereço).
Caso uma *url* (endereço web) seja passada como atributo do node da rede, *Geph*i se torna um *browser* que integra a rede com a Web. Para efetivá-lo é preciso utilizar um *plugin* (https://marketplace.gephi.org/plugin/linkfluence-plugin/).

Gexf
http://gexf.net/format/
É o formato especifico do *Geph*i. Encontra-se no formato XML, que faz a conversão usando o XSLT. Existem também bibliotecas em *python, Javascript, Java, Perl, R, C++* para construir *gexf files*, mas a documentação não é perfeita e ainda necessita de ajustes. Mas, quando falamos de hierarquia, é bastante dinâmico. O interessante é a biblioteca liberada no *github* (https://github.com/raphv/gexf-js) que deixa facilmente colocar um gráfico g*exf* na web (como exemplo, co-autores brasileiros que publicam sobre "dengue": http://vlab4u.info/doencas%20negligenciadas/Dengue/denguecore1/Resultados/Rede/gexf/index.html). É interessante salientar que esta representação não inclui os gráficos dinâmicos,e a hierarquia é limitada a 2D (3D seria o ideal). Mas este ponto pode ser visto como contribuição, já que o projeto está em "aberto".

Command line execution
https://gephi.github.io/toolkit/
Existe uma versão para rodar o *Gephi* 100% em linha de comando, caso seja utilizado para trabalho de "*back office*".

SPARQL plugin
https://marketplace.gephi.org/plugin/semanticwebimport/
Permite tratar arquivos locais RDF (http://pt.wikipedia.org/wiki/Resource_Description_Framework) e *vufind* (usado pela BDTD nova versão), sendo *Zotero* compatível. Deve possibilitar exportação de dados da bibliografia em RDF. Embora este *plugin* seja delicado de se operar, permite também fazer buscas SPARQL (http://en.wikipedia.org/wiki/SPARQL) e montar a rede *Gephi* no tempo real da busca. Um exemplo: usando a base DBPEDIA, que está disponível em http://blog.ouseful.info/2012/07/03/visualising-related-entries-in-wikipedia-using-gephi/.
DBpedia é um esforço da comunidade *crowdsourced* para extrair informações estruturadas da Wikipedia e disponibilizar essas informações na *web*. DBpedia permite fazer consultas sofisticadas contra Wikipedia e unir os diferentes conjuntos de dados na *Web* com os dados da Wikipédia. Seria interessante realizar um teste de busca de palavras-chave da BDTD na DBPEDIA com SPARQL para tentar mapear as áreas de conhecimento das teses na wikipedia.

Mapear Twitter com Gephi
http://matthieu-totet.fr/Koumin/tools/naoyun/

Tratamento dos campos texto integral

Quando não há palavras-chave, ou quando a consistência das mesmas não é boa, não é preciso recuar frente ao tratamento de campos de "texto integral" como títulos, resumo, descrição, ou até artigos completos. Existem soluções simples e elegantes que podem ser utilizadas e que não precisam de análise linguística detalhada envolvendo gramática para sua execução..

Treecloud

http://treecloud.univ-mlv.fr/

Instituição mantenedora

Philippe Gambette, professor adjunto em Paris (http://igm.univ-mlv.fr/~gambette/indexENG.php), criou este aplicativo durante o doutorado, mantendo desde então o acesso ao referido *software*.

Comentários

A entrada do software é muito simples: um arquivo *txt* contendo um texto a se analisar (títulos) por linha. A interface *Windows* só serve para construir a linha de comando necessária para executar o *script python*, que se conecta a um programa Java de representação. Então, deve ser possível incluir num "*back office*" de *site web.* O autor providenciou uma versão com essas características (http://treecloud.univ-mlv.fr/cgi-bin/NuageArbore_FR.cgi). O site do autor menciona que "Desde 2014, o módulo de "construção árvore-nuvem" lida com caracteres Unicode e inclui várias operações de pré-processamento linguístico (supressão *stopword*, detecção de expressões multi-palavra) realizada por Unitex" http://www-igm.univ-mlv.fr/~unitex/, e adicionadas por Claude Martineau http://hal-upec-upem.archives-ouvertes.fr/Public/afficheRequetePubli.php?auteur_exp=Martineau&collection_exp=LIGM&CB_auteur=oui&CB_titre=oui&CB_article=oui&langue=Francais&tri_exp=annee_publi&tri_exp2=typdoc&tri_exp3=date_publi&ordre_aff=TA&Fen=Aff&css=../css/VisuRubriqueEncadre.css". Esta melhoria ainda não foi adequadamente testada. Antes do seu surgimento era necessária uma limpeza do texto com os softwares notepad++ e wReplace para cautelosamente retirar diacriticos, caracteres numéricos e pontuação. Um resultado do uso com base nos títulos de artigos publicados mencionados no Lattes está disponível em: http://vlab4u.info/uninove/administracao/analisa/PR_titulos_prod_bibliografica/TreecloudText.txt.jaccard.colored.pdf.

Cowo e Vosviewer

https://github.com/seinecle/Cowo/blob/master/README.md
http://www.vosviewer.com/

Instituição mantenedora

- COWO esta no Github https://github.com/seinecle/Cowo e foi elaborado por Clement Levallois, um dos mantenedores do Gephi. Este aplicativo, embora longe de ser

perfeito, pode ser melhorado rapidamente (em java para poder trazer/utilizar uma técnica interessante para análise de texto integral: os *Ngrams* (http://en.wikipedia.org/wiki/N-gram que detalharemos mais abaixo).

- VosViewer foi criado por Nees Jan van Eck e Ludo Waltman http://www.vosviewer.com/aboutus/, e é mantido pelo *Centre for Science and Technology Studies* da Universidade de Leiden (Netherland) http://www.cwts.nl/Home, um dos maiores laboratórios de bibliometria que se conhece.

Comentários

Este software livre esta sendo desenvolvido em Java. A técnica usada é muito semelhante à utilizada no *treecloud,* embora o gráfico produzido pareça muito diferente. Um exemplo de resultado está disponível em http://vlab4u.info/uninove/administracao/analisa/PR_titulos_prod_bibliografica/cowo.pdf (mesmos dados do apresentado anteriormente e demonstrado para o software *treecloud*). VosViewer pode ser utilizado sem cowo (Ele somente é utilizado para um tratamento de entrada diferente. O arquivo da entrada tem que ser o mesmo que foi preparado para *treecloud*. Existe a possibilidade de uma lista de "stopwords", mas o formato da lista é diferente (confere manual).

JsLDA

https://github.com/mimno/jsLDA

Instituição mantenedora

David Mimno (Cornell University, USA - http://mimno.infosci.cornell.edu/), liberou uma versão Javascript do algorithmo "*Latent Dirichlet allocation*" (http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation), baseado em um modelo probalístico e vetorial para encontrar tópicos num conjunto de "textos integrais". A vantagem é ser "independente do idioma", embora inclua a possibilidade de ter lista de *stopwords.* Existem vários algoritmos recentes baseados nestas técnicas.

Comentários

O formato de entrada é muito simples e semelhante ao arquivo usado para *treecloud* ou Vosviewer, permite colocar um "*labe*l" na entrada de cada linha. Embora as técnicas de modelo de espaço vetorial não sejam recentes (Salton em 1975, em http://en.wikipedia.org/wiki/Vector_space_model), as capacidades das máquinas atuais deixam revisitar este tipo de modelo, já que possuem capacidade de cálculo suficiente. Seria de grande importância criar ferramentas de tecnologias híbridas incluindo vários modelos, cada um com a sua contribuição.

IRAMUTEQ

http://www.iramuteq.org/

Instituição mantenedora

Criado e mantido por Pierre Ratinaud (http://www.lerass.com/?profile_cct=pierre-ratinaud), Iramuteq esta no *sourceforge* na base do GNU GPL (v2).

Comentários

Especializado em análise de "texto integral" (principalmente entrevistas semi-direcionadas), Iramuteq usa algoritmos "clássicos" (sem custo) como a "*divisive Hierarchical clustering*" do software Alceste (http://pt.wikipedia.org/wiki/Alceste_%28software%29) ou a "*correspondence analysis*" (http://en.wikipedia.org/wiki/Correspondence_analysis), ambos baseados na detecção de "segmentos de texto repetidos", construindo os scripts automaticamente para o R. Existe uma documentação inicial em português (http://www.iramuteq.org/documentation/fichiers/tutoriel-en-portugais), elaborada por professores da UFSC (http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4787871Z8 e http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4256976T8). Essa solução está, na minha opinião, ainda em fase de teste para a língua portuguesa. O formato de entrada, embora na forma de arquivo txt, é bem específico, mas o interessante é que o arquivo permite fazer um "mix" de "texto completo" e variáveis estruturantes (IES, UF, Data, etc.).

Ngrams

Embora pouco conhecido para aplicações pessoais, a tecnologia dos *Ngrams* é amplamente usada para fins de "*approximate maching"* http://en.wikipedia.org/wiki/N-gram#n-grams_for_approximate_matching. *Ngrams* tem sido usada em larga escala por:

- *Google*: https://books.google.com/ngrams/, https://books.google.com/ngrams/info,
- *COCA* : http://www.ngrams.info/intro.asp. Estes *Ngrams* estão baseados, em sua maioria, em "*corpus*" que está publicamente disponível. Há 450 milhões de palavras no *Corpus of Contemporary* Inglês Americano (COCA). Com esses dados *Ngramas* (2, 3, 4, 5 sequências de palavras, com sua frequência), você pode realizar consultas poderosas *off-line* - sem a necessidade de acessar o *corpus* através da interface *web*. Existe uma base dos *Ngrams* liberada em português http://www.ngrams.info/portuguese.asp.
- *Microsoft* : http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx
- *W3C* : http://www.w3.org/TR/ngram-spec/

Existe em bibliotecas de programação:

- *Python*: (http://www.nltk.org/api/nltk.html?highlight=ngrams#nltk.util.ngrams) - a biblioteca NLTK (*Natural language toolkit* http://www.nltk.org/) inclui tal funcionalidade, embora a NLTK seja complicada de usar (*full linguistic features*)
- *PERL*: http://search.cpan.org/dist/Text-Ngrams/Ngrams.pm

Cowo é uma tentativa de aplicativo para detectar os *Ngrams* para análise. Mas, considerando os monogramas e somente os *Ngrams* em nível de palavras (e não de caracteres), os resultados não mostram claramente o interesse da metodologia. Sendo em código aberto, uma melhoria deste aplicativo é possível acoplando uma "*lematização"* http://en.wikipedia.org/wiki/Lemmatisation ou *sitematização* http://en.wikipedia.org/wiki/Word_stem", baseada nos *Ngrams* nas representações *Treecloud, VosViewer, Js-LDA,* o que poderia ser de grande contribuição.

Campos que permitem fazer relações

Hoje existem bases de dados que dão acesso a grandes e valiosos recursos de informação. Como mantenedor de base de dados, pode ser valioso ter acesso e usar este recurso para melhorar as bases (vocabulário controlado), ou incluir complementos de informação visando aumentar o valor / consistência da(s) base(s). Aqui serão expostos 3 recursos deste tipo. A pretensão é que não seja exaustivo.

Lattes

A plataforma Lattes, mesmo sendo imperfeita, se tornou uma plataforma imprescindível para a ciência brasileira. Também se constituiu em uma forma engenhosa de se conseguir uma base para levantar a produção cientifica brasileira com baixo custo. Ela se baseia no modelo 2.0, da qual cada um participa (neste caso os pesquisadores,) sem grande controle externo. Hoje ela contempla mas de 1,5 milhão de currículos de doutores brasileiros. Um fato muito importante na plataforma Lattes é a presença de uma identificação personalizada de cada currículo com um número único (ID Lattes). Por exemplo, meu currículo na plataforma está com o número de 10 dígitos "K4775374H1", e com o número de 16 digitos "4754764003480925", com apenas um nome e um CPF vinculado a eles. Os 2 sistemas de numeração são equivalentes e existem em um sistema de conversão para se poder chegar de um até o outro. Torna-se, portanto, um sistema valioso para correção de nomes de pesquisadores brasileiros, pois com o(s) número(s), é possível encontrar o nome correto, tal como cadastrado na plataforma Lattes pelo pesquisador. Em anexo, foram liberados *scripts* (em *python*) que permitem, à partir de uma lista de nomes, encontrar os ID Lattes 10 de cada nome. Para os casos de ambiguidade é necessário que sejam resolvidos manualmente (transformação dos ID 10 em ID 16). Este conjunto ajudaria a completar o *link* para acesso a todos os currículos Lattes da base BDTD (doutorando ou mestrando, orientador, demais membros da banca). Permitindo baixar como campo os ID Lattes 16, permitiria-se ao usuário executar o ScriptLattes à partir da saída do campo. Existem inúmeras aplicações interessantes: produção cientifica de um conjunto de membros de bancas, produção cientifica de doutorando 10 anos após defesa, produção cientifica de orientadores/programa/IES, identificação das contribuições de um grupo de pesquisadores á sociedade através de suas produções técnicas e tecnológicas, identificação das redes de colaboração entre pesquisadores, visualização dos futuros resultados das pesquisas em andamento (*forecasting*), dentre outras.

Quando o ScriptLattes é executado, ele utiliza como "entrada" o ID Lattes do pesquisador, e baixa o nome exato do Lattes. Olhando o ScriptLattes, seria fácil, especialmente com ajuda de programador habilitado para tal, isolar a parte do código que recupera o nome exato do Lattes com ID 16. Tal etapa poderia ser utilizada para corrigir os nomes, na BDTD, de doutorando, mestrando, orientador e demais membros da banca para se conseguir corespondência exata com os nomes constantes no Lattes.

Wikipedia/DBpedia

Embora existam críticas à *Wikipedia* http://pt.wikipedia.org/wiki/Cr%C3%ADticas_%C3%A0_Wikip%C3%A9dia), ela não pode ser considerada 100% "ruim", pois:

- Representa quase 46 milhões de dólares (somente a Inglesa), em economia, se olharmos o equivalente financeiro do esforço dos colaboradores das enciclopédias tradicionais fornecido pelos contribuintes da wikipedia (http://en.wikipedia.org/wiki/User:UBX/wikivalue e http://infodisiac.com/blog/2008/10/quantifying-volunteer-contribution/)
- Representa 100 vezes o volume da "Britanica" (somente a inglesa) (http://en.wikipedia.org/wiki/Wikipedia:Size_in_volumes)
- Esta disponível em 287 idiomas, das quais 52 são as mais "expressivas" com mais de 100.000 entradas http://meta.wikimedia.org/wiki/List_of_Wikipedias

Então, ela se torna a maior fonte de informação enciclopédica e de acesso livre e gratuito do mundo. E cabe a nós melhorá- la ou completá-la, se julgamos que ela deva ser aperfeiçoada. Mas, no entanto, a *Wikipedia* pode ser de grande ajuda para a ciência da informação, base(s) de dados, entre outros com o projeito DBPedia e com a API da *Wikipedia.*

DBpedia

http://dbpedia.org/About

É um esforço da comunidade (*crowdsourced*) para extrair informações estruturadas da *Wikipedia* e disponibilizar essas informações na web. *DBpedia* permite consultas sofisticadas na *Wikipedia*, e ainda possibilita interligar os diferentes conjuntos de dados na *Web* com dados da *Wikipédia*. Espera-se que este trabalho seja o mais adequado para a enorme quantidade de informações disponibilizadas na *Wikipédia,* podendo ser utilizado em algumas novas maneiras interessantes. Além disso, ele pode inspirar novos mecanismos para navegar, fazer vinculação e melhorar a própria enciclopédia. É bom citar:
Ligação da *DBPedia* com outros conjuntos de dados (http://wiki.dbpedia.org/Interlinking).
Acesso a *DBpedia* na *web* (http://wiki.dbpedia.org/OnlineAccess) (enriquecer a BDTD com dados complementares: geoposicionamento, relações de conceitos), e a possibilidade de *"linkar*" um texto com entradas da *Wikipedia* para documentá-lo (https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki), o que possibilitaria, por exemplo, ilustrar com a W*ikipedia* todos os resumos *(abstracts*) das dissertações e teses da BDTD. Experimente o demo em http://dbpedia-spotlight.github.io/demo/.

API da Wikipedia

As *Application Programming Interface* (em português: Interface de Programação de Aplicações (http://pt.wikipedia.org/wiki/API) permite, por meio de programação, usar os recursos de um servidor. A *Wikipedia* libera uma API (http://en.wikipedia.org/w/api.php) que autoriza fazer varias ações com a *Wikipedia*. Existe também uma área de teste (http://en.wikipedia.org/wiki/Special:ApiSandbox). Algumas ideias usando o API da *Wikipedia*:

- *Faviki* (http://www.faviki.com/pages/welcome/) é uma ferramenta de *bookmarking social* (http://pt.wikipedia.org/wiki/Social_bookmarks) que permite que você use conceitos *Wikipédia* como *tags*. Infelizmente não se tornou muito popular. Mas o conceito é interessante por oferecer uma normatização das palavras-chave escolhidas (em quantidade, qualidade, e definição). Poderia ser utilizado na colocação de palavras-chave numa base de dados (BDTD ou outra), até criando um novo campo de palavras-chave. Este recurso poderia ser utilizado não somente na digitação, mas também para rever todo um vocabulário existente. Os *links* seguintes fornecem um *start* de solução de programação http://www.labnol.org/internet/tools/using-wikipedia-api-demo-source-code-example/3076/ e http://anexiledderryman.com/post/52858733180/html5-datalist-control-dynamic-auto-complete-with.
- O "*book creator*" da *Wikipedia* (http://en.wikipedia.org/wiki/Help:Books), Pedia *press* (http://pediapress.com/code/), e *Wikipedia,* liberaram a possibilidade para o usuário de criar e baixar *e-books* (em vários formatos), sem direitos autorais, e contendo qualquer conjunto de páginas da *Wikipedia*. Poderia, por exemplo, ser usado com a base de Repostas Técnicas para fornecer documentação completa, ou para documentar uma tese ou dissertação. É necessário verificar mais a fundo se é possível automatizar o processo por programação.
- Poderia também ser propiciada tradução automatizada em Inglês e Francês das palavras-chave (normalizadas da *Wikipedia*), pois a API permite encontrar as páginas *Wikipedia* em vários idiomas.

IPC

A Classificaçao Internacional de Patente (http://www.wipo.int/classifications/ipc/en/), mantida pela WIPO (Organização Internacional de propriedade Intelectual), é um sistema complexo de descrição do conteúdo das patentes (que não tem palavras-chave). A IPC é atribuída a todas as patentes, qualquer que seja a língua ou o país de depósito. Torna-se, então, um sistema único de descrição da matéria patentária. Ela é mantida em dois idiomas, o Inglês e o Francês, é consultável *online* e liberada para *download.* Para uma abordagem mais simples, existem 2 dicionários que deixam conectar palavras-chave com a classificação. Estes dicionários se chamam "*Catchwords*" e existem em Inglês (http://web2.wipo.int/ipcpub/#&notion=cw) e francês (http://web2.wipo.int/ipcpub/#&notion=cw&lang=fr). Não se trata de tradução, já que são

complementares. Um exemplo elegante do uso que pode ser feito deste recurso é o "*green Inventory*" (http://www.wipo.int/classifications/ipc/en/est/). O "Inventário IPC Verde" foi desenvolvido por uma Comissão de Peritos do IPC, a fim de facilitar a busca por informações de patente relacionadas com as chamadas Tecnologias Ambientalmente Saudáveis (ESTs), conforme listado pela Convenção-Quadro das Nações Unidas sobre Mudanças Climáticas (UNFCCC). BDTD, A base de Respostas Técnicas poderia ter uma ligação com o mundo das patentes.

Para aumentar o conhecimento

Zotero

Um gerenciador de referências bibliográficas de código aberto, arquivador de documentos (https://www.zotero.org/), automatizador de citações e referenciamento bibliográfico, que funciona muito bem acoplado ao Mozilla Firefox (https://www.mozilla.org/pt-BR/). Possibilita, quando baixadas as referências e salvas em uma biblioteca interna, anexar os arquivos .pdf relacionados a cada uma das obras. Atualmente possibilita armazenar 300 mega de informações por usuário, permitindo inclusive que diversos usuários compartilhem referências no site do próprio programa, dinamizando o processo e aumentando a produtividade em pesquisa. Atende a mais de 7 mil estilos de citação e é constantemente discutido e melhorado por uma ampla gama de usuários ao redor do mundo. É superior a seus concorrentes por ser absolutamente gratuito e de fácil operação.

Orange

http://orange.biolab.si/

Visualização de dados em código aberto e de análise para iniciantes e especialistas. A mineração de dados se dá através de uma programação visual ou script *Python*. Possui componentes para aprendizado de máquina. Oferece extras para bioinformática e mineração de texto. Embalado com características para análise de dados. Na minha opinião, é mais direcionado para a ciência da computação do que para a ciência da informação, embora as coisas evoluam muito rapidamente. No mestrado sob minha responsabilidade na França, agora há aulas de programação *Python* como parte da grade curricular.

Papermachines

http://papermachines.org/

Já que a nova plataforma BDTD inclui formatos compatíveis com o *software* de gestão automática de bibliografias *Zotero* (https://www.zotero.org/), uma coisa a ser testada é o *plugin* do *Zotero* "*Paper Machines*" para ver como um aluno qualquer poderia estudar o ineditismo do trabalho que ele imagina, baixando da BDTD referências de teses e dissertações, e visualizando direitamente com o *plugin*. Por exemplo :

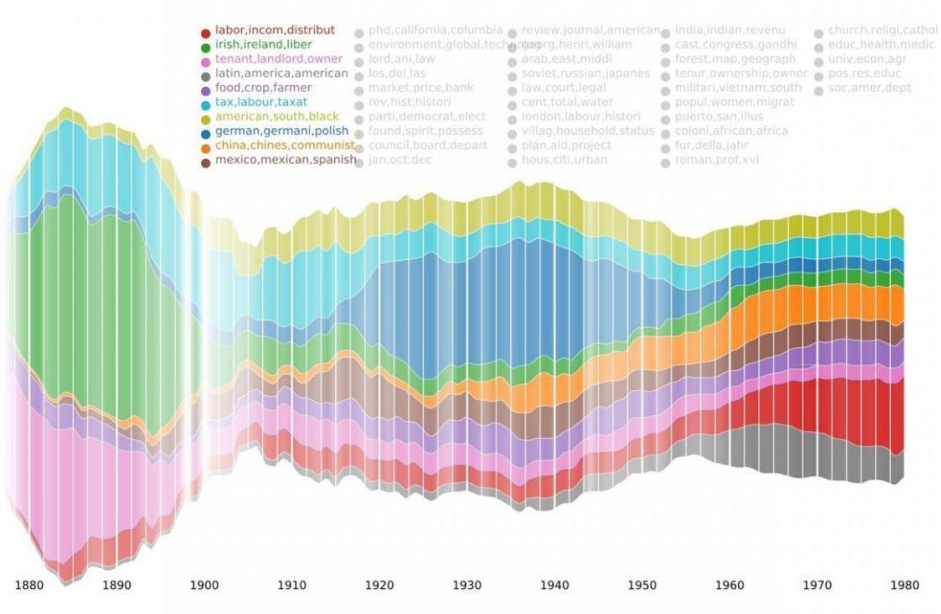

Figura 5: Evolução temporal do uso de palavra chaves usando *Paper Machines.*

Figura 6: "*Word cloud*" gerada automaticamente usando o *Paper Machines.*

Proposta de capacitação da equipe

Este relato inteiro constitua o manual do curso (inclusive o conteúdo das paginas web mencionadas) a ser dispensado para capacitação da equipe, uma vez selecionadas as soluções conveniente para o IBICT (na aplicação das bases da instituição) e as equipes a ser treinado. Este curso pode ser (conforme conveniente para o IBICT) efetuado em presencial (primeiro período possível entre 12/11 e 23/11); ou na forma de apoio permanente pelas conexões distantes (Skype ou visio-conferencia).

Recomendações finais

As soluções apresentadas atendem aos requisitos de código aberto e software livre, mas são em sua maioria desenvolvidos na esfera acadêmica. Embora sejam livres e gratuitos, acredito que é preciso ter a consciência do esforço feito pelos autores dos softwares apresentados no desenvolvimento deste trabalho. Então, ao usá-los, é preciso ter em mente o desenvolvimento de políticas para "retribuir" aos autores , fortalecendo e estimulando seus esforços. Isto poderia ser feito:

- Citando sempre o trabalho desses autores quanto for utilizado, não somente nos artigos científicos mas também nas paginas *web*. Para ajudar neste quesito poderiam ser levantadas as referências científicas do autor para ajudar a divulgá-los. Também ajudaria a embasar mais suas escolhas na teoria e na ampliação do conhecimento do produto.
- Chamando para consultorias os autores das soluções selecionadas a serem implantadas. Os autores, não somente ficarão felizes (pois isto pode se constituir em uma possibilidade de reconhecimento do trabalho), bem como ajudarão a melhorar o produto, ou possibilitar a sua implementação.
- Contribuindo na melhoria do produto e disponibilizando as melhorias para o autor e/ou para a comunidade na *web*.

Então como já foi mencionado na introdução, estas soluções nem sempre são mais baratas, mas as despesas são dosadas de maneira mais adequada (no tempo e no espaço), e são resultados de mudanças de políticas mais do que estritamente de redução de custos.

Este relatório foi co-elaborado numa plataforma colaborativa (*Google drive*) que permitiu uma interação "em tempo real". Agradeço a participação de todos, particularmente à Kira, Lillian, amigas de coração, Carol e Renato que tiveram a coragem de tornar o meu Português razoável.