Uma arquitetura para mecanismos de buscas na web usando integração de esquemas e padrões de metadados heterogêneos de recursos educacionais abertos em repositórios dispersos

Murilo Gleyson Gazzola

Murilo Gleyson Gazzola

Uma arquitetura para mecanismos de buscas na web usando integração de esquemas e padrões de metadados heterogêneos de recursos educacionais abertos em repositórios dispersos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestra em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA.*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Cristina Dutra de Aguiar Ciferri

USP – São Carlos
Dezembro de 2015

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

G289a

Gazzola, Murilo Gleyson Gazzola
Uma arquitetura para mecanismos de buscas na web usando integração de esquemas e padrões de metadados heterogêneos de recursos educacionais abertos em repositórios dispersos / Murilo Gleyson Gazzola Gazzola; orientadora Cristina D. A. Ciferri Ciferri. -- São Carlos, 2015.
133 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2015.

1. Recuperação de informação. 2. Banco de dados. 3. Engenharia de software. I. Ciferri, Cristina D. A. Ciferri, orient. II. Título.

Murilo Gleyson Gazzola

An architecture for web search engines using integration of heterogeneous metadata schemas and standards of open educational resources in scattered repositories

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. FINAL VERSION.

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Cristina Dutra de Aguiar Ciferri

USP – São Carlos
December 2015

Agradecimentos

Agradeço inicialmente à Deus, aos meus pais Elenice Aparecida Ferrari Gazzola e Paulo Cesar Gazzola, pela educação, amor, carinho, dedicação dados a mim. Ao meu irmão Paulo Cesar Gazzola Junior pelo apoio e incentivo. Ao meu tio Jovelino Gazzola, pela motivação, apoio e incentivo na área de pesquisa acadêmica. E, obrigado por tolerarem minha ausência e entenderem minha dedicação aos estudos.

À minha orientadora, Profa. Dra. Cristina Dutra de Aguiar Ciferri, que sempre me manteve motivado, confiante e apoiou desde o início esta dissertação. Sempre demonstrando paciência, dedicação, ensinamentos que teve ao longo da carreira e pelas correções ao longo de todo o trabalho. À Profa. Dra. Itana Maria de Souza Gimenez, pelos ensinamentos, correções, conselhos, dedicação e apoio mesmo a distância. E, que foi a pessoa fundamental para iniciar minha carreira no mestrado e ao tema.

À minha banca de qualificação e de defesa, profa. Dra. Carmem Satie Hara, profa. Dra. Ellen Francine Barbosa e profa. Dra. Marcela Xavier Ribeiro, pelas correções e encaminhamento do projeto na época muito importante do trabalho. Com certeza, fizeram toda a diferença pelas ótimas sugestões, comentários e dedicação que foram determinantes para os avanços na estrutura,

desenvolvimento e resultados.

À minha futura esposa e namorada há 8 anos, Cinthia Suemy Tomyama, pelo apoio, companheirismo, incentivo, compreensão e pelos ensinamentos da cultura japonesa sempre demonstrando equilíbrio, equidade e uma ótima cozinheira com pratos maravilhosas da culinária japonesa.

Ao professor Dr. Ricardo Rodrigues Ciferri, pelas correções realizadas nesta dissertação e sempre interessado no assunto, além de ser uma pessoa estimulado e organizado nas atividades inter-núcleos de pesquisa do GBD e do GBdI para os avanços interdisciplinares e de pesquisas. Agradeço aos meus amigos Nathan S. Hartman, Alessandro Yovan Bokan Garay, Roque E. Lopez Condori, Marco Antonio Sobrevilla Cabezudo, e todos meus amigos do Grupo de Bases de dados e de Imagens (GBdI) e do Núcleo Interinstitucional de Linguística Computacional (NILC) do ICMC da USP e Grupo de Banco de Dados (GBD) da UFSCar que me ajudaram de alguma forma.

Pelo apoio do instituto ICMC/USP que financiou toda a infraestrutura necessária, como também apresentação do artigo desta dissertação. Ao CNPQ, FAPESP e outros órgãos de fomento à pesquisa que de alguma forma ajudaram na conclusão deste trabalho.

Resumo

Recursos Educacionais Abertos (REA) podem ser definidos como materiais de ensino, aprendizagem e pesquisa, em qualquer meio de armazenamento, que estão amplamente disponíveis por meio de uma licença aberta que permite reuso, readequação e redistribuição sem restrições ou com restrições limitadas. Atualmente, diversas instituições de ensino e pesquisa têm investido em REA para ampliar o acesso ao conhecimento. Entretanto, os usuários ainda têm dificuldades de encontrar os REA com os mecanismos de busca atuais. Essa dificuldade deve-se principalmente ao fato dos mecanismos de busca na Web serem genéricos, pois buscam informação em qualquer lugar, desde páginas de vendas até materiais escritos por pessoas anônimas. De fato, esses mecanismos não levam em consideração as características intrínsecas de REA, como os diferentes padrões de metadados, repositórios e plataformas existentes, os tipos de licença, a granularidade e a qualidade dos recursos. Esta dissertação apresenta o desenvolvimento de um mecanismo de busca na Web especificamente para recuperação de REA denominado SeeOER. As principais contribuições desta pesquisa de mestrado consistem no desenvolvimento de um mecanismo de busca na Web por REA com diferenciais entre os quais se destacam a resolução de conflitos em nível de esquema oriundos da heterogeneidade dos REA, a busca em repositórios de REA, a consulta sobre a procedência de dados e o desenvolvimento de um *crawler* efetivo para obtenção de metadados específicos. Além disso, contribui na inclusão de busca de REA no cenário brasileiro, no mapeamento de padrões de metadados para mecanismos de busca na Web e a publicação de uma arquitetura de um mecanismo de busca na Web. Ademais, o SeeOER disponibiliza um serviço que traz um índice invertido de busca que auxilia encontrar REA nos repositórios dispersos na Web. Também foi disponibilizada uma API para buscas que possibilita consultas por palavras chaves e o uso de palavras booleanas. A forma de validação em mecanismos de busca na Web, como um todo, e de forma quantitativa e específica por componentes foi feita em grau de especialidade. Para validação de qualidade foram considerados 10 participantes com grupos distintos de escolaridade e área de estudo. Os resultados quantitativos demonstraram que o SeeOER é superior em 23.618 REA indexados em comparação a 15.955 do Jorum. Em relação à qualidade o SeeOER demonstrou ser superior ao Jorum considerando a função penalizada e o *score* utilizada nesta pesquisa.

Palavras-chaves: mecanismo de busca na Web, recursos educacionais abertos, integração de dados, procedência de dados.

Abstract

Open Educational Resources (OER) has been increasingly applied to support students and professionals in their learning process. They consist of learning resources, usually stored in electronic device, associated with an open license that allows reuse, re-adaptation and redistribution with either no or limited restrictions. However, currently the Web search engines do not provide efficient mechanisms to find OER, in particular, because they do not consider the intrinsic characteristics of OER such as different standards of metadata, repositories and heterogeneous platforms, license types, granularity and quality of resources. This project proposes a Web search engine, named SeeOER, designed to recover OER. Main features of SeeOER are: schema-level conflict resolution derived from the heterogeneity of OER, search for Brazilian OER repositories, query considering data provenance and the development of an effective crawler to obtain specific metadata. In addition, our project contributes to the inclusion of the search OER research issues in the Brazilian scenario, to the mapping of metadata standards to Web search engine. In addition, SeeOER provides a service which internally has an inverted index search to find the OER which is different from traditional Web repositories. We also provide an API for queries which make it possible to write queries based on keywords and boolean. The validation of the search engine on the Web was both qualitative and quantitative. In the quantitative validation it was observed in level of specialty of the search engines components. In conclusion, the quality and quantitative results experiments showed that SeeOER is superior in OER indexed 23,618 compared to 15,955 the Jorum. In relation to the quality SeeOER shown to be superior to Jorum 27 points considering the metric used in project.

Keywords: Web search engine, open educational resources, metadata standards, integration schemes.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo

1

Introdução

Recursos Educacionais Abertos (REA)[1] podem ser definidos como materiais de ensino, aprendizagem e pesquisa, em qualquer meio de armazenamento, que estão amplamente disponíveis por meio de uma licença aberta que permite reuso, readequação e redistribuição sem restrições ou com restrições limitadas (Atkins et al., 2007; Gimenes et al., 2012; Kanwar et al., 2011). REA podem incluir cursos completos, partes de cursos, módulos, guias para estudantes, anotações, livros didáticos, artigos de pesquisa, vídeos, instrumentos de avaliação, recursos interativos como simulações e jogos de interpretação, bancos de dados e aplicativos, dentre outros recursos. O conceito foi cunhado em 2002 pela UNESCO (Wiley, 2002) e está relacionado ao conceito de objetos de aprendizagem que são pequenos recursos digitais modulares focados em objetivos educacionais (Freire et al., 2008; Wiley, 2002).

Os REA enfatizam o conceito de abertura, com destaque para licenças abertas como as do tipo *Creative Commons*[2] e GNU Free Documentation License[3]. Isso significa que não existem custos associados à licença ou ao compartilhamento do recurso quando do reuso de seu conteúdo, como ocorre em materiais que utilizam direitos autorais tradicionais (Spector et al., 2007). Em especial, os REA vislumbram a ampla disseminação de seus conteúdos para promover as ações denominadas "4*R*" (Wiley, 2010): (i) reusar: o direito de usar o conteúdo em sua forma original ou modificada (ex.: fazer uma cópia); (ii) revisar: o direito de adaptar, ajustar, modificar, ou alterar o conteúdo (ex.: traduzir o conteúdo para outra língua); (iii) remixar: o direito de combinar o conteúdo original ou

[1] Open Educational Resources (OER)
[2] `http://creativecommons.org/licenses/`
[3] `http://www.gnu.org/copyleft/fdl.html`

o conteúdo revisado com outro conteúdo para criar algo novo (ex.: incorporar o conteúdo em um *mash up*); e (iv) redistribuir: o direito de compartilhar cópias do conteúdo original, das revisões ou mixagens com outros (ex.: disponibilizar uma cópia do conteúdo para um amigo).

O *software* de acesso e disponibilização de REA é referenciado por vários termos como *framework*, plataforma, ambiente ou simplesmente *software* ou aplicação (Gimenes et al., 2012). Nesta dissertação, é usado o termo repositório REA para referenciar o local de acesso, armazenamento e disponibilização de REA, enquanto que o termo plataforma de REA é usado para representar um repositório REA com extensões para oferecer funcionalidades adicionais aos usuários, como exemplo a possibilidade de criar um espaço público para publicação de seus materiais. Exemplos de plataformas de REA incluem OpenLearn (*The Open University*)[4], Connexions (*Sharing Knowledge and Building Communities*)[5] e TheOrangeGrove[6]. Já exemplos de repositórios incluem OCW-MIT[7], OCW-UNICAMP[8] e Domínio Público[9]. Esses repositórios e plataformas são desenvolvidos usando padrões de metadados, dentre os quais os mais utilizados são Dublin Core Metadata Element Set e o IEEE Learning Object Metadata (McClelland, 2003).

O uso e disseminação de REA na Web vêm contribuir para diminuir a desigualdade educacional, oriundos de diversos fatores como: as condições geográficas, jovens e adultos sem acesso à Educação de qualidade (Hilu et al., 2015) e os altos custos de livros e revistas científicas. Além disso, o interesse por REA é cada vez mais crescente, pois eles constituem um dos principais elementos da educação aberta que visa permitir o acesso gratuito ao conteúdo educacional de forma global. Como exemplo, pode-se citar a declaração da cidade do Cabo, a qual reúne pesquisadores de diferentes países, e tem como objetivo acelerar esforços para promover REA, práticas tecnológicas e de ensino na educação[10]. Atualmente, diversas instituições de ensino e pesquisa têm investido no uso de REA para a disponibilização de conteúdo relacionado à educação, como The Open University[11] (Little et al., 2011; Okada, 2007), Stanford University[12], MIT[13], Unicamp[14], FGV[15], UNESP[16]

[4] http://www.open.edu/openlearn/
[5] http://cnx.org/
[6] http://www.theorangegrove.org
[7] http://ocw.mit.edu/
[8] http://www.ocw.unicamp.br/
[9] http://www.dominiopublico.gov.br/
[10] http://www.capetowndeclaration.org/read-the-declaration
[11] http://openlearn.open.ac.uk/
[12] http://class2go.stanford.edu
[13] http://ocw.mit.edu
[14] http://www.ggte.unicamp.br/e-unicamp/public/
[15] http://www5.fgv.br/fgvonline/Cursos/Gratuitos
[16] http://www.unesp.br/unespaberta

e USP[17] (Gazzola et al., 2014). Outra vertente da educação aberta são os *Massive Open Online Courses* (MOOC), os quais podem oferecer uma educação alternativa e de qualidade tanto na complementação da educação tradicional quanto na formação continuada (Matkin, 2013).

1.1 Contexto e Motivação

Os mecanismos atuais de busca na Web dificultam a identificação de REA e, portanto, prejudicam a sua disseminação e incorporação em práticas educacionais. Essa dificuldade deve-se a dois principais fatores. Primeiro, os mecanismos de busca na Web são genéricos, assim buscam informação em qualquer lugar, desde páginas comerciais até definições escritas por pessoas anônimas (por exemplo, Wikipédia). Segundo, eles não levam em consideração as características intrínsecas de REA. Características intrínsecas de REA são dados dispostos pelo usuário e armazenados no repositório por meio dos metadados, é uma forma de diferenciar de uma página na Web com um REA. Alguns dos principais problemas específicos na área de recuperação de REA na Web são descritos a seguir. Existem diferentes *padrões de metadados, repositórios e plataformas* disponíveis. Os REA têm sido construídos sem a utilização adequada desses padrões, repositórios e plataformas, o que tem gerado diversos problemas de heterogeneidade, tanto em nível de esquema quanto em nível de instância. Por exemplo, segundo Dietze et al. (2012), os diferentes repositórios são isolados uns dos outros e usam como base diferentes tipos de aplicação.

1.2 Objetivos

Os objetivos desta dissertação são identificar os principais padrões de metadados, desenvolver e avaliar uma arquitetura para um mecanismo de busca na Web por REA que considere os diferentes padrões de metadados instanciados nos repositórios REA, a integração em nível de esquemas e a inclusão dos diversos repositórios e plataformas nacionais e internacionais na Web por REA. A metodologia usada está descrita no Capítulo 6. Foi desenvolvida uma arquitetura para este mecanismo e no desenvolvimento de seus componentes específicos enfatizou-se obtenção de REA na Web, recuperação dos diferentes padrões de metadados instanciados de forma heterogênea, obtenção da procedência dos dados e resolução de conflitos em nível de esquema oriundos do uso de diferentes padrões de metadados, repositórios e plataformas.

[17] `http://eaulas.usp.br/portal/home`

Esses objetivos levaram ao estabelecimento da seguinte tese.

Tese: Um mecanismo de busca na Web especificamente projetado para levar em consideração as características intrínsecas de REA torna o resultado da busca vertical e focado para REA.

Assim, a hipótese formulada para este projeto, conforme segue.

Hipótese: A recuperação de REA e seus metadados na Web, a resolução de conflitos em nível de esquema e em nível de instância oriundos do uso de diferentes padrões de metadados, repositórios e plataformas de REA difere da resolução de conflitos de mecanismos de busca na Web existentes.

REA possuem padrões de metadados, repositórios e plataformas com características particulares, os quais introduzem heterogeneidades específicas que devem ser tratadas. Ademais, REA encontram-se armazenados em repositórios e plataformas que disponibilizam o acesso aos seus recursos por meio de interfaces de consultas criadas especificamente para esse fim. Nesse contexto, as páginas contendo os recursos solicitados são geradas em resposta às consultas realizadas por meio dessas interfaces. Mecanismos de busca na Web genéricos, como os descritos na Seção 5.1, não consideram essas particularidades. Portanto, grande parte dos recursos disponíveis permanece escondido dos mecanismos de busca. Isto prejudica a incorporação de REA nas práticas educacionais. Deve-se considerar também que mecanismos de busca na Web por REA, como os descritos na Seção 5.2, são limitados, pois não tratam a heterogeneidade dos REA atualmente disponíveis.

De um ponto de vista mais abrangente, a pesquisa desenvolvida nesta dissertação de mestrado visa incentivar as práticas de utilização e produção de REA na educação, pois cria um mecanismo para facilitar a identificação desses recursos que pode ser utilizado por professores e aprendizes das mais diversas área do conhecimento. Portanto, vislumbra-se também que o mecanismo de busca na Web desenvolvido tenha um alto impacto social.

1.3 Estrutura da Dissertação

Além deste capítulo introdutório, essa dissertação de mestrado possui mais seis capítulos, estruturados da seguinte forma.

- No Capítulo 2 são descritos os REA e seu padrões de metadados.
- No Capítulo 3 são descritos conceitos relacionados à recuperação de informação dentro do contexto desta dissertação. Também será abordada a recuperação de

informação na Web, tratando-se de *Web Crawler* e uma arquitetura teórica de um mecanismo de busca na Web.

- No Capítulo 4 são detalhados conceitos relacionados à integração de dados e à procedência de dados.
- No Capítulo 5 são descritos trabalhos correlatos e detalhadas as justificativas para o desenvolvimento do projeto.
- No Capítulo 6 é apresentada a arquitetura desenvolvida de um mecanismo de busca na Web, considerando uma abordagem de recuperação de REA e os padrões de metadados inspecionados na Web os quais facilitam a catalogação, pesquisa e reutilização de REA. São apresentados a arquitetura e os detalhes de cada componente.
- No Capítulo 7 são apresentados os experimentos e os resultados obtidos.
- No Capítulo 8 são apresentadas as conclusões deste trabalho.

Capítulo

2

Recursos Educacionais Abertos

A proposta do mecanismo de busca desenvolvido visa identificar na Web, REA em repositórios e plataformas heterogêneos, os quais são desenvolvidos de acordo com diferentes padrões de metadados.

Nesse contexto, na Seção 2.1 é definido o termo REA, são identificados quais formatos de REA são encontrados na Web, é definida a importância de REA como recursos livres e reutilizáveis para o ensino e aprendizagem, e é salientada a repercussão do termo REA no Brasil. Na Seção 2.2 são descritos os padrões de metadados que facilitam a catalogação, pesquisa e reutilização de REA, os quais serão usados no desenvolvimento deste projeto. Foram inspecionados alguns repositórios de REA, considerando os repositórios que armazenam REA de diversos formatos de arquivos, como: imagens, animações, arquivos de áudio, vídeos e outros.O capítulo é finalizado na Seção 2.3, com as considerações finais.

2.1 Definição de REA

Como descrito no Capítulo 1, REA podem ser definidos como materiais de ensino, aprendizagem e pesquisa, em qualquer meio de armazenamento, que estão amplamente disponíveis por meio de uma licença aberta que permite reuso, readequação e redistribuição sem restrições ou com restrições limitadas (Atkins et al., 2007; Gimenes et al., 2012; Kanwar et al., 2011). Em 2002 a UNESCO cunhou esse termo em um fórum internacional (Wiley, 2002), do qual vários países participaram. Embora muitos países já tenham adotado o uso de REA, isso ainda não é uma prática acadêmica convencional devido a alguns inibidores. Um inibidor é a dificuldade de encontrar REA na Web, de forma que atenda às necessidades dos usuários. Outra dificuldade refere-se ao fato de que,

quando encontrados materiais para o ensino e aprendizagem, muitas vezes eles não são academicamente úteis ou não têm um nível acadêmico aceitável (Abeywardena e Chan, 2013).

A ideia dos REA é tornar o conhecimento do mundo um bem público tendo a tecnologia em geral, e especificamente a Web, como uma fonte de acesso a esse conhecimento. Entende-se, assim, que os REA devem permitir seu compartilhamento, uso e reuso. O movimento de REA foi financiado inicialmente pela Fundação Hewllet que considerou o movimento uma filosofia de democratização do conhecimento por meio da Web (Smith e Casserly, 2006) (Wiley et al., 2014). Mike Smith, diretor do programa de educação da Fundação Hewllet disse: *"O coração dos recursos educacionais abertos é uma ideia simples e poderosa de transformar o conhecimento do mundo em um bem público e a Web é uma oportunidade extraordinária para que todos possam compartilhar, usar e reutilizar esse conhecimento"* (Smith e Casserly, 2006).

Existem diversas definições de REA, diferentes modelos de compartilhamento, diversos modelos de produção, como também muitos desafios para REA. Devido à importância do tema, em um documento recente (UNESCO, 2009) (Wiley et al., 2014), a Organização das Nações Unidas (ONU) definiu os principais problemas relativos ao desenvolvimento e ao uso de REA: i) *o problema da qualidade* de REA; ii) *o problema da descoberta* de como encontrar REA; iii) *o problema da sustentabilidade* de como financiá-los; iv) *o problema da localização e re-contextualização* de REA; e v) *o problema do remix, isto é, a dificuldade de identificar a granularidade de alteração do conteúdo por outras pessoas e o nível de alteração*. De acordo com o documento da ONU, estes problemas devem ser enfrentados de maneira que os REA possam cumprir seu potencial de contribuir para o desenvolvimento humano.

Apesar dos primeiros REA terem sido publicados em formatos textos ou documentos baseados em um formato texto, isso não significa que os REA possuem limitações sobre os tipos de mídias ou os tipos de arquivos a serem usados. Muitos REA modernos são liberados em diferentes formatos, como em imagens, clipes de filmes, animações, conjunto de dados e arquivos de áudios, dentre outros. Eles fornecem, portanto, um material multimídia rico para o uso e reutilização, os quais são disponibilizados por meio de grandes repositórios como Youtube (vídeos)[18], Flickr[19] (imagens) e iTunes (podcasts), sobre o regime de licenciamento *Creative Commons (CC)* (Abeywardena e Chan, 2013). O conteúdo dos REA também podem incluir cursos completos, partes de cursos, módulos, guias para estudantes, anotações, livros didáticos, artigos de pesquisa, instrumentos de avaliação,

[18] `http://www.youtube.com`
[19] `http://www.flickr.com`

recursos interativos como simulações e jogos de interpretação, aplicativos, dentre outros recursos.

Neste projeto foi feita uma pesquisa no Google Trends[20] para gerar uma série temporal sobre o uso do termo REA, como mostra a Figura 2.1. Nela são usados dois termos, *"Open Education Resources"* (termo em inglês para REA) representado pela linha vermelha e "Recursos Educacionais Abertos" representado pela linha azul. No eixo x é apresentado o interesse, usando-se uma escala de 20 à 100, no qual o valor 100 representa o interesse máximo das pesquisas feitas no mecanismo de busca do Google, enquanto que no eixo y é representado o tempo que foram feitas as buscas na escala de anos. É possível observar nesse gráfico que o termo em inglês é muito pesquisado desde 2007, enquanto o termo em português só veio a sobressair no ano de 2013.

Figura 2.1: Série temporal gerado no Google Trends mostrando a repercussão do termo REA (azul) em contrapartida do termo *Open Educational Resources* (vermelho).

Os REA têm um potencial para se tornar uma fonte importante de material didático e de pesquisa, especialmente para a educação superior, pois grandes organizações mundiais estão a favor de sua expansão em escala global, como a UNESCO, Comunidade de Aprendizagem (COL), Organização para a Cooperação e Desenvolvimento Econômico (OECD), e o Conselho Internacional de Educação à Distância (ICDE) (Abeywardena e Chan, 2013).

[20] http://www.google.com/trends

2.2 Padrões de Metadados específicos usados no See-OER

Na década de 1990, o reconhecimento da necessidade de reutilização de materiais educativos gerou o desenvolvimento de padrões de metadados para compartilhamento e armazenamento de objetos de aprendizagem. Os metadados são dados que descrevem um recurso físico ou eletrônico (McClelland, 2003). Eles podem ser usados para auxiliar o gerenciamento das coleções de documentos, imagens e outras informações em um repositório.

No desenvolvimento deste projeto foram usados padrões de metadados que facilitam a catalogação, pesquisa e reutilização de REA, a saber: o Dublin Core Metadata Element Set (DCMES) (seção 2.2.1), IEEE Learning Object Metadata (IEEE/LOM) (seção 2.2.2), Protocolo *Open Graph* (OGP) (seção 2.2.3), MathML e Vídeo Sitemaps (seção 2.2.4). Por fim, na subseção 6.1.1, é feito um levantamento dos metadados atualmente utilizados em repositórios REA.

2.2.1 Dublin Core Metadata Element Set (DCMES)

O padrão DCMES foi criado pela *Dublin Core Metadata Initiative* (DCMI) para facilitar a busca e a recuperação de REA. Sua versão atual, versão 1.1, inclui 15 elementos bem definidos para descrever as propriedades mais importantes de um recurso, que são: título, autor, assunto, descrição, editor, colaborador, data, tipo, formato, identificador, origem, idioma, relação, cobertura e direitos. A descrição desses elementos é apresentada na Tabela 2.1. Todos os elementos são recomendados pelo DCMI, mas nenhum deles é obrigatório, consequentemente, determinados recursos podem não apresentar recursos suficientes para sua localização.

O DCMES é utilizado em diversos repositório e também é chamado apenas de "Dublin Core", sendo o termo *Dublin* originário de um *workshop* realizado em 1995 em Dublin, Ohio, EUA; e *Core* por causa dos elementos que são amplos e genéricos, usados para descrever uma grande coleção de recursos (DCMI, 2012). Por exemplo, ele foi adotado como padrão para o compartilhamento de metadados pela Open Archives Initiative (McClelland, 2003), uma organização que promove padrões de interoperabilidade para facilitar a disseminação eficiente de conteúdos. Além disso, o DSpace, um dos *software* mais utilizados para construção de repositórios institucionais, também utiliza o DCMES como metadados padrão (Lousangfa et al., 2008). Outras organizações e instituições que usam esse padrão incluem a Biblioteca do Congresso e da Fundação Nacional de Ciência dos

Elemento	Descrição
Title	Título do material.
Creator	Criador do material - pode ser uma pessoa ou uma organização.
Subject	Assunto do material - pode se usar uma expressão que classifique o material, palavras-chaves ou tópicos.
Description	Descrição do material - um resumo.
Publisher	Editor - uma pessoa, organização ou entidade que auxiliou na publicação do recurso.
Contributor	Outro contribuinte - uma pessoa, organização ou entidade que auxiliou da construção do material.
Date	Data - data de publicação ou criação do material, definida no formato AAAA-MM-DD, o qual é um formato recomendado ISO 8601.
Type	Tipo de recurso - representa a natureza ou gênero do conteúdo do material. Tipos podem incluir termos descrevendo categorias, funções, gêneros, ou níveis de agregação para o conteúdo.
Format	Formato - o formato pode ser digital ou físico. No caso de digital, deve ser informado o MIME-Type, como audio/mp4. No caso de físico, deve-se informar os detalhes, como exemplo as dimensões do material.
Identifier	Identificador - uma chave primária, por exemplo ISBN (Internacional Standard Book Number) ou DOI (Digital Object Identifier).
Source	Fonte - uma referência de onde o recurso foi retirado.
Language	Idioma - duas letras referentes ao idioma mais duas letras opcionais para diferenciar países nos quais há diferenças de dialeto, de acordo com as recomendações ISO 339 e 3166. Por exemplo, pode-se citar en, en-uk, pt-br.
Relation	Relação - referência a um recurso relacionado.
Coverage	Cobertura - extensão ou alcance do recurso - por exemplo a localização espacial, um período de tempo, ou uma jurisdição.
Rights	Gestão de direitos - informação dos direitos do autor sobre o recurso.

Tabela 2.1: Os 15 elementos do DCMES (Borba, 2000).

Estados Unidos, o Instituto Nacional de Informática do Japão, a Biblioteca da Universidade de Helsinki, a Biblioteca Nacional de Austrália, da Alemanha e do Canadá, e a Comissão de Sistemas de Informação Conjunta do Reino Unido.

Na Figura 2.2 é ilustrado um modelo de informação abstrato do metadados DCMES. Um modelo de informação é independente de qualquer sintaxe de codificação em particular, e permite uma melhor compreensão dos tipos de descrições que estão sendo codificadas, além de facilitar o desenvolvimento de mapeamentos e traduções de diferentes metadados. A Figura 2.2 representa um diagrama baseado na representação clássica da UML, na qual uma seta fechada deve ser lida como "é" ou "é um"; e uma linha que começa com um *diamante* deve ser lida como "contém um" ou "tem a". Esse diagrama mostra que, para cada *recurso descrito*, é usado um ou mais pares de *propriedade-valor*. Uma *propriedade-valor* é composta por uma *propriedade* e um *valor*. Uma *propriedade* está associada a um literal ou uma entidade física, digital ou conceitual. Cada *valor* representa o valor de um determinado recurso, que pode ser um *valor literal* ou um *valor não literal*. Um *valor não literal* é o valor que possui uma entidade física, digital ou conceitual, enquanto que um *valor literal* é o valor na sua forma bruta, que não representa nenhuma entidade, mas apenas uma sequência de caracteres *Unicode* como uma forma lexical para denotar um recurso, juntamente com uma *tag* opcional do idioma.

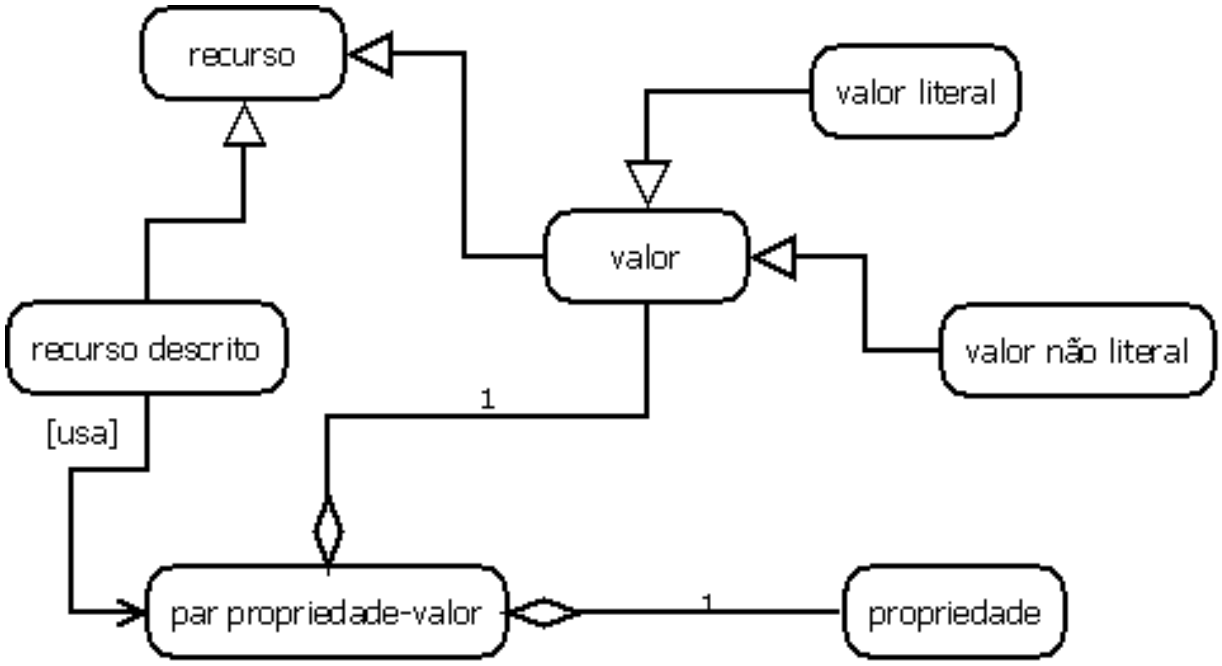


Figura 2.2: Modelo de informação abstrato do metadados DCMES

No caso do DCMES (ou DC), os repositórios REA podem representar seus metadados para o público com o formato variado, assim o formato como os metadados são instanciados não é sempre a mesma. Por exemplo, o Connexions instancia seus metadados por meio do formato XML, enquanto que o repositório Teses USP instancia seus metadados por meio do formato HTML/XHTML. Outros formatos incluem: HTML/XML, RDF/XML, XML e DC-DS-XML. Cada um desses formatos é descrito em mais detalhes a seguir.

O formato HTML/XHTML é recomendando pela DCMI. Ele utiliza somente duas *tags* para instanciar os elementos e os atributos: a *tag* <meta> para os elementos e a *tag* <link> para os atributos. A Figura 2.3 mostra a instanciação de metadados no formato HTML/XHTML sem utilizar nenhum esquema, enquanto a Figura 2.4 mostra um exemplo de instanciação de metadados no formato HTML/XHTML usando um esquema DC para instanciação dos elementos[21].

```
<head>
<meta name="description" content="Metadatas">
<meta name="keywords" content="HTML">
<meta name="author" content="W3C">
</head>
```

Figura 2.3: Metadados instanciados em HTML/XHTML sem utilizar nenhum esquema.

```
...
<link rel="schema.DC"
href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/">
<meta name="DC.language" xml:lang="pt" scheme="DCTERMS.RFC1766">
<meta name="DC.subject" content="Recuperação da informação" xml:lang="pt">
<meta name="DC.subject" content="Information recovery" xml:lang="en">
...
```

Figura 2.4: Metadados do repositório "Teses USP" instanciados no formato HTML/XHTML utilizando esquema da Dublin Core.

[21]Metadados da página `http://www.teses.usp.br/xml.php?id=tde-23042007-220548`

O formato RDF[22]/XML também é recomendado pelo DCMI. No caso do formato RDF, ele é declarado por meio de uma tripla composta por *propriedade (predicate), recurso (subject) e valor da propriedade (object)*. Nesse caso, as instâncias são escritas usando *tags* do XML, permitindo que as declarações sejam analisadas sintaticamente por um analisador XML e visualizadas por aplicativos que interpretam o XML (Ferreira, 2006). A Figura 2.5 mostra um exemplo que descreve semanticamente o título de uma página, validado no W3C Validator[23]. Nesse exemplo é ilustrado o modelo de dados representado pela tripla composição o documento em RDF/XML e uma representação gráfica do código RDF/XML.

Modelo de Dados (Tripla)

Recurso	Propriedade	Valor da Propriedade
http://www.usp.br/	http://purl.org/dc/elements/1.1/title	"Universidade de São Paulo"

Documento em RDF/XML

```
1: <?xml version="1.0"?>
2: <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3:   xmlns:dc="http://purl.org/dc/elements/1.1/">
4:   <rdf:Description rdf:about="http://www.usp.br/">
5:     <dc:title>Universidade de São Paulo</dc:title>
6:   </rdf:Description>
7: </rdf:RDF>
8:
```

Grafo do Modelo de Dados

Figura 2.5: Exemplo de modelo de dados, documento RDF/XML e o grafo do modelo de dados.

O terceiro formato que pode ser usado para a instanciação é o XML, no qual os metadados são descritos diretamente no arquivo XML e não mais em HTML/XHTML ou RDF/XML. O DCMI recomenda que os repositórios utilizem algum esquema XML que defina um *container* (recipiente) para os recursos, mas não especifica qual esquema XML deve ser usado para esse *container* (Powell e Johnston, 2003). Um *container* é descrito em XML como um grupo de elementos que representa o registro. Um exemplo de *container* XML é o esquema da OAI-DC[24], no qual é usado pelo Connexions. A Figura 2.6 exibe um exemplo do esquema da OAI-DC usado pelo Connexions, no qual o *conteiner* contém os

[22] *Resource Description Framework* - RDF não deve ser considerado uma linguagem, mas um modelo de dados para descrição de recursos na Web de forma semântica, através da adoção de metadados (Ferreira, 2006).

[23] `http://www.w3.org/RDF/Validator/`

[24] `http://www.openarchives.org/OAI/2.0/oai_dc.xsd`

seguintes elementos: <ListRecords>, <record>, <metadata>, <identifier>, <header>, <datestamp>, <oai-dc> e <dc>.

Além disso, o DCMI recomenda que os elementos tenham uma codificação com um nome qualificado, assim o elemento deve estar associado com o espaço de nomes do esquema declarado. Por exemplo: <dc:title>*Digital Humanities* 2.0 : *A Report on Knowledge*</dc:title> em vez de <dc:title value="*Digital Humanities* 2.0 : *A Report on Knowledge*" />. Além disso, quando existe mais de um valor para o mesmo elemento, é recomendado repetir o elemento em XML, em vez de utilizar algum marcador de separação. O DCMI também cita que muitos repositórios podem misturar o DCMES com propriedades tiradas de outros esquemas, pois é uma forma de estender o DCMES. Por exemplo, é possível misturar DCMES com o IMS, assim outros elementos como o elemento "tempo de aprendizagem" também podem ser incluídos.

```
...
<ListRecords>
<record>
<header>
<identifier>oai:cnx.org:m34246</identifier>
<datestamp>2010-06-08T14:56:21Z</datestamp>
</header>
<metadata>
...
<oai-dc:dc ...>
// 15 Elementos do DCMES
<dc:title>Digital Humanities 2.0 : A Report on Knowledge</dc:title>
<dc:creator>Todd Presner</dc:creator>
<dc:subject>Digital</dc:subject>
...
</oai-dc:dc>
</metadata>
</record>
```

Figura 2.6: Metadados Dublin Core em XML do Connexions usando o *container* do esquema da OAI-DC

A última forma de instanciar o DCMES é por meio do formato DC-DS-XML (Dublin Core - *Description Set* - XML). Com o DC-DS-XML é possível utilizar o GRDDL (*Gleaning Resource Descriptions from Dialects of Languages*) recomendado pela W3C

[24] `http://cnx.org/content/OAI?verb=SearchRecords&metadataPrefix=oai-dc&query:list=digital&b-start:int=10&b-size=10`

para descrever um conjunto de convenções para associar com um documento XML, juntamente com um algoritmo para extração de descrições de dialetos de vários idiomas de um conjunto de dados RDF.

2.2.2 IEEE Learning Object Metadata (IEEE/LOM)

O padrão IEEE/LOM foi desenvolvido em um esforço conjunto do Comitê de Padrões de Tecnologia de Aprendizagem IEEE em colaboração com DCMI e outras organizações (McClelland, 2003). A principal diferença do IEEE/LOM em relação ao DCMES é que o IEEE/LOM tem uma abordagem hierárquica para a criação de metadados (McClelland, 2003). Ele também é usado como referência para criação de outros metadados, como é o caso do *Canadian Core Learning Resource Metadata Application Profile* (CanCore), UK *Learning Object Metadata Core* e o *Sharable Content Object Reference Model* (SCORM) (Koutsomitropoulos et al., 2010). Seus elementos fornecem um meio de desenvolver descrições mais detalhadas de REA. A versão 1.0 do IEEE/LOM organiza 60 elementos em 9 categorias: geral, ciclo de vida, meta-metadados, técnica, educacional, direitos, relação, anotação e classificação. Por exemplo, para a categoria geral, definem-se os atributos identificador, título, linguagem, descrição, palavra-chave, cobertura (ou seja, o tempo, a cultura, a geografia ou região a que se aplica), estrutura e nível de agregação.

A *Instructional Management System* (IMS) continuou os trabalhos da IEEE/LOM (Koutsomitropoulos et al., 2010) e criou seu próprio padrão, o qual é chamado IMS Learning Resource Metadata. Esse padrão é equivalente ao IEEE/LOM, sendo que a principal diferença refere-se à taxonomia utilizada nos metadados (Gimenes et al., 2012). Por exemplo, pode-se citar o *classification.purpose* que representa o objetivo educacional do recurso, que é um elemento da taxonomia criada pela IMS.

O esquema conceitual do IMS ou IEEE/LOM é hierárquico e possui 3 níveis. Na Figura 2.7 é uma amostra desses níveis, começando da esquerda para direita, do nível superior para o inferior (Phil Barker, 2006).

Na Figura 2.8 são mostrado os elementos e a estrutura do IMS por meio de um diagrama. Como é possível observar a estrutura e organização desse padrão de metadados é diferente da maioria dos padrões de metadados visto até este momento e diferente dos padrões OGP e Vídeo Sitemaps. Os 3 níveis de hierarquia presente no IMS não é possível observar na Figura 2.8 devido a modelagem(Phil Barker, 2006). Porém, é mostrado os elementos associadas aos tipos de elementos, tais como “LangString” e “DateTime” .

A quantidade de metadados presente no IMS é para facilitar a interoperabilidade e o compartilhamento de recursos educacionais. Porém, depende dos fornecedores de

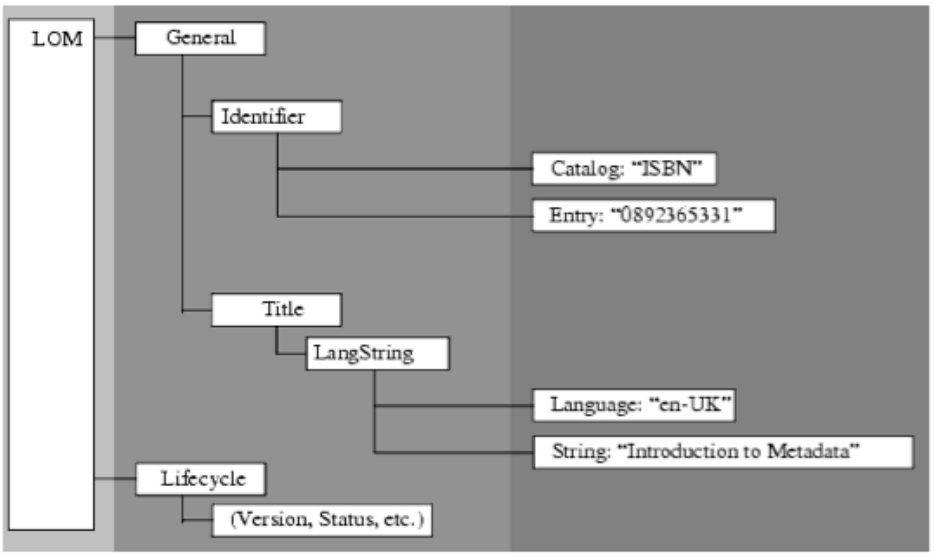


Figura 2.7: Esquema conceitual em 3 níveis do IMS (Phil Barker, 2006).

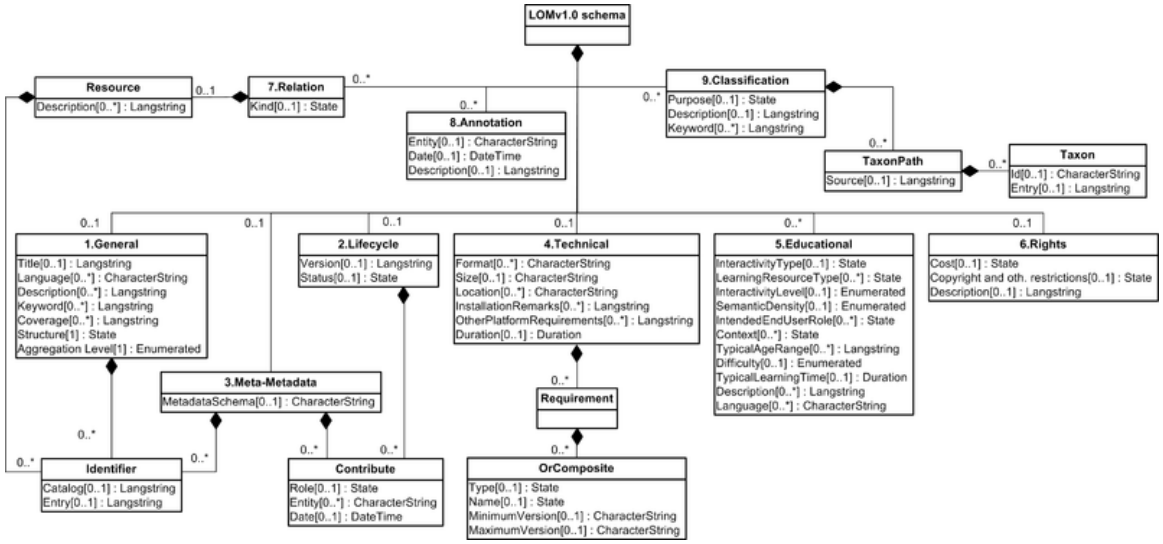


Figura 2.8: Modelagem da organização dos metadados do padrão IMS (Phil Barker, 2006).

recursos educacionais e desenvolvedores estarem familiarizados com tais taxonomias para que isso de fato ocorra. Esse metadados estão de acordos com diversas normas, como por exemplo a ISO 2788 (Phil Barker, 2006). Contudo, nem todos esses metadados podem ser vistos como "melhores práticas" pois vão depender de que área e especialidade que estão se tratado, é possível observar que em alguns casos muitos desses metadados serão preenchidos, em outros casos poucos serão preenchidos.

2.2.3 Protocolo *Open Graph* (OGP)

O Protocolo *Open Graph* (OGP) foi criado pelo Facebook. Ele foi inspirado no Dublin Core, Microformatos e RDFa (RDF em HTML com adição de atributos) (Facebook, 2013b). O OGP requer apenas 4 informações essenciais: o título, o tipo e um URL (Graham, 2012). Porém, ele possui muitos outros metadados opcionais como metadados para localização (latitude, longitude) e para vídeos (resolução e o tipo), dentre outros (Facebook, 2013a).

Na Tabela 2.2 mostra os elementos essenciais do OGP e que aparecem em diversos repositórios REA. O atributo **título** (*og:title*) representa o título do REA a ser compartilhado. O atributo **tipo** (*og:type*) é o tipo do objeto e não exatamente seu formato de arquivo. Por exemplo, se for um artigo será do tipo *article*. O atributo **localidade** (*og:locale*) representa a procedência do REA, com o idioma e o território. Por exemplo, português e do Brasil é pt_BR. O atributo **perfil** de quem está compartilhado o REA (og:profile_id) é um identificador único usado no Facebook *(Facebook ID)* o qual o usuário pode ser seguido futuramente. Também considera-se um dado de procedência do usuário.

Atributo	Tipo de dado
fb:profile_id	Inteiro
og:description	String
og:locale	String
og:site_name	String
og:url	String
og:title	String
og:type	String

Tabela 2.2: Elementos essenciais do OGP

O OGP representa uma forma de formalizar os metadados usados pelo Facebook, principalmente para a inclusão de elementos na "linha do tempo" criado pela empresa. Esses metadados consistem em um meio de facilitar a comunicação entre a rede social do Facebook com as páginas e aplicativos criados fora do seu ambiente. Vários repositórios REA o utilizam, como o *Connexions* e o *Khan Academy*.

Além disso, o tipo de objeto *article*, muito usado para compartilhar os REA, também é descrito por meios dos elementos que constam na Tabela 2.3. No contexto de REA é usado o atributo **autor** (*article:author*) para representar um vetor de URL(s) do perfil do autor ou autores e também é possível usar os IDs dos Facebook. O atributo de **seção**

(*article:section*) é usado para referenciar qual contexto o artigo pertence. Por exemplo, Letras é instanciado por meio <*meta property="article:section" content="Letras"*>.

Atributo	Tipo de dado
article:author	Vetor de String
article:expiration	Data e Hora
article:modified_time	Data e Hora
article:published_time	Data e Hora
article:publisher	String
article:section	String
article:tag	Vetor de String

Tabela 2.3: Elementos de um artigo (*article*) no OGP.

2.2.4 Vídeo Sitemaps

Vídeo sitemaps representa uma forma de se incluir metadados em vídeos e áudios. Ele é um arquivo XML que lista os URLs de um site junto com metadados adicionais sobre cada URL (Sitemaps, 2008). No entanto, esses metadados não são suficientes para descrever vídeos e áudios.

Como resultado, o Google estendeu os atributos tradicionais do Sitemap e adicionou novos metadados para descrição de vídeos e áudios. Dentre esses metadados, os obrigatórios são o URL da página do vídeo (*tag loc*), um URL que aponta para um arquivo de imagem miniatura do vídeo (*thumbnail−loc*), o título do vídeo, a descrição do vídeo, um URL que aponta para o arquivo verdadeiro de mídia do vídeo e um URL que aponta para um *player* de vídeo. Além disso, existem outros metadados opcionais como palavras-chaves para o vídeo, categoria e preço para o *download* do vídeo.

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:video="http://www.google.com/schemas/sitemap-video/1.1">
   <url>
     <loc>http://www.example.com/videos/some_video_landing_page.html</loc>
     <video:video>
       <video:thumbnail_loc>http://www.example.com/thumbs/123.jpg</video:thumbnail_loc>
       <video:title>Grilling steaks for summer</video:title>
       <video:description>Alkis shows you how to get perfectly done steaks every
         time</video:description>
       <video:content_loc>http://www.example.com/video123.flv</video:content_loc>
       <video:player_loc allow_embed="yes" autoplay="ap=1">
         http://www.example.com/videoplayer.swf?video=123</video:player_loc>
       <video:duration>600</video:duration>
       <video:expiration_date>2009-11-05T19:20:30+08:00</video:expiration_date>
       <video:rating>4.2</video:rating>
       <video:view_count>12345</video:view_count>
       <video:publication_date>2007-11-05T19:20:30+08:00</video:publication_date>
       <video:family_friendly>yes</video:family_friendly>
       <video:restriction relationship="allow">IE GB US CA</video:restriction>
       <video:gallery_loc title="Cooking Videos">http://cooking.example.com</video:gallery_loc>
       <video:price currency="EUR">1.99</video:price>
       <video:requires_subscription>yes</video:requires_subscription>
       <video:uploader info="http://www.example.com/users/grillymcgrillerson">GrillyMcGrillerson
         </video:uploader>
       <video:live>no</video:live>
     </video:video>
   </url>
</urlset>
```

Figura 2.9: Exemplo do *vídeo sitemaps* (Google, 2011).

Para o uso dos metadados, o Google criou um esquema[25] XML que define esses novos elementos e atributos (Google, 2011). Esses metadados são utilizados por alguns repositórios, como o e-Aulas USP[26].

A Figura 2.9 mostra um exemplo do uso do **vídeo sitemaps**. Cada arquivo do vídeo sitemaps deve possuir no máximo 50.000 entradas (Google, 2011). Além disso, o vídeo sitemaps não suporta aninhamento de arquivos de vídeos.

2.3 Considerações Finais

Neste capítulo foram discutidos conceitos básicos relacionados a REA. Foram descritos a definição do termo REA, o seu potencial para se tornar uma importante fonte de recursos livres e reutilizáveis para o ensino e aprendizagem, as tendências sobre o uso do termo no contexto brasileiro. Também foram detalhados os principais metadados utilizados para instanciar os REA em repositórios e plataformas, a saber DCMES, IEEE/LOM, OGP e o Vídeo Sitemaps. Cada esquema de metadados possui uma forma de instanciação e

[25] `www.google.com/schemas/sitemap-video/1.1/sitemap-video.xsd`
[26] `eaulas.usp.br/portal/VMSResources/sitemaps/sitemaps1.xml`

diferentes atributos para descrever o metadados, os quais podem ser similares ou não. Os esquemas que descrevem os metadados e a forma de instanciação são importantes para o projeto de mestrado, visto que eles serão usados no mecanismo de busca a ser proposto.

No próximo capítulo será feita uma descrição de conceitos básicos relacionados à recuperação de informação, especificamente por recuperação de informação na Web. Esses conceitos serão fundamentais para o entendimento de um mecanismo de busca na Web por REA, que foi desenvolvido neste projeto.

Capítulo

3

Recuperação de Informação

A Recuperação da Informação (RI) é uma área de pesquisa muito ampla. Na Ciência da Computação, seu foco é tornar à informação mais acessível aos interessados (Baeza-Yates e Ribeiro-Neto, 2011). De acordo com Salton e Harman (2003), RI trata várias questões como a estrutura, a representação, a organização, o armazenamento, a busca e a recuperação de informações. Com relação aos mecanismos de busca, esses autores consideram que eles sejam implementações de RI, as quais utilizam diversas técnicas de buscas.

Em termos de pesquisa, os estudos de RI podem ser divididos em duas grandes áreas distintas e complementares. A primeira é o computador como objeto central e a segunda é o homem como elemento principal. No computador como objeto central, destacam-se pesquisas relacionadas à construções de índices para RI, ao processamento de consultas e ao desenvolvimento de algoritmos de ordenação para os resultados de pesquisa, dentre outros tópicos. No homem como elemento principal, destacam-se o estudo do comportamento do usuário na busca, o conhecimento do usuário antes da pesquisa e a organização dos resultados dos mecanismos de busca (Baeza-Yates e Ribeiro-Neto, 2011).

Esta pesquisa de mestrado concentra-se na primeira abordagem - o computador é objeto central, pois ele vislumbra a proposta de um mecanismo de busca na Web por REA. Na proposta desse mecanismo, os metadados de REA serão considerados. Em geral, esses metadados estão disponíveis na Web em arquivos XML ou páginas XHTML, podendo estar instanciados junto com dados estruturados, não estruturados ou semiestruturados.

Este capítulo está estruturado da seguinte forma. Na Seção 3.1 são detalhados os conceitos de dados estruturados, semiestruturados e não estruturados. Em seguida, na Seção 3.2, é descrita a RI na Web, com destaque para a descrição do *Web crawler* e a

arquitetura de um mecanismo de busca na Web. O capítulo é finalizado na Seção 3.3, com as considerações finais.

3.1 Dados Estruturados, Semiestruturados e Não Estruturados

Os **dados estruturados** possuem um formato fixo e rigoroso definido por meio de um esquema projetado para eles (Elmasri e Navathe, 2011). Se no esquema foi definido que um atributo deve ser do domínio dos $\mathbb{Z}$'s, então todos os dados para esse atributo devem, rigorosamente, pertencer ao conjunto dos $\mathbb{Z}$'s.

Dados não estruturados	Dados semiestruturados	Dados estruturados
Exemplo: textos livres	Exemplo: Páginas HTML	Exemplo: Banco de dados relacionais

Tabela 3.1: Tabela das estruturas de dados

Os **dados semiestruturados** são de caráter intermediário em que possuem pelo menos alguma estrutura. As páginas HTML e XML são exemplos de dados semiestrurados. Os dados semiestruturados possuem uma representação mais flexível e mais adaptativa. E, difere dos dados estruturados pois não possuem restrições rígidas impostas como nos bancos de dados relacionais.

Também existem os **dados não estruturados**, os quais não possuem nenhum esquema ou possuem apenas uma indicação muito limitada sobre o tipo de dados que está sendo usado. (Elmasri e Navathe, 2011). Os textos livres são exemplos de dados não estruturados.

A Tabela 3.1 resume os dados estruturados, semiestruturados e não estruturados. Nesta dissertação são considerados, na RI, os dados semiestruturados, que são os metadados, e os dados não estruturados. No caso dos dados não estruturados, serão considerados os textos livre. Para os vídeos, imagens e áudios, serão considerados apenas os metadados associados a eles.

3.2 Recuperação de Informação na Web

A Web foi criada conceitualmente em 1989, por Tim Berners-Lee, no CERN[27] na Suíça. Em 1991, foi lançado o primeiro servidor Web, chamado de *World Wide Web*,

[27]Organização Europeia para a Pesquisa Nuclear

mas referenciado pela maioria dos livros apenas como Web (Baeza-Yates e Ribeiro-Neto, 2011). A Web se expandiu rapidamente, e hoje conta com mais de 46 bilhões[28] de páginas da Web indexadas[29].

Para encontrar informações neste gigantesco conjunto de páginas Web, é necessário o uso de um mecanismo de busca na Web. Os mecanismos de busca na Web se tornaram bastante populares e continuam sendo alvos de estudos (Baeza-Yates e Ribeiro-Neto, 2011). Muitos mecanismos de busca são construídos nos Estados Unidos e com foco em documentos no idioma inglês. Porém, outros mecanismos de busca são desenvolvidos para alguns idiomas específicos no qual seu alfabeto é diferente do inglês, como o caso do idioma chinês, russo e árabe. Por exemplo, alguns mecanismos de busca na Web, específicos para outros idiomas, incluem o Baidu[30] da China, Yandex[31] da Rússia e o Naver[32] da Coreia do Sul.

Existem vários desafios enfrentados pelos mecanismos de busca na Web, como a distribuição dos dados, a grande porcentagem de dados voláteis, o grande volume de dados, os dados redundantes e não estruturados, a qualidade dos dados (existem muitas informações publicadas sem acurácia, obsoletas, escritas de forma errada, etc) e a heterogeneidade dos dados (vários tipos de mídias, formatos específicos, uma variedade de idiomas e alfabetos) (Baeza-Yates e Ribeiro-Neto, 2011). Além desses desafios, existem outros desafios relacionados à interação homem-máquina. Uma delas é como o usuário expressa a sua consulta, por exemplo, por meio de palavras chaves ou uma frase. Uma consulta expressa exatamente como falada naturalmente pode resultar em páginas que não reflitam a necessidade de informação do usuário. Um outro desafio refere-se à interpretação dos resultados por parte do usuário, ou seja, como o mecanismo de busca na Web deve apresentar seus resultados de modo que facilite o usuário a encontrar rapidamente o que ele está procurando.

Como pode ser observado, existem vários desafios para construção de um mecanismo de busca na Web. Na Seção 3.2.1 é descrito o conceito de *Web Crawlers*, que é um ponto chave de um mecanismo de busca na Web, enquanto que na Seção 3.2.2 é detalhado a arquitetura de um mecanismo de busca na Web, a qual serve de base para o desenvolvimento do presente projeto.

[28] `http://www.worldwidewebsize.com/` - informações de 2015

[29] Foi considerado o indexador do Google, pois possui uma quantidade elevada de páginas da Web indexadas.

[30] `http://www.baidu.cn`

[31] `http://www.yandex.ru`

[32] `http://www.naver.com`

3.2.1 *Web Crawler*

Um *Crawler* para Web (também conhecido por *"Web Crawler"*, *"Web Spider"*, *"Web Robot"*, ou simplesmente "bot") é uma aplicação que faz o *download* das páginas da Web de forma automática (Baeza-Yates e Ribeiro-Neto, 2011).

Em 1993, foi criado o primeiro mecanismo de busca na Web por Martijn Koster. Chamado de "ALIWEB" *(Archie-Like Index of the Web)*, esse mecanismo exigia dos sites um índice de todas as páginas locais que o site possuía e quisesse que o "ALIWEB" pesquisasse. Desde que a maioria dos sites não publicavam esse índice, o "ALIWEB" enfrentou dificuldades para mostrar a sua eficiência e eficácia. Então, nesse mesmo ano, foi criado um algoritmo para percorrer a Web coletando URLs dos sites e fazendo o *download* das páginas, para posteriormente construir um índice de busca para essas páginas. Esse algoritmo foi chamado de WWWW (*World Wide Web Wanderer*) e constituiu o primeiro *Crawler* para Web.

Em grande parte, o sucesso de um mecanismo de busca na Web está relacionado à eficiência do *Crawler* que ele possui. Além disso, o *Crawler* dentro de um mecanismo de busca na Web, é um componente dependente do tipo de mecanismo que está sendo construído (Baeza-Yates e Ribeiro-Neto, 2011). Existem *Crawlers* para mecanismo de busca na Web genéricos e para mecanismo de busca na Web vertical, dentre outros tipos. No caso de um mecanismo de busca na Web genérico, o *Crawler* precisa enfocar no balanceamento da cobertura e na qualidade das páginas. A cobertura refere-se à quantidade de páginas que são analisadas pelo *Crawler* para que possa responder às diferentes consultas no mecanismo. Já a qualidade, refere-se à qualidade das páginas que foram analisadas, as quais devem ser de altíssima qualidade devido a quantidade de páginas que devem ser tratadas.

Com relação ao *Crawler* para mecanismo de busca na Web vertical, chamado de *Crawler* vertical, seu foco é um conjunto particular da Web e não toda a Web. Esse subconjunto de páginas pode ser definido, por exemplo, pela geografia, pelo idioma e pelo tópico. Um *Crawler* vertical é usado para agregar dados de diferentes fontes, usualmente fontes similares. É muito comum usar o *Crawler* vertical para o comércio eletrônico, pois sua função principal é fazer o *download* das informações de diferentes catálogos de sites de comércio eletrônico (Baeza-Yates e Ribeiro-Neto, 2011).

Outro exemplo de *Crawler* é o *"feed Crawler"*, que é usado por agregadores de notícias baseados na Web. Ele periodicamente analisa um conjunto de sites na Web pré-especificados em busca de atualizações em RSS/RDF.

O *Crawler* desenvolvido nesta dissertação teve enfoque no *Crawler* para mecanismo de busca na Web por REA com balanceamento de cobertura como descrito na Seção 6.3.

Os tipos de *Crawlers* podem ser classificados por meio de uma taxonomia, como mostra a Figura 3.1. Nessa taxonomia, existem 3 eixos principais:

- Novidade: em alguns casos, as informações sofrem mudanças constantes, e o *Crawler* deve obter novas atualizações sempre que possível. Em outros casos, essa necessidade não existe, pois a informação pode ser antiga.
- Qualidade: alguns *Crawlers* tem objetivos particulares, como uma porção de alta qualidade da Web, enquanto outros visam a abrangência a ser coberta e não especificamente a qualidade das páginas.
- Volume: alguns *Crawlers* tem interesse em uma grande fração da internet, enquanto outros sacrificam a amplitude da internet para enfocar a qualidade e/ou a novidade.

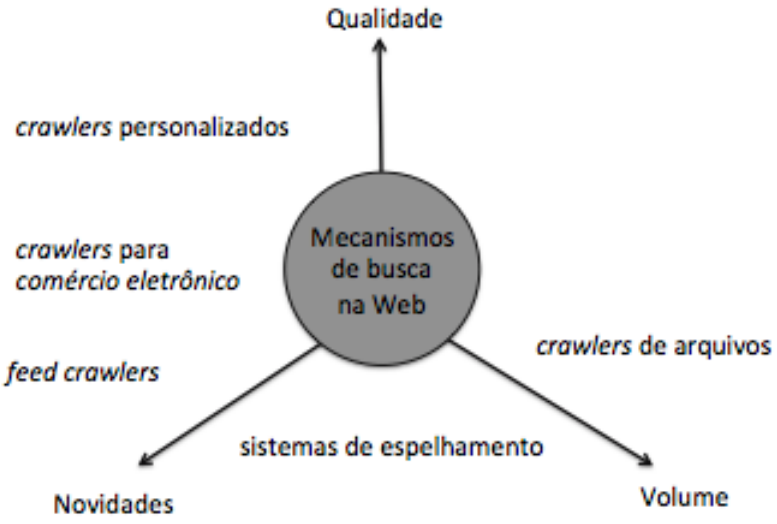


Figura 3.1: Tipos de Crawler (adaptado de Baeza-Yates e Ribeiro-Neto (2011)).

Baeza-Yates e Ribeiro-Neto (2011) também citam a importância do tempo que um *Crawler* deve retornar à página para ver se houve mudanças. É possível ter uma estimativa de mudanças da página para que o *Crawler* tome novas decisões no futuro, baseando-se nas alterações anteriores de uma página. Para isso, existe um parâmetro λ_p que corresponde à taxa de variação média de uma página p. Assuma que N_p é a quantidade de vezes que uma página p foi visitada, X_p é a quantidade de visitas que observaram mudanças na página, S_p é o tempo que passou desde a primeira visita na página e T_p é um acumulador do

tempo total de mudanças feitas na página p. Então, é possível estimar a taxa de variação média pela fórmula:

$$\lambda_p \approx \frac{(X_p - 1) - \frac{X_p}{N_p log(1 - X_p/N_p)}}{T_P} \approx \frac{X_p}{S_p}$$

A taxa de variação média pode ser estendida para que se possa ter uma visualização mais ampla das mudanças de uma página. Para isso, utiliza-se uma integral com limites inferior e superior em relação ao tempo, que integra a função Ano, composta pela taxa de variação média (λ_p) e a quantidade de tempo que o *Crawler*(t) rastreou a página p.

$$Anos(\lambda_p, t) = \int_0^t \lambda_p e^{\lambda_p x}(t - x)dx$$

Por fim, o *Crawler* pode ser considerado um assunto ainda bastante pesquisado. Pois, a Web está em constante crescimento do volume de dados. Além disso, existem pesquisas voltadas a acessibilidade, vunerabilidade e propostas para reduzir o tempo e o custo do rastreamento na Web.

3.2.2 A Arquitetura de um Mecanismo de Busca na Web

A Web cria novos desafios para a RI, dentre os quais destacam-se (Sanderson e Croft, 2012): (i) o rastreamento deve ser rápido para reunir os documentos na Web e mantê-los atualizados; (ii) o espaço de armazenamento deve ser utilizado de forma eficiente para armazenar os índices e, opcionalmente, os próprios documentos; (iii) o sistema de indexação deve processar centenas de dados de forma eficiente; (iv) as consultas devem ser tratadas rapidamente, a uma taxa de centenas de milhares por segundo; e (v) os documentos recuperados devem ser de alta precisão, mesmo que o número de documentos relevantes seja muito grande.

De acordo com Baeza-Yates e Ribeiro-Neto (2011), mecanismos de busca na Web devem realizar todo o processamento da consulta e o *ranking* sem acessar a fonte dos documentos. Dessa forma, evita-se o acesso a páginas remotas por meio da rede no momento da consulta, o que seria muito lento. Isso impacta diretamente na indexação e nos algoritmos de busca, bem como na complexidade das linguagens de consulta.

Não existe uma arquitetura universal que defina um mecanismo de busca na Web, porém existem semelhanças entre os vários mecanismos de busca existentes. Duas funções principais que convergem entre eles são o **processo de indexação** e o de **consulta**. No processo de indexação é construída uma estrutura de dados para a busca, enquanto no

processo de consulta é construída uma lista ordenada dos documentos, usando a consulta de busca feita pelo usuário e a estrutura de dados indexada.

Os principais componentes do processo de indexação são mostrados na Figura 3.2. O componente ***Web Crawler***, ele foi descrito na Seção 3.2.1. O componente de **transformação do texto** converte esses dados para que fiquem em um formato de um índice de termos. Esse componente pode possuir diversas funções, como: i) analisador léxico; ii) remoção da *stopping words*; iii) *stemming*; iv) extração e análise de links; v) extrator de informação; e vi) classificador. O componente de **construção do índice** tem a tarefa de criar o índice ou uma estrutura de dados semelhante para que possa futuramente fazer uma busca otimizada. De acordo com Baeza-Yates e Ribeiro-Neto (2011), os índices invertidos são comumente usados principalmente se tratando de mecanismos de busca na Web.

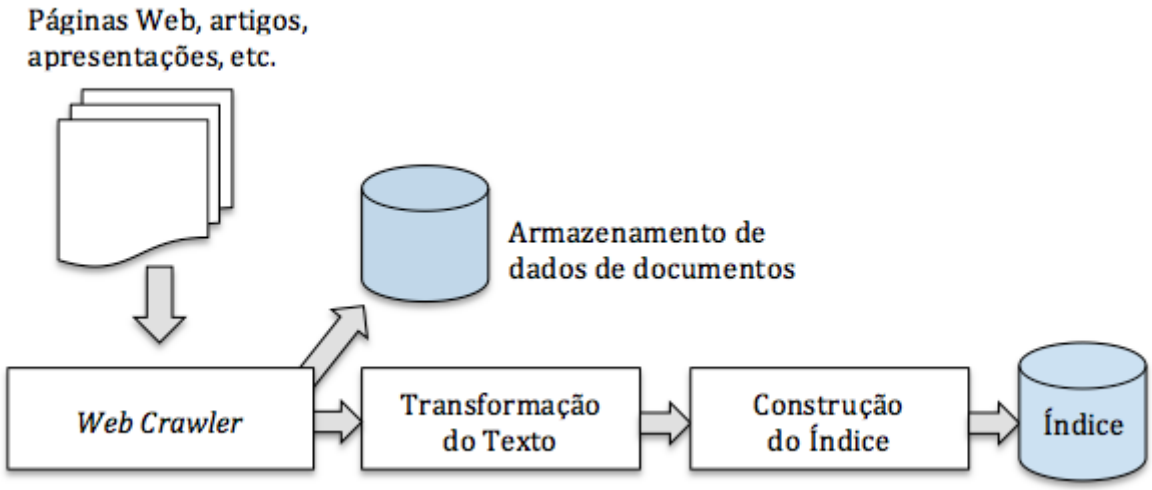


Figura 3.2: Processo de indexação (adaptado de Croft et al. (2011)).

O processo de consulta é ilustrado na Figura 3.3. Seus principais componentes são a interface de consulta do usuário, o *ranking* e avaliação. A **interface de consulta** do usuário fornece um meio de interação entre o usuário que faz a consulta e o mecanismo de busca. Esse componente possui várias funções, como transformar a consulta do usuário em termos de índice e exibir o *ranking* dos documentos. Isso inclui, por exemplo, gerar resumos dos documentos resultantes ou retirar fragmentos desses documentos. Além disso, esse componente também inclui uma variedade de técnicas para refinar a consulta para que melhor represente a necessidade de informação.

O componente de ***ranking*** (classificação) é importante para o mecanismo de busca, pois ele gera um *ranking* dos documentos usando uma pontuação baseada em um modelo

de recuperação. O *ranking* deve ser **eficiente**, uma vez que muitas consultas podem ser feitas e, consequentemente, o *ranking* calculado. O *ranking* também deve ser **eficaz**, visto que a qualidade da classificação determina se o mecanismo de busca alcançou o objetivo de encontrar a informação relevante. Croft et al. (2011) entendem que a eficiência do *ranking* depende dos índices e a efetividade depende do modelo de recuperação.

Para avaliar e monitorar a efetividade e eficiência do mecanismo é necessário um componente, que é o de **avaliação**. Esse componente deve gravar e analisar o comportamento do usuário com dados de log, para que se possa ter informações para ajustar o *ranking*. No entanto, essa captura de informação deve ser informada e aceita pelo usuário antes dele usar o mecanismo de busca, para preservar a privacidade e os direitos dos usuários.

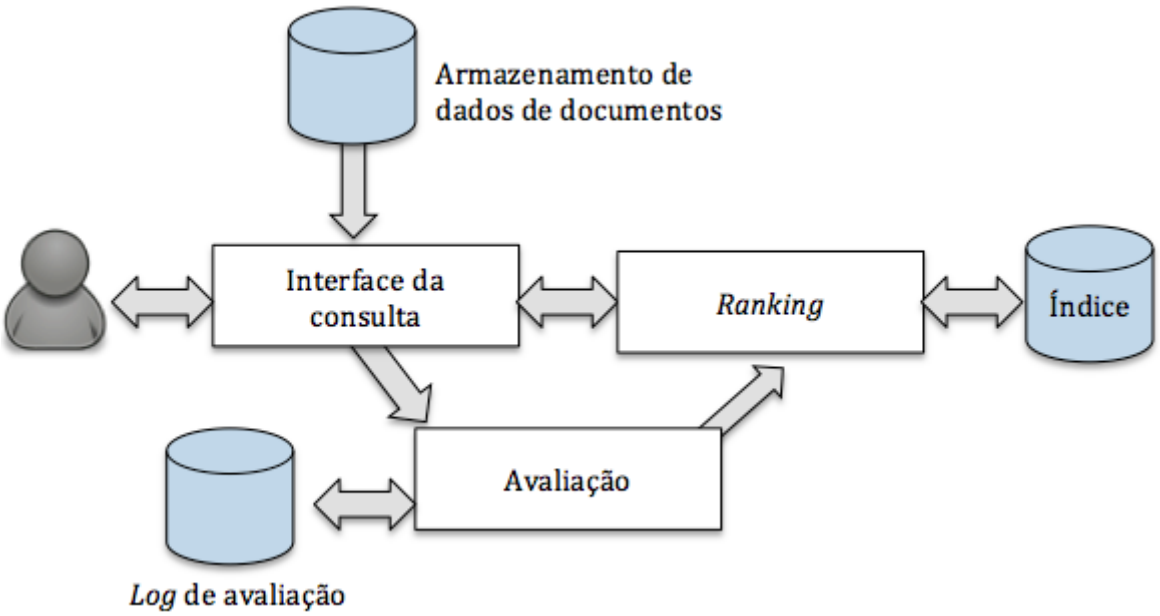


Figura 3.3: Processo de consulta (adaptado de Croft et al. (2011)).

Por fim, na Figura 3.4 é ilustrado a arquitetura geral da execução de um mecanismo de busca na Web, considerando que os processos de consulta e de indexação já foram descritos. O usuário realiza uma consulta de acordo com a sua necessidade de informação. Na sequência, o mecanismo de busca retornará a resposta à consulta, a qual deve ser refinada caso a resposta não seja condizente com a necessidade de informação.

Na Figura 3.4 também é ilustrado um exemplo de consulta, na qual a atividade do usuário é encontrar uma maneira para complementar o aprendizado adquirido nas aulas de cálculo. Usando essa atividade como base, o usuário formula a sua necessidade de informação que refere-se à complementação das aulas de cálculo por meio de exercícios.

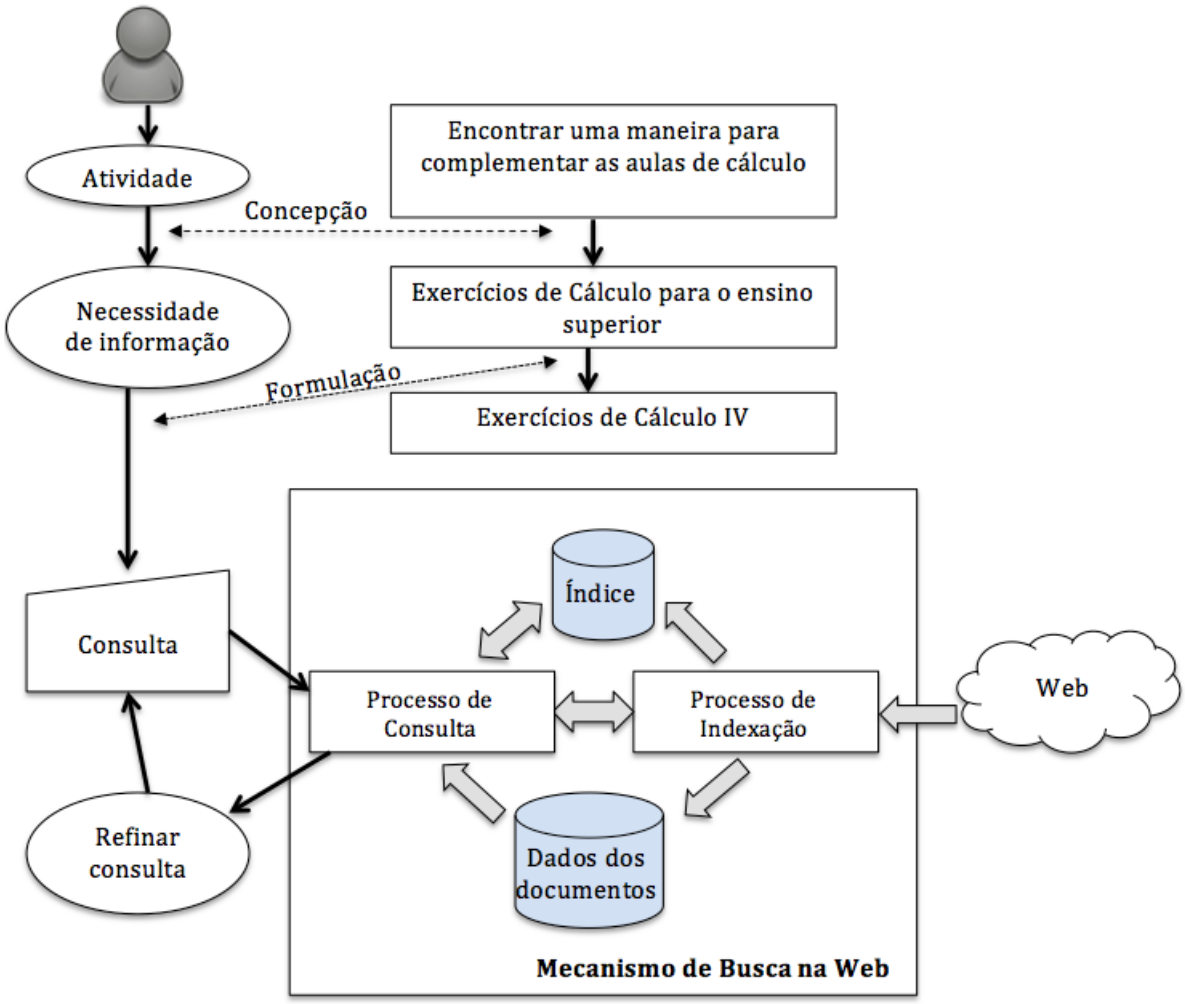


Figura 3.4: Visão geral de um mecanismo de busca genérico na Web.

Assim, o usuário escreve sua consulta representando sua necessidade de informação, que são exercícios de cálculo IV, que representam ser do ensino superior.

3.3 Considerações Finais

Neste capítulo foram descritos conceitos relacionados à RI dentro do contexto desta dissertação. Inicialmente foi feita uma breve introdução de RI. Em seguida foram descritos dados estruturados, semiestruturados e não estruturados. Após isso, foi abordada a recuperação de informação na Web, tratando-se de *Web Crawler*, que é o componente chave de um mecanismo de busca na Web. Dentro do contexto de RI na Web, foi detalhada uma arquitetura teórica de um mecanismo de busca.

O presente capítulo teve como objetivo introduzir o embasamento teórico relacionado à proposta de um mecanismo de busca na Web por REA, que consiste no objetivo deste trabalho. No próximo capítulo o embasamento teórico será referente a aspectos de integração e procedência dos dados.

Capítulo

4

Procedência e Integração de dados

Um problema enfrentado por aplicações que acessam várias fontes de dados autônomas e heterogêneas, como os repositórios REA, é a integração dos dados. Uma técnica que pode auxiliar na resolução desse problema é a procedência dos dados. Dentro deste contexto, este capítulo está organizado da seguinte maneira. Na Seção 4.1 são descritos os principais conceitos relacionados à procedência, enquanto que na Seção 4.2 são descritos conceitos relacionados à integração de esquemas e de instâncias em aplicações heterogêneas. O capítulo é finalizado na Seção 4.3, com as considerações finais.

4.1 Procedência dos Dados

A procedência dos dados possibilita identificar as fontes de dados e os processos de transformação aplicados aos dados, basicamente é um conjunto de metadados que pode estar em seu estado de criação ou já ter passado por algum processo de transformação (Benjelloun et al., 2008; Buneman et al., 2000; Glavic e Ditt, 2007; Munroe et al., 2006). São várias as motivações do ponto de vista de integração para se armazenar a procedência dos dados, dentre as quais destacam-se (Demsky, 2011; Freire et al., 2008; Ikeda et al., 2012; Tan, 2004): verificar o histórico dos dados, assegurar a qualidade dos dados integrados inferindo, por exemplo, que dados obtidos de fontes confiáveis têm maior probabilidade de serem corretos, realizar processos de auditoria dos dados e de atribuição de autoria aos proprietários dos dados, retificar fontes de dados que estão incorretas, e reproduzir decisões de integração em situações nas quais as fontes de dados não podem ser alteradas devido a direitos autorais.

Quatro aspectos devem ser considerados para o desenvolvimento de modelos de procedência dos dados. O primeiro deles considera "quais dados de procedência armazenar", e inclui a definição de quais tipos de dados de procedência devem ser armazenados e qual a granularidade desses dados. Com relação aos tipos de dados, existem várias classificações na literatura, as quais referem-se ao armazenamento de informações sobre as fontes de dados e sobre os processos de transformação pelos quais os dados foram submetidos (Buneman et al., 2001; Del Rio e Silva, 2007; Widom, 2005; Zhao et al., 2006). Com relação à granularidade, os dados sobre a procedência podem ser coletados em diversos níveis de detalhamento. Em um banco de dados relacional, por exemplo, eles podem ser armazenados em nível de tabela (maior granularidade), tupla (média granularidade) ou atributo (menor granularidade) (Glavic e Ditt, 2007). Além disso, os dados sobre procedência podem ser vistos como resultados de operações, as quais podem ter diferentes tipos de dados e granularidades.

Após a definição de quais dados armazenar, deve-se estabelecer estratégias para enfocar os três outros aspectos. O aspecto "como coletar os dados de procedência" indica se deve haver a interferência do usuário ou se a coleta deve ser feita automaticamente (Archer et al., 2009; Buneman et al., 2006b). O aspecto "como armazenar os dados de procedência", refere-se ao fato de que a procedência pode ser armazenada juntamente com o dado ao qual ela se refere (Widom, 2005), ou separadamente (Buneman et al., 2006a; Zhao et al., 2006). Também pode-se investigar uma forma de se reduzir o espaço de armazenamento (Anand et al., 2009; Chapman et al., 2008; Heinis e Alonso, 2008). O último aspecto é tornar os dados de procedência disponíveis para que os usuários possam consultá-los, e assim "como consultar os dados de procedência".

O uso de procedência de dados na recuperação de REA pode ser uma ferramenta relevante para fornecer informações adicionais sobre seu conjunto de dados.

Alguns outros autores Pearson (2002), Cameron (2003), Simmhan et al. (2005) consideram que as aplicações de procedência de dados devem ser dividas em 5 categorias, não sendo exclusivas entre si: i) **qualidade dos dados** que se refere a qualidade e confiabilidade dos dados com base em sua origem e as transformações ocorridas; ii) **trilha de auditoria** que é usado para rastrear o caminho dos dados da origem ao destino para determinar erros na geração dos dados como outras formas possíveis de auditoria; iii) **replicação** que são informações detalhadas da procedência as quais permitem sua replicação e derivação dos dados; iv) **atribuição** que estabelece uma relação do direito do autor e a propriedade dos dados; e v) **informacional** que fornece a partir dos metadados um contexto para interpretar os dados.

4.2 Integração de Dados

Integração de dados é um problema enfrentado por aplicações que precisam acessar várias fontes de dados autônomas e heterogêneas (Halevy et al., 2005, 2006). Ele envolve questões em dois níveis: esquema e instância. No *nível de esquema*, a necessidade de se especificar correspondências entre esquemas de fontes de dados heterogêneas que se referem a uma mesma entidade surge devido à não uniformização desses esquemas (Bhattacharjee e Jamil, 2012; Nguyen et al., 2011; Unal e Afsarmanesh, 2010). Por exemplo, uma fonte pode tratar entidades *localização* de forma genérica, enquanto que outra fonte pode considerar diferentes tipos de localização, como *a cultura* e *a região*. Além disso, esquemas de fontes heterogêneas podem possuir nomes de atributos distintos para representar um mesmo conceito. Por exemplo, uma fonte pode armazenar a data em atributos com nome *data*, enquanto outra fonte pode armazenar a mesma informação com o nome *timestamp*. Outro conflito em nível de esquema refere-se ao fato de que atributos que representam o mesmo conceito podem estar armazenados em diferentes tipos de dados. Por exemplo, em uma fonte a data pode ser armazenada no formato *mês/ano*, enquanto que em outra fonte a data pode ser armazenada apenas como *ano*.

No *nível de instância*, são dois os principais problemas (Prabhakar et al., 1993): ambiguidade na identificação de entidades e conflito de valores de atributos. A *ambiguidade na identificação de entidades* consiste em determinar quais entidades de fontes distintas são similares e, portanto, referem-se à mesma entidade no mundo real. A pesquisa sobre ambiguidade na identificação de entidades é bastante extensa, e tem sido denominada como resolução de entidades e reconciliação de referências (Ferreira et al., 2012; Shu et al., 2011; Whang e Garcia-Molina, 2012; Zhu et al., 2010). Basicamente, as técnicas existentes na literatura visam a geração de agrupamentos de entidades que têm certo grau de similaridade entre si e que, portanto, têm alta probabilidade de serem a mesma entidade do mundo real.

Ainda com relação ao nível de instância, o *conflito de valores de atributos* refere-se ao fato de que diferentes fontes podem possuir valores conflitantes para atributos de entidades similares. Nesses casos, mesmo identificando-se que duas ou mais entidades são similares, elas podem armazenar valores heterogêneos para um mesmo atributo. Assim, para cada agrupamento pode ser útil gerar uma entidade integrada que represente todas as entidades similares, contendo apenas dados integrados que sejam os mais corretos possíveis. Isso é chamado na literatura de fusão de dados, e tem tido bastante enfoque recentemente (Cao et al., 2013; Dong et al., 2010; Fan et al., 2013).

Especificamente com relação à integração de metadados, existe uma função de mapeamento chamada de *Crosswalk*, a qual é considerada um mapeador semântico. O *Crosswalk* é usado para traduzir diferentes conjuntos de elementos de metadados. Os elementos em cada conjunto de metadados são correlacionados com os elementos do outro conjunto de metadados que possuem o mesmo significado ou que possuem um significado semelhante (Reitsma et al., 2012).

O *Crosswalk* faz uma comparação entre os elementos de dois padrões de metadados heterogêneos e constrói uma associação conceitual e estrutural em relação aos elementos dos dois padrões. A Figura 4.1 ilustra a arquitetura de um *Crosswalk* que faz a integração entre os esquemas de metadados X e Y. Como pode ser visto, a função de mapeamento *Crosswalk* tem como entrada dois esquemas de metadados heterogêneos. Em seguida é realizado o *Crosswalking*, o qual obtêm informações de tabelas de mapeamento estruturais já existentes para realização da integração. Por fim, gera-se um esquema *Crosswalk* para os dois tipos de metadados, o qual representa um esquema integrado. Por tanto, o *Crosswalk* realiza apenas integração de esquemas e não realiza integração de instâncias.

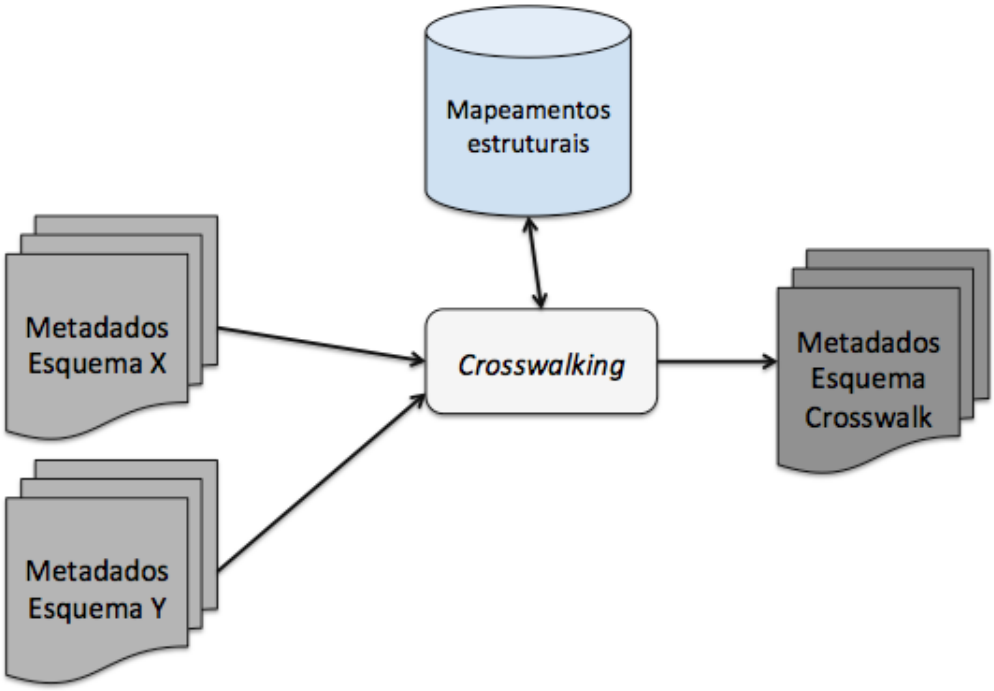


Figura 4.1: *Crosswalk*: Integração de dois esquemas de metadados.

Por exemplo, considere o exemplo ilustrado na Figura 4.2, no qual são representados metadados de repositórios REA considerando o esquema OGP e o esquema DCMES. Os únicos elementos que são similares nesses dois esquemas são os elementos *title* e *type*. Os

demais elementos não são equivalentes. O esquema integrado é ilustrado levando-se em consideração essa similaridade.

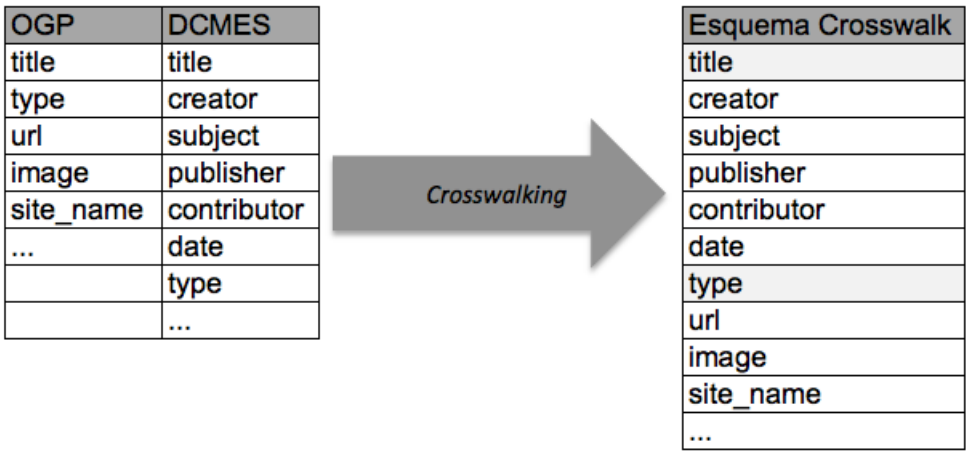


Figura 4.2: *Crosswalk*: Integração de dois esquemas de metadados REA.

4.3 Considerações Finais

Neste capítulo foram descritos conceitos relacionados à procedência e integração de dados. Foi feita uma descrição do uso de procedência de dados no contexto de REA. Em seguida, foram descritos aspectos relacionados à integração de dados, em nível de esquema e instância. Por fim, foi descrita a função de mapeamento *Crosswalk*, a qual é usada para resolver problemas de integração de esquemas de metadados REA heterogêneos. No próximo capítulo, capítulo 5, são descritos trabalhos correlatos a esta pesquisa de mestrado.

Capítulo

5

Trabalhos Correlatos

Conforme descrito na Seção 1.2, esta pesquisa enfoca na proposta de um mecanismo de busca na Web por REA. Neste capítulo são descritos os trabalhos correlatos considerando as seguintes perspectivas relacionadas a presente dissertação de mestrado, que são: (i) mecanismos de busca genéricos na Web (Seção 5.1); e (ii) mecanismos de busca na Web por REA (Seção 5.2). Para cada perspectiva, também são destacadas as limitações desses trabalhos correlatos considerando o contexto deste projeto. Na Seção 5.3 são destacados os diferenciais do mecanismo de busca desenvolvido.

5.1 Mecanismos de Busca Genéricos na Web

Com relação aos mecanismos de busca genéricos na Web, pode-se destacar os trabalhos de Brin e Page (2012) e de Hogan et al. (2011). O trabalho de Brin e Page (2012) refere-se ao protótipo do mecanismo de busca do Google, cuja arquitetura e principais funcionalidades são descritas na Seção 5.1.1. Na Seção 5.1.2 é descrito o mecanismo de busca na Web Semântica de Hogan et al. (2011).

5.1.1 Brin e Page (2012)

É possível visualizar a Web como um conjunto de páginas, onde cada página possui um ou mais *links* para outra página. A Figura 5.1 mostra um exemplo de um pequeno conjunto de páginas Web representado por meio de um grafo, em que as páginas são representadas pelos vértices e os *links* pelas arestas.

A estrutura interna, a estrutura de *links* e o conteúdo de páginas são usadas pelos mecanismos de busca na Web para a recuperação de informação. Analisando a estrutura de *links* das páginas, é possível classificá-las e ordená-las, como faz o algoritmo PageRank (Brin e Page, 2012). O PageRank é um algoritmo de pontuação que ordena os

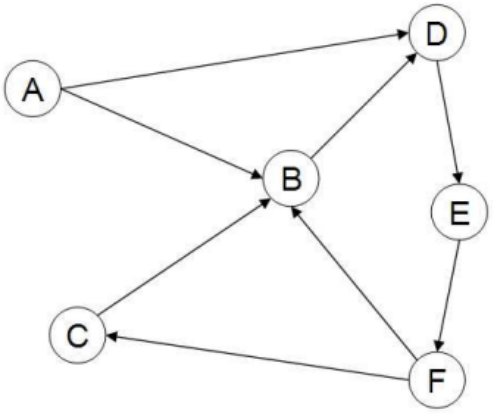


Figura 5.1: Um simples exemplo de seis páginas representadas pelos vértices de A à F.

resultados de busca usando como base a importância de cada documento. Cada página Web pode ter uma série de *links* de saída (arestas de saída da página) e *links* de volta (arestas de entrada da página), como visto na Figura 5.1. O algoritmo atribui uma pontuação a cada página, de forma que essa pontuação é determinada em termos da quantidade de *links* de volta que a página possui. Quanto maior a pontuação, mais *links* de volta uma página possui e maior é sua importância. O algoritmo forma uma distribuição de probabilidade, de modo que a soma dos valores de pontuação de todas páginas do conjunto procurado seja 1.

O cálculo da pontuação do PageRank segue uma técnica iterativa, pois o algoritmo trata a Web como um modelo de Markov. Considere A uma determinada página, $C(A)$ o número de *links* de saída da página A, $(T_1, T_2, ..T_n)$ as páginas que apontam para a página A e d o fator de amortecimento no intervalo $0 < d < 1$ (normalmente é usado 0,85). O valor do PageRank para A é calculado como:

$$PageRank(A) = (1-d) + d * (\frac{PageRank(T_1)}{C(T_1)} + ... + \frac{PageRank(T_n)}{C(T_n)}) \qquad (5.1.1)$$

Embora esse mecanismo de busca recupere resultados relevantes, ele não é adequado para a busca por REA. Isso se deve ao fato de que os REA encontram-se armazenados em repositórios e plataformas que disponibilizam o acesso aos seus recursos por meio de interfaces de consultas criadas especificamente para esse fim. Nesse contexto, as páginas

contendo os recursos solicitados são geradas em resposta às consultas realizadas por meio dessas interfaces. Como resultado, essas páginas não possuem *links*, o que faz com que grande parte dos recursos disponíveis permaneça escondido do mecanismo de busca. Além disso, o mecanismo de Brin e Page (2012) retorna todas as páginas da Web de forma generalizada e assim dificulta encontrar os REA na Web.

5.1.2 Hogan et al. (2011)

O trabalho de Hogan et al. (2011) introduz a arquitetura e a implementação de um mecanismo de busca semântico na Web (SWSE). Essa arquitetura inclui os seguintes componentes: *crawler*, tratamento do dado, indexador e interface do usuário para busca, sendo que as funcionalidades genéricas desses componentes foram descritas no Capítulo 3. A principal característica do SWSE é que ele trabalha com dados Web no formato RDF. Assim, o seu *crawler* se diferencia dos tradicionais por buscar na Web apenas por dados estruturados, especificamente arquivos XMLs, que é o formato mais usado para descrever o RDF.

Apesar das vantagens introduzidas pelo SWSE, ele lida apenas com RDF e não lida com outras formas de instanciação, que é um dos fatores principais que prejudicam a recuperação dos padrões de metadados. Além disso, Hogan et al. (2011) trata o mecanismo de busca apenas para uma busca semântica .

5.2 Mecanismos de Busca na Web por REA

Existem poucos trabalhos na literatura que visam propor mecanismos na Web especificamente para a busca por REA. A seguir, são descritos os seguintes mecanismos de busca verticais na Web por REA, a saber: (Seção 5.2.1) mecanismo de busca proposto por Warpechowski (2005) que consiste recuperar OA em um repositório local; (Seção 5.2.2) mecanismo de busca proposto por Bissell et al. (2009) que recupera REA a partir de um conjunto repositórios que utilizam *feeds*; (Seção 5.2.3) mecanismo de busca proposto por Abeywardena et al. (2013) que recupera os REA a partir de uma lista estática de repositórios; (Seção 5.2.4) mecanismo de busca com objetivo de recuperar planos de ensino para Ciência da Computação e que sejam REA; e (Seção 5.2.5) um mecanismo de busca proposto pelo Comitê de Sistemas de Informação Conjunta (JISC) do Reino Unido, baseado no DSpace, *ElasticSearch* e SOLR.

5.2.1 Warpechownski (2005)

O trabalho de Warpechowski (2005) define técnicas para a recuperação de metadados de objetos de aprendizagem (OA), com a mínima intervenção do usuário, resultando na indexação e recuperação desses objetos. Essas técnicas são definidas com base na estrutura e funcionamento do AdaptWeb, que é um ambiente de aprendizagem que disponibiliza material instrucional.

As técnicas utilizadas na recuperação de metadados são: i) **análise do OA** que obtém as informações do cabeçalho de arquivo do OA; ii) **pré-definição de metadados** que busca na base a existência de outro OA do mesmo tipo; iii) **inferência** que usa uma base de inferência e as informações do OA que pretende-se recuperar.

As técnicas propostas são direcionadas a um tipo específico de metadados, o padrão IEEE/LOM, referente a um único repositório local, o AdaptWeb. Essas técnicas possuem como limitação, portanto, o fato de não considerarem a recuperação de REA armazenados em diferentes repositórios e plataformas, o que introduz a necessidade da resolução de conflitos em nível de esquema e em nível de instância oriundos da heterogeneidade dos REA.

5.2.2 Bissell et al. (2009)

Bissell et al. (2009) introduzem um mecanismo de busca que recupera recursos a partir de um conjunto de repositórios REA que utilizam *feeds*. Os *feeds* fornecem uma lista de URLs que indicam onde determinados recursos podem ser encontrados. O *crawler* desse mecanismo utiliza os *feeds* para realizar um rastreamento direcionado dos recursos dentro de cada repositório incorporado ao mecanismo. Ele recupera cada recurso e adiciona o seu conteúdo a um índice, que pode então ser usado para recuperar resultados relevantes a partir dos termos buscados.

A principal limitação desse trabalho é que ele é direcionado a um conjunto limitado de repositórios REA, os quais necessariamente devem prover *feeds*. Entretanto, pode ser que nem todos os repositórios REA forneçam *feeds*. Adicionalmente, esse trabalho correlato não lida com a heterogeneidade dos REA atualmente disponíveis.

5.2.3 Abeywardena et. al (2013)

OERScout é uma proposta recente de mecanismo de busca por REA na Web (Abeywardena et al., 2013). Ele recupera os REA a partir de uma lista estática de repositórios, conforme mostra a Tabela 5.1.

Repositório	Universidade
Connexions	Rice University
OCW Athabasca	Athabasca University
OCW Capilano	Capilano University
OCW USQ	University of Southern Queensland
OCT Open Content	University of Cape Town
OpenLearn	The Open University
WikiEducator	COL & Ottago Polytechnic
Unow	University of Notthingham
TESSA	Multiple African Universities
OER AVU	African Virtual University
WOU OER	Wawasan Open University

Tabela 5.1: Tabela de REA usado pelo OERScout.

A Figura 5.2 mostra o funcionamento do OERScout. Esse mecanismo de busca utiliza uma abordagem tradicional para classificação de texto, a qual extrai todas as palavras do documento removendo formatação e pontuação para formar o *Corpus*. Em seguida, é realizada uma "tokenização" dos termos, removendo as *stop words* e gerando uma lista de termos. Com relação ao rastreamento de REA, são considerados apenas os índices virtuais disponíveis que estão no formato XML (*Sitemap*). Portanto, não foi desenvolvido um *crawler* específico. O OERScout foi implementado (interface e os algoritmos) usando Microsoft Visual Basic.NET e o SGBD MySQL.

Para atribuir pesos aos termos gerados, o OERScout utiliza um modelo clássico de RI, chamado de técnica Matriz Termo-Documento (TDM). Considerando que $tf_{i,j}$ é o número de ocorrências do termo i no documento j, N é o número total de documentos, df_i é o número de documentos que contém o termo i, o peso $w_{i,j}$ é dado pela fórmula $w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$.

Abeywardena et al. (2013) também utilizam um subconjunto do TDM em forma de matriz chamado KDM (***K**eyword-**D**ocument **M**atrix*) para sugestão de palavras-chaves na pesquisa. O KDM foi construído normalizando os valores de TF-IDF para os termos do TDM, aplicando o princípio de Pareto (80:20[33]).

O OERScout possui diversas limitações em comparação ao projeto proposto. A primeira é que ele não descreve como realiza o rastreamento de REA na Web. Em particular, esse é um grande desafio encontrado por mecanismos de busca. Ademais, o OERScout é limitado a um conjunto de repositórios REA que devem possuir um índice virtual (*sitemaps*) no formato XML. Portanto, ele não lida com a heterogeneidade dos REA atualmente disponíveis.

[33]Valor empírico, encontrado por meio da seleção dos recursos

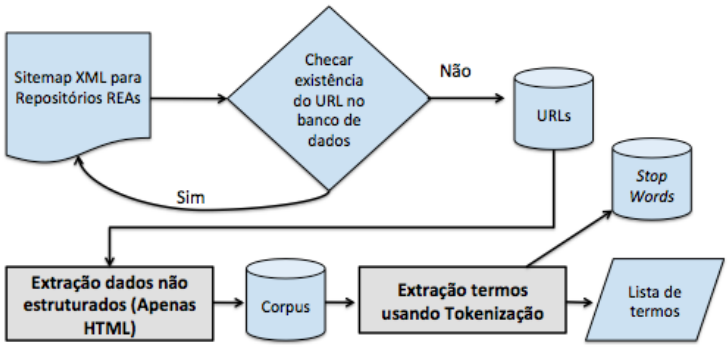


Figura 5.2: Funcionamento do OERScout (adaptado de Abeywardena et al. (2013)).

5.2.4 Rathod e Cassel (2013)

Rathod e Cassel (2013) descrevem um mecanismo de busca na Web usando classificadores. O objetivo do mecanismo consiste em recuperar especificamente planos de ensino para Ciência da Computação e que sejam REA. Para tanto, os autores criaram uma coleção de planos de ensino e treinaram vários classificadores de aprendizagem de máquina. O uso de classificadores de aprendizagem de máquina foi motivado por duas razões. A primeira refere-se ao fato de que um plano de ensino não é estritamente definido e que os programas de ensino variam em conteúdo. Com o aprendizado de máquina, os algoritmos aplicados têm a capacidade de aprender e adaptar-se. A segunda razão diz respeito à grande escala de dados da Web e à necessidade de algoritmos automatizados que possam rapidamente ler milhares de páginas e tomar decisões inteligentes sobre elas.

Foram usados recursos de mineração de textos, como o *bag-of-words* e o *stemming*, na tarefa de classificação. Rathod e Cassel (2013) citam que não é eficaz utilizar apenas o *bag-of-words*, visto a grande quantidade de palavras irrelevantes (*stop words*) e a ineficácia de trabalhar com os classificadores com grande quantidade de palavras nos conjuntos. Então, foi feito um corte das *stop-words* e utilizado o algoritmo de *stemming* para reduzir as palavras à sua raiz, nos conjuntos criados na *bag-of-words*. A Figura 5.3 mostra uma visão geral da arquitetura de Rathod e Cassel (2013).

Para aprendizagem de máquina foram usados três algoritmos de classificação: (i) Árvores de decisão (DT); (ii) *Naïve Bayes* (NB); e (iii) *Support Vector Machines* (SVM). Além disso, foram usados dois algoritmos para o agrupamento, o k-vizinhos mais próximos (KNN) e as florestas aleatórias. O objetivo foi identificar três grupos: o plano de ensino, a disciplina e os componentes do plano. Com todos os classificadores treinados, os dados rastreados pelo *crawler* foram analisados e obteve-se cerca de 2.946 planos de

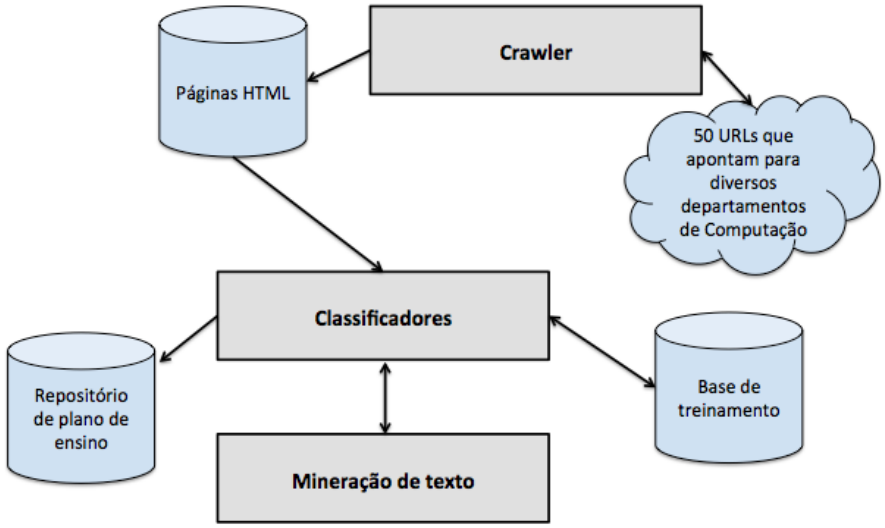


Figura 5.3: Visão geral da arquitetura do mecanismo de busca de Rathod e Cassel (2013) (adaptado de Rathod e Cassel (2013)).

ensino das 88.003 páginas. Por fim, foi criado um repositório para realizar buscas nos planos de ensino (incluindo também a busca facetada).

O trabalho Rathod e Cassel (2013) buscou uma forma de recuperar planos de aulas na Web. Porém, o trabalho se restringe apenas em planos de aulas e não consideram nenhum padrão de metadados. Por outro lado, o mecanismo desenvolvido nesta dissertação de mestrado trabalha em rastrear os REA na Web, considerando os diferentes padrões de metadados e realiza a integração dos padrões de metadados heterogêneos.

5.2.5 Jorum (2013)

Jorum[34] é um mecanismo de busca por REA na Web proposto pelo Comitê de Sistemas de Informação Conjunta (JISC) do Reino Unido, baseado no DSpace, *ElasticSearch* e SOLR (Jorum, 2013a). Este mecanismo também disponibiliza um serviço de armazenamento de materiais, o qual é restrito à comunidade institucional do Reino Unido. Para função de repositório, o Jorum utiliza especificamente a plataforma DSpace (Jorum, 2013c), com suporte apenas à licença Creative Commons. O padrão de metadados utilizado pelo repositório é o Dublin Core estendido (Jorum, 2013a). Na Figura 5.4 é mostrado a arquitetura do Jorum, a qual é dividida em quatro componentes: i) interface do usuário, a qual é capaz de realizar uma busca facetada por algumas características intrínsecas dos REA; ii) repositório de materiais, o qual é restrito para a comunidade institucional

[34] `http://www.jorum.ac.uk/`

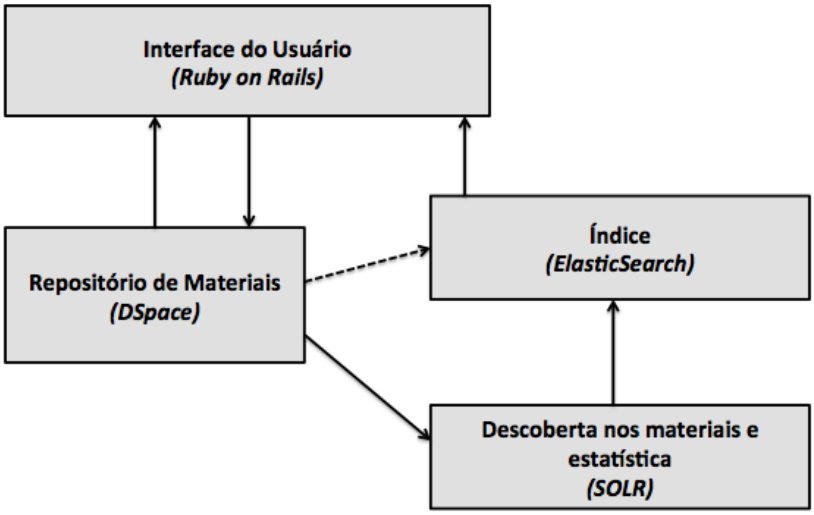


Figura 5.4: Visão geral do JORUM, um mecanismo de busca por REA (adaptado de Jorum (2013a)).

realizar o *upload* de materiais REA; iii) descoberta (reconhecimento) de REA e estatística sobre o mecanismo, o qual utiliza uma lista de repositórios REA pré-definidos; e iv) índice utilizado pelo *ElasticSearch* para retornar o resultado.

A Figura 5.5 mostra um exemplo de busca facetada. Desse modo, é possível recuperar REA usando alguns atributos comuns entre eles, como: comunidade, instituição, autor, licença, palavras-chaves, assuntos de curso superior e outros assuntos.

O projeto é mantido pelo MIMAS (*Manchester Information and Associated Services*) associado à Universidade de Manchester (Jorum, 2013c). O Jorum possui 15.760 REA indexados no seu sistema para consulta (Jorum, 2013b).

Porém, seu mecanismo de busca na Web possui algumas limitações. A primeira é que ele não possui *crawler*. A segunda, é a consequência da primeira, ele não considera a heterogeneidade dos padrões de metadados. A terceira limitação, devido a essa pré-definição ele deixa de indexar muitos repositórios existentes como nota a última análise feita com 15.760 REA indexados.

5.2.6 BioOER (2015)

O BioOER é um mecanismo de busca por REA voltado para vídeos de biomedicina. Zhao et al. (2015) cita que que os mecanismos de busca na Web tradicionais não recuperam REA de forma eficiente, e não é trivial a construção de um mecanismo de busca na Web para recuperação de REA.

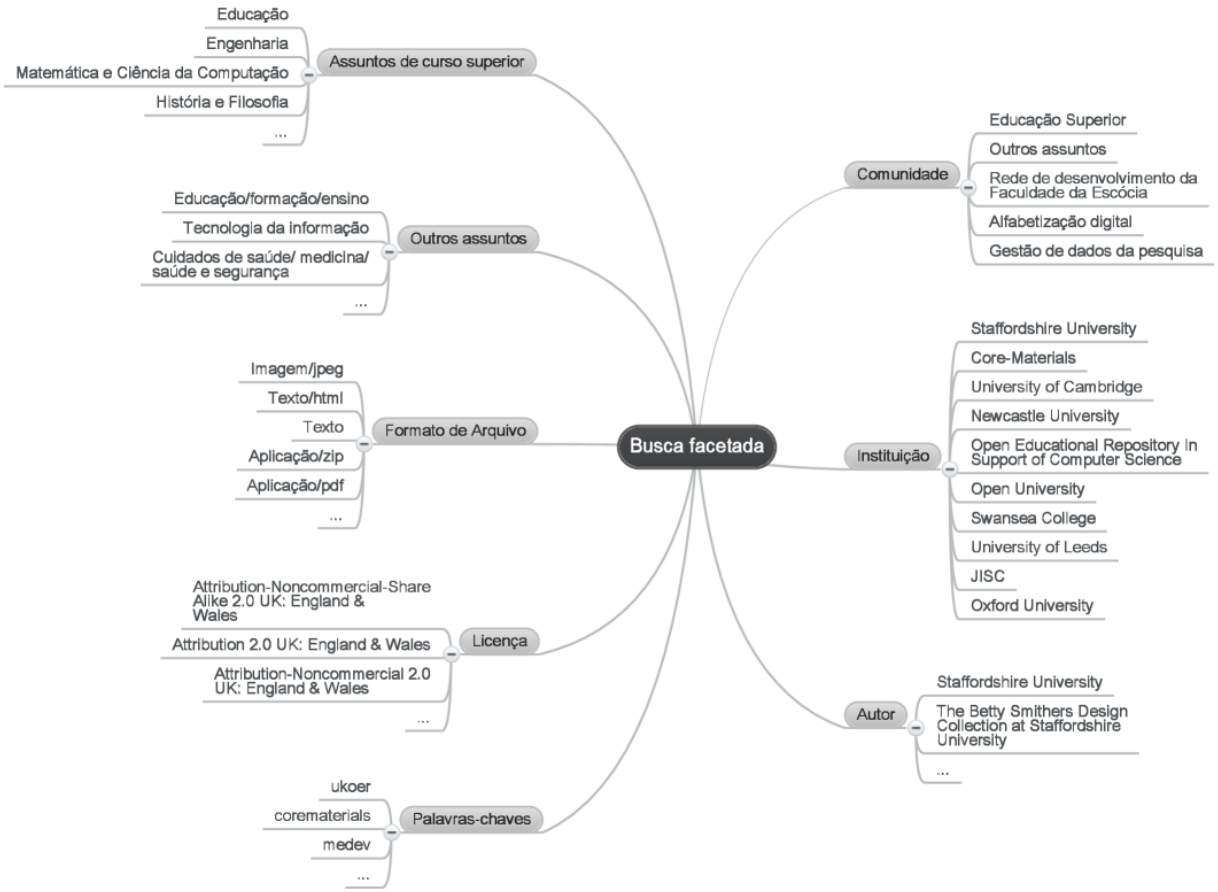


Figura 5.5: Busca facetada utilizada pelo mecanismo de busca Jorum.

O BioOER coletou e processou 25.000 vídeos do Youtube e extraiu os textos da descrição dos vídeos. A Figura 5.6 mostra a arquitetura construída por Zhao et al. (2015). A extração do conteúdo é feito de forma automatizada, construída especificamente para os vídeos do Youtube. O índice do conteúdo do vídeo é armazenado em um repositório baseado em banco de dados relacional (MySQL). A interface de consulta faz o acesso a essa base de dados e retorna para o usuário o resultado de busca.

O BioOER difere do SeeOER em diversos aspectos. O SeeOER foi projetado para um mecanismo de busca na Web em grande escala e não apenas para um repositório, como é o caso do BioOER que é voltado apenas para o YouTube. Também, o SeeOER considera a heterogeneidade dos repositórios. Os metadados de vídeos também são considerados pelo SeeOER e não apenas um padrão. Como é o caso do BioOER.

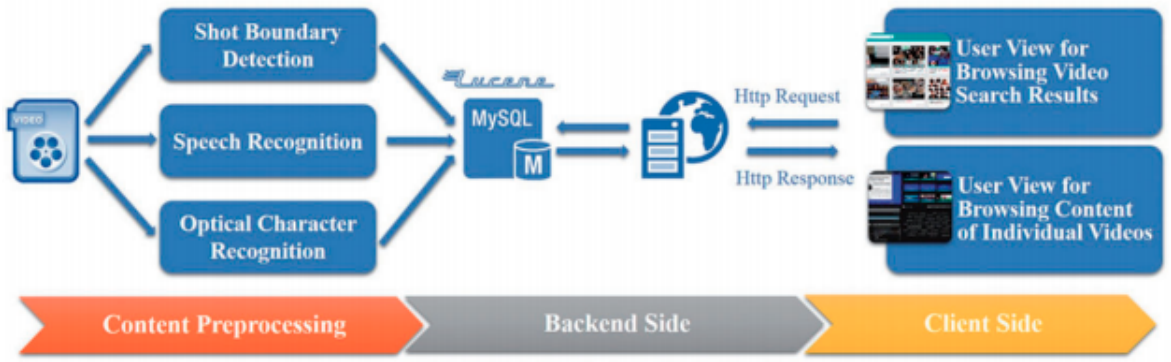


Figura 5.6: Mecanismo de Busca BioOER (Zhao et al., 2015).

5.3 Considerações Finais

Apesar de existirem trabalhos correlatos sobre mecanismos de busca na Web genéricos, mecanismos de busca na Web por REA, como os descritos respectivamente nas seções 5.1 e 5.2. Porém, nenhum dos trabalhos correlatos consideram um um mecanismo de busca por REA que incorpore aspectos de padrões de dados heterogêneos, processo de integração e as características intrínsecas de REA.

Este projeto de mestrado visa avançar no estado da arte por meio de um mecanismo de busca na Web especificamente voltado à recuperação de REA que supre as limitações existentes na literatura. Como diferenciais, com enfoque no *Crawler* por REA na Web e os padrões metadados, na resolução de conflitos em nível de esquema oriundos do uso de diferentes repositórios e plataformas.

No próximo capítulo são detalhados o mecanismo de busca na Web por REA e seus componentes desenvolvidos nesta dissertação de mestrado.

Capítulo

6

Uma arquitetura para mecanismos de buscas na Web por REA usando integração de esquemas

Este capítulo apresenta uma arquitetura de um mecanismo de busca na Web por REA, em que foram projetados componentes específicos para recuperação de REA na Web. O mecanismo de busca desenvolvido visa identificar na Web REA em repositórios e plataformas heterogêneos, os quais são desenvolvidos de acordo com diferentes padrões de metadados. O SeeOER [35] desenvolvido ao longo desta pesquisa de mestrado contribuirá para o avanço nas pesquisas nas áreas de REA e afins.

No Capítulo 5 foram descritos trabalhos correlatos a esta pesquisa que são de mecanismos de busca genéricos na Web, de mecanismos de busca na Web especificamente por REA, de componentes específicos para o tratamento de recuperação de REA na Web. Para cada trabalho correlato, foram destacadas as suas limitações, as quais motivaram o desenvolvimento desta pesquisa de mestrado.

Inicialmente foi feita uma análise dos padrões de metadados usados por REA na Web a partir dos quais foram selecionados para serem usados no SeeOER como descrito na seção 2.2. Em seguida, foi concebida a arquitetura SeeOER com base em um mecanismo na Web em grande escala e desenvolvido um protótipo. Sementes foram escolhidas para os testes realizados a partir de pesquisas na Web, essas sementes constam no apêndice deste trabalho. Foram considerados os repositórios brasileiros como também de outros

[35] ***S**earch **E**ngin**E** for **O**pen **E**ducation **R**esources*

países. Em seguida os REA foram capturados e comparados aos trabalhos correlatos. Foram realizados experimentos quantitativos e qualitativos.

Nos experimentos qualitativos e nos experimentos quantitativos os dados obtidos dos trabalhos correlatos foram retirados do Jorum (Jorum, 2014) e do SeeOER. Os outros mecanismos de busca correlatos não foram mencionados na comparação, pois estes não possuem uma versão aberta ou para pesquisa, o que dificulta o acesso e a comparação real. A complexidade aumenta, principalmente, por tratarem de bases diferentes, expressões de buscas distintas e não apenas um número comparativo.

Este capítulo está organizado da seguinte forma. Na Seção 6.1 são descritos as diretrizes de desenvolvimento do SeeOER. Os componentes do SeeOER e a arquitetura são divididos na arquitetura (Seção 6.2), nesta arquitetura se encontra diversos componentes, que se destacam o *crawler* (Seção 6.3), o componente de integração (Seção 6.4), a indexação (Seção 6.5) e a consulta (Seção 6.6).

6.1 Diretrizes de projeto do SeeOER

A arquitetura tem como base inicial uma arquitetura de busca em grande escala (Brin e Page, 2012; Siqueira, 2013), os padrões de metadados e a fundamentação teórica da Seção 3.2. Porém, com diversas diferenças e avanços na área de recuperação de REA. O SeeOER possui arquitetura diferenciada dos trabalhos correlatos, seguem nas próximas subseções seus detalhes.

6.1.1 Padrões de metadados investigados

Nesta pesquisa de mestrado foram investigados, como uma amostra diferencial, alguns tipos de metadados usados por alguns repositórios REA. Esses repositórios foram retirados da lista de repositórios REA que estão no Apêndice B. Foram inspecionados os repositórios REA e encontrados os tipos de metadados usados por eles.

Os repositórios armazenam REA de diversos formatos de arquivos, como: textos, imagens, animações, arquivos de áudio, vídeos e outros. Com essa pequena porção de repositórios REA é possível verificar a diversidade entre os repositórios, desde os formatos de arquivos até os tipos de metadados utilizados. A Tabela 6.1 resume os resultados obtidos. O tipo de metadados *Personalizado* refere-se aos tipos de metadados não catalogados.

Cada esquema de metadados possui uma forma de instanciação e diferentes atributos para descrever o metadados, os quais podem ser similares ou não. Os esquemas que descrevem os metadados e a forma de instanciação foram importantes para todo projeto.

Nome do repositório	Principal formato de arquivos	Metadados
Banco de Imagens Geográficas	imagens	ISO 19115:2003
Flickr	imagens	EXIF e Personalizado
Biblioteca Digital de Ciência	animações	Não encontrado
Banco Internacional de Objetos Educacionais	imagens e arquivos de áudio	DCMES
Connexions	textos	DCMES, IMS, CNX, MathML e OGP
OCW-MIT	vídeos e textos	OGP e Personalizado
Teses USP	textos	DCMES
Khan *Academy*	vídeos	OGP e MathML
e-Aulas USP	vídeos	Vídeo Sitemaps

Tabela 6.1: Repositórios REA heterogêneos com diferentes metadados.

A Tabela 6.2 sintetiza os principais padrões encontrados na Web dos repositórios REA (Gazzola et al., 2014). Além disso, são mostrados os elementos descritores o qual faz referência aos campos que possibilita descrever os REA. Por fim, a possibilidade de extensão do padrão o que possibilita adicionar mais elementos descritores.

Padrão de metadados	Instanciação	Elementos descritores	Possibilidade de estender?
OGP	1 formato	4 elementos	Sim
Vídeo Sitemaps	1 formato	6 elementos	Sim
Dublin Core	5 formatos	15 elementos	Sim
IEEE/LOM	Não descrevem formatos	60 elementos divididos em 9 categorias	Sim

Tabela 6.2: Padrões de metadados encontrados em repositórios REA

Para o desenvolvido do SeeOER foram utilizados todos os padrões de metadados mencionados na Tabela 6.2, que são: OGP, Vídeo Sitemaps, Dublin Core ou DC, IMS ou IEEE/LOM. A possibilidade de extensão e a possibilidade de mais de uma forma de instanciação dos padrões de metadados aumentaram a complexidade de recuperação da estruturação dos dados.

6.1.2 Componente Crawler

O *crawler* desenvolvido pelo SeeOER é descrito na Seção 6.3 e específico por padrões de metadados de REA. Os padrões foram definidos na Seção 2.2. Esse *crawler* desenvolvido é único, específico, e diferencia da arquitetura de Brin e Page (2012); Siqueira (2013). O

funcionamento específico do *crawler* de Brin e Page (2012); Siqueira (2013) são pouco detalhados por eles. Porém, existem diversas diferenças, desde os padrões de metadados de REA, nas sementes, fronteiras, e na forma de tratamento dos dados estruturados e não estruturados realizados pelo SeeOER.

Em Hogan et al. (2011) o *crawler* é específico para Web semântica e utiliza apenas o formato *RDF*. Como mencionado na Seção 2.2, o RDF é apenas uma forma de instanciação do DCMES e não considera nenhum outro padrão de metadados. Além disso, ele não descreve detalhes específicos sobre seu *crawler*. Enquanto, o *crawler* desenvolvido no SeeOER considera diversos formatos de instanciação e diferentes padrões de metadados. No trabalho de Warpechowski (2005) não é utilizado um *crawler*, tendo em vista que funciona localmente e não na Web. E, são os próprios usuários que inserem os objetos educacionais. O trabalho de Bissell et al. (2009) é baseado em *feeds*, como descrito na Seção 5.2.2 e não considera a heterogeneidade dos padrões de metadados. A sua implementação e maiores detalhes não são especificados por Bissell et al. (2009). Rathod e Cassel (2013) não descreve seu *crawler*. Por fim, Jorum (2013), utiliza o DSpace como base para o mecanismo de busca. Como descrito na Seção 5.2.5. E, não utiliza um rastreador para Web.

6.1.3 Componente de Integração de Esquemas

O trabalho Chen et al. (2015) cita semelhanças usando o *Crosswalk* com metadados. Porém, no SeeOER foi desenvolvido uma forma para resolução de *Crosswalk* considerando graus de semelhanças. O qual não se viu na literatura. E, assim permite uma maior flexibilidade entre os diversos metadados existentes. A maior dificuldade esteja na validação específica deste componente. O qual ele está diretamente integrado com o mecanismo de busca. Mas, os resultados do mecanismo permitem verificar que os padrões de metadados existentes não foram prejudicados. O Componente de Integração de Esquemas desenvolvido para o SeeOER é descrito na Seção 6.4.

6.1.4 Componente Indexador

O Componente Indexador desenvolvido pelo SeeOER é descrito na Seção 6.5. Ele teve como bases o modelo de indexação de Brin e Page (2012); Siqueira (2013) e a fundamentação teórica. Porém, foram realizadas modificações em relação à representação de forma binária e nos detalhes de implementação.

Bissell et al. (2009); Hogan et al. (2011); Jorum (2013a); Rathod e Cassel (2013) não citam detalhes de indexação.

6.1.5 Componente da Interface de Consulta

O modelo de consulta teve como base o modelo de Jorum (2013a). Porém, com diversas mudanças em relação aos padrões de metadados e saídas específicas de formatação da API.

Em Brin e Page (2012); Siqueira (2013) trabalha sem padrões de metadados, o qual não era viável para o SeeOER. Em Bissell et al. (2009); Hogan et al. (2011); Rathod e Cassel (2013) não especificam detalhes específicos sobre sua interface de consulta.

6.2 Arquitetura

O SeeOER é derivado da arquitetura de um mecanismo de busca na Web em grande escala, como citado nas diretrizes de projeto do SeeOER. Sua arquitetura possui dois processos principais, o **processo de indexação** e o **processo de consulta**.

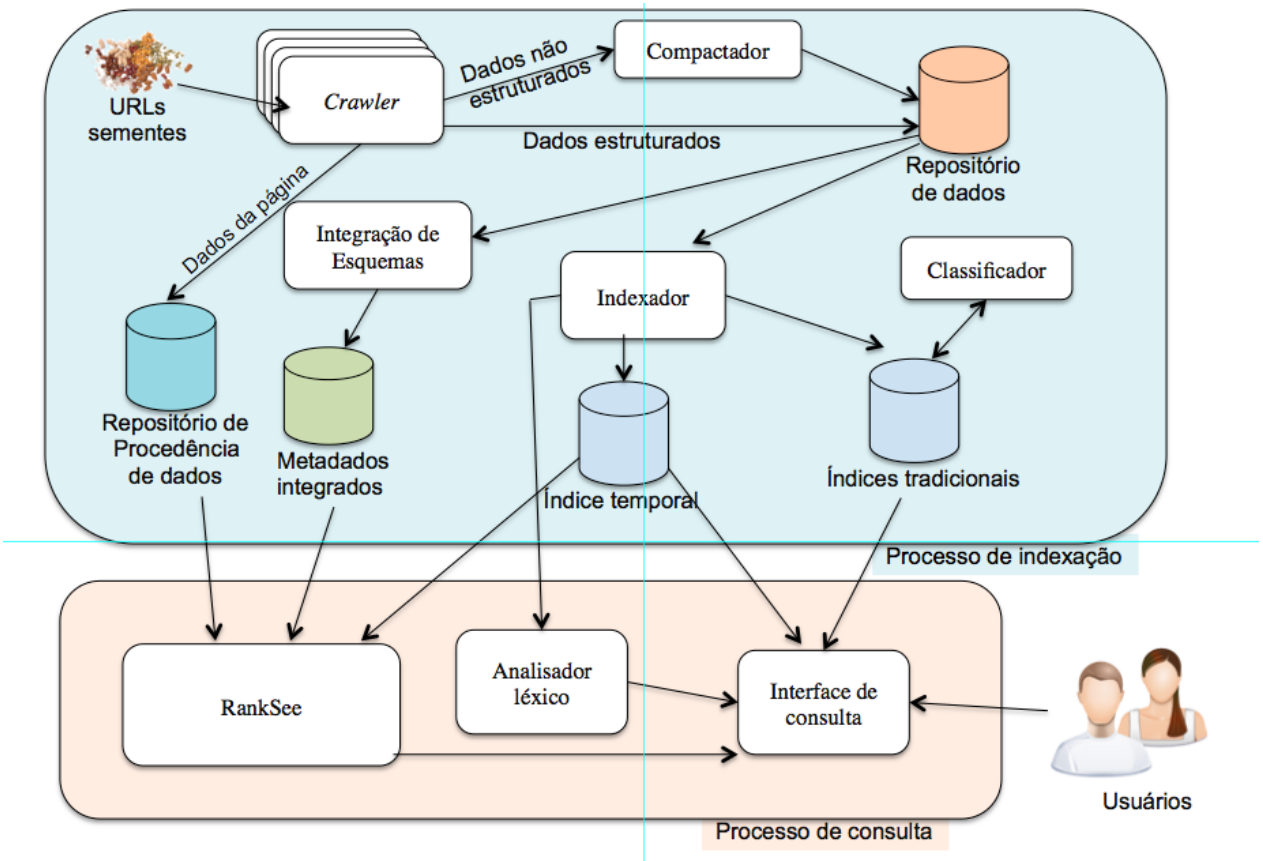


Figura 6.1: Arquitetura do SeeOER

As funcionalidades dos componentes da arquitetura são descritas a seguir. O processo de indexação inicia com uma lista de URLs sementes a serem rastreadas pelo ***Crawler***, o

qual faz o *download* das páginas referentes a esses URLs de forma distribuída. Na Seção 6.3 é descrito o funcionamento detalhado do ***Crawler*** usado para obtenção de padrões de metadados de REA.

Os dados estruturados representam os metadados, enquanto que os dados não estruturados representam o próprio material, o qual pode estar em diversos formatos (por exemplo, HTML, PDF, DOC). Os dados não estruturados são enviadas ao *compactador*, o qual comprime e armazena-os no repositório, de forma que cada URL é associado a um número de identificação único. Os dados estruturados são armazenados no repositório, mas de forma que possam ser uniformizados futuramente. Os dados das páginas, como origem do servidor, última atualização, entre outros dados, são armazenados no repositório de procedência de dados.

O componente de ***integração de esquemas*** tem como objetivo criar um esquema integrado representativo, que represente a maioria dos metadados. Esse esquema representativo é armazenado no repositório de *metadados integrados*. Na Seção 6.4 é descrito o algoritmo desenvolvido para esta tarefa.

No **processo de consulta** o componente ***RankSee*** terá como funcionalidade recuperar e ranquear os REA, levando em consideração características intrínsecas de REA os quais podem auxiliar na busca e atendam às requisições dos usuários.

Entre o **processo de consulta e o processo de indexação** o **Indexador** tem várias funções. Ele lê o repositório de dados, descompacta os documentos e os analisa. Cada documento é convertido em um conjunto de ocorrências de texto chamado de *hits*. Cada *hit* possui a palavra, a posição no documento, uma aproximação do tamanho da fonte e a capitalização (maiúsculo ou minúsculo). Esses *hits* são distribuídos na forma de um índice para frente (*forward index*) no repositório de índices tradicionais. O **Classificador** realiza, dentre outras operações, a geração de um índice invertido que é usado pela **Interface de consulta** conjuntamente com o Analisador Léxico e o RankSee. O índice temporal armazena o ID dos documentos, status de funcionamento, quantidade de iterações e última atualização.

6.3 Crawler

O *Crawler* é um componente da arquitetura do SeeOER. O *Crawler* também pode ser chamado de **Coletor**. Em síntese o processo se dá por meio da captura de hyperlinks e processamento de páginas. Porém, o *Crawler* deve obedecer as restrições do arquivo *robots.txt*, caso esse arquivo exista na raiz do repositório a ser coletada. O arquivo de

robots é conhecido como um protocolo a ser seguido pelos *Crawlers* em que especifica algumas regras de acesso ao site ou repositório a serem coletadas.

Na arquitetura do SeeOER foi projetado um *Crawler* para metadados das páginas dos repositórios REA, como também os dados de procedências das páginas capturadas. O funcionamento do *Crawler* é descrito no algoritmo apresentado na Figura 6.2, e detalhado a seguir.

entrada: S = páginas sementes, P_n = profundidade, E_n = esquemas de metadados, S_m=estado da máquina

```
1  recupera-estado(S_m)
2  fila-de-URLs← S
3  repeat
4  |  página ← x ∈ fila-de-URLs
5  |  armazena-Proc(pagina_dados)
6  |  metadados_pagina ← leitura-metadados(página)
7  |  if metadados_pagina ∈ E_n then
8  |  |  armazena-DEst ({metadados_pagina} ⊄ repositorio-DEst)
9  |  end if
10 |  conteúdo ← Download(página)
11 |  (textos,links_p, estruturas, ...) ← Analisar(conteúdo)
12 |  links = ∀w ∈ links_p | w ≤ P_n
13 |  fila-de-URLs ← Adicionar-novos-links (fila-de-URLs, links)
14 |  salvar-estado(S_m)
15 until (fila-de-URLs ≠ ∅);
```

Figura 6.2: Algoritmo de um *Crawler* por REA

O *Crawler* recebe como entrada algumas páginas sementes do repositório, a profundidade, os esquemas de metadados e o estado da máquina. A profundidade, neste contexto, significa a quantidade máxima de níveis permitidas que o *Crawler* pode percorrer. O diretório raiz (/) é o nível zero e os sub-diretórios seguintes são os níveis maiores. O estado da máquina representa a possibilidade do *Crawler* ser pausado. Inicialmente, o *Crawler* retoma o estado da máquina, o qual pode ser vazio no começo (linha 1). As páginas sementes são incluídas na fila (linha 2). O laço de repetição, em seguida, desenfileira os URLs até que não exista nenhum URL para ser retirado da fila (linha 3 à 14). Nesse laço de repetição, são feitas as seguintes ações: uma página x é atribuída da fila-de-URLs para ser analisada (linha 4); são armazenados dados da página para o repositório de procedência (linha 5); é feita uma leitura dos metadados da página escolhida (linha 6); Se os metadados pertencerem em algum esquema de metadados (linha 7), então esses metadados que ainda não existem no repositório são armazenados, sendo considerados

dados estruturados (*DEst*) (linha 8). Todas as informações não estruturadas também são armazenadas para que possam ajudar na recuperação dos REA (linha 10 e 11). Em seguida, os links capturados da página passam por uma seleção. São selecionados todos os links exceto aqueles que tiverem profundidade maior que o permitido (linha 12). Esses links são adicionados na fila-de-URLs, nessa inclusão apenas os links que ainda não estão na fila são adicionados (linha 13). O estado da máquina é salvo (linha 14).

A função de armazena-Proc($pagina_{dados}$) do algoritmo, obtém dados externos e dados internos de procedência. Os **dados externos** são IP, Protocolo, Servidor, URL e data de última modificação. Esses dados externos são obtidos diretamente pelo servidor de resposta ao *Crawler* (Cliente). Como segue na Figura 6.3.

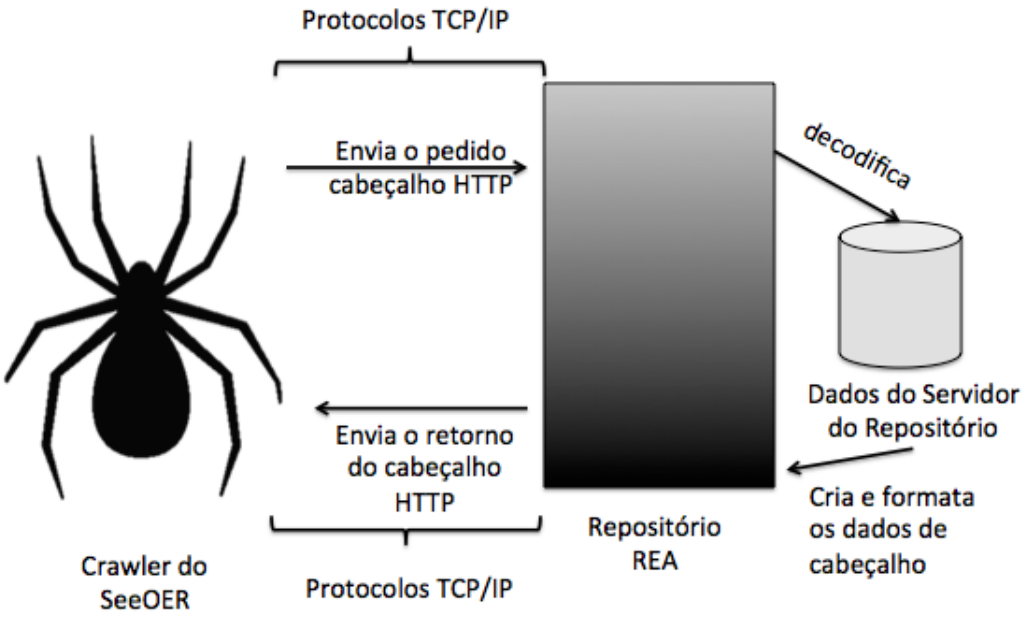


Figura 6.3: Obtendo dados externos para procedência

```
GET http://www.ocw.unicamp.br/ HTTP/1.1
Accept : text/html
If-Modified-Since : Saturday, 15-January-2015 03:23:11 GMT
User-Agent : Agent SearchEngine_for_OER-Crawler
```

Figura 6.4: Pedido HTTP

As figuras, Figura 6.4 e Figura 6.5, mostra o pedido HTTP realizado pelo Crawler do SeeOER e o retorno enviado pelo repositório da OCW da Unicamp.

Os **dados internos** de procedência são obtidos diretamente nas páginas do repositório em que se encontra os REA. O SeeOER considera os seguintes dados internos de proce-

```
HTTP/1.0 200 OK
Date : 25 Ago 2015 17:38:32 GMT
Server : Microsoft-IIS/3.0
Content-Type : text/HTML
Content-Length : 1028
Last-Modified : 25 Ago 2015 16:21:10 GMT
```

Figura 6.5: Retorno HTTP

dência de um material REA, quando disponível: i) TimeZone do material; ii) autor da alteração; iii) data da modificação; e iv) versão da alteração. A Figura 6.6 é um exemplo de dados internos usados como dados de procedência pelo SeeOER.

As fronteiras do *Crawler* do SeeOER foram estimadas por meio da seguinte expressão regular:

$$\boxed{([a-z0-9])^*.<\textit{URL-canônica}>/([a-z0-9]^*)^*}$$

O que significa que diversas páginas de um repositório (um mesmo URL canônico) podem ser rastreados. Porém, quando o *Crawler* encontra um link fora da sua fronteira, por exemplo um link de propaganda que não possui uma URL canônica que satisfaça a expressão regular, o *Crawler* não incluirá na fila de URLs a serem futuramente visitadas.

A Figura 6.7 mostra o funcionamento do *crawler* e rodando as *threads* para obtenção de uma grande escala de metadados e páginas. É possível observar uma quantidade elevada de páginas lidas e que estão na fila. Além disso, o uso das fronteiras é essencial para o funcionado adequado do *crawler*.

A Figura 6.8 é o diagrama de objetos do Crawler. O diagrama de objetos representa o estado do sistema em um dado momento e é um diagrama estático ou estrutural. Neste caso, o diagrama está mostrando um caso genérico em um dado momento. Outros detalhes técnicos não foram considerados, como por exemplo robots e *threads*.

Para o parser do html foi considerado o *BeautifulSoup*, o melhor analisador léxico considerado por Richardson (2015). O BeautifulSoup gera um grafo direcional da página. Na Figura 6.9 mostra o grafo gerado pelo *BeautifulSoup*. As nuvens representam a possibilidade dos padrões de metadados estarem instanciados nesses locais pelos repositórios. Essas nuvens demonstram a nebulosidade desses padrões possíveis para o *Crawler* obter dados importantes como metadados instanciados. Para isso, foram criados 6 regras de instanciações de padrões de metadados de REA em repositórios heterogêneos.

ξ_1 Regra InstaMeta

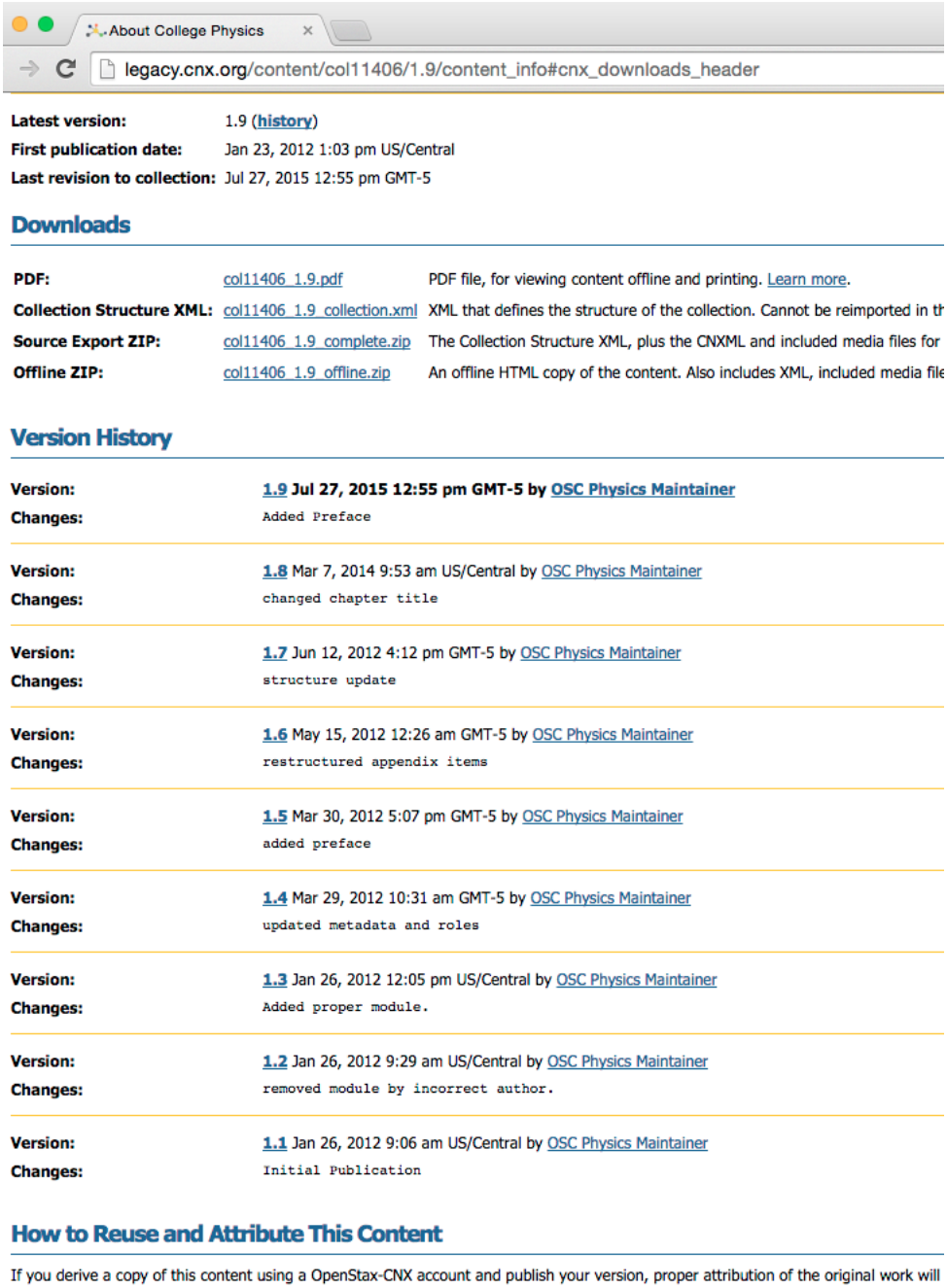


Figura 6.6: Dados internos de procedência

$\xi_1 = \forall$ <html>
<head>
<meta name="$\alpha.Atributo_1$" content="$Conteudo_1$" .../>$^{\psi}$
<meta name="$\alpha.Atributo_n$" content="$Conteudo_n$" .../>$^{\psi}$
...

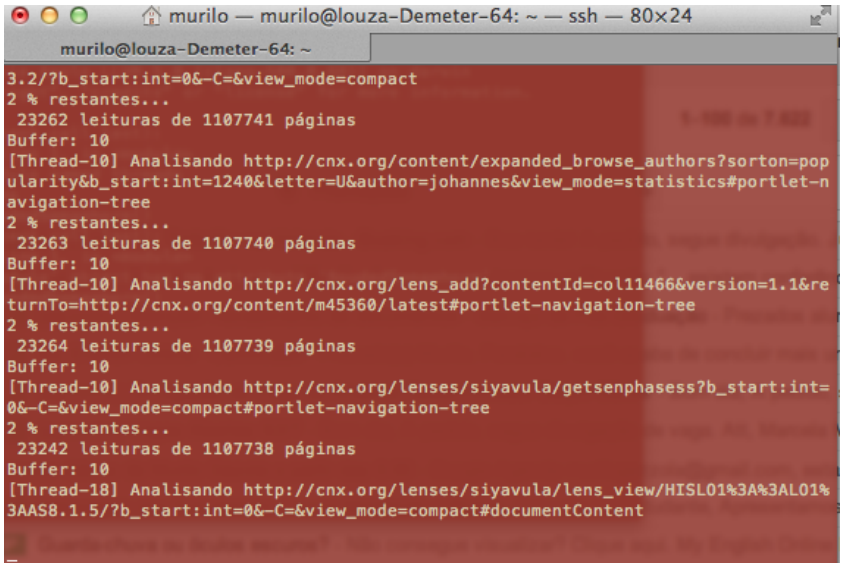


Figura 6.7: Crawler SeeOER em funcionamento

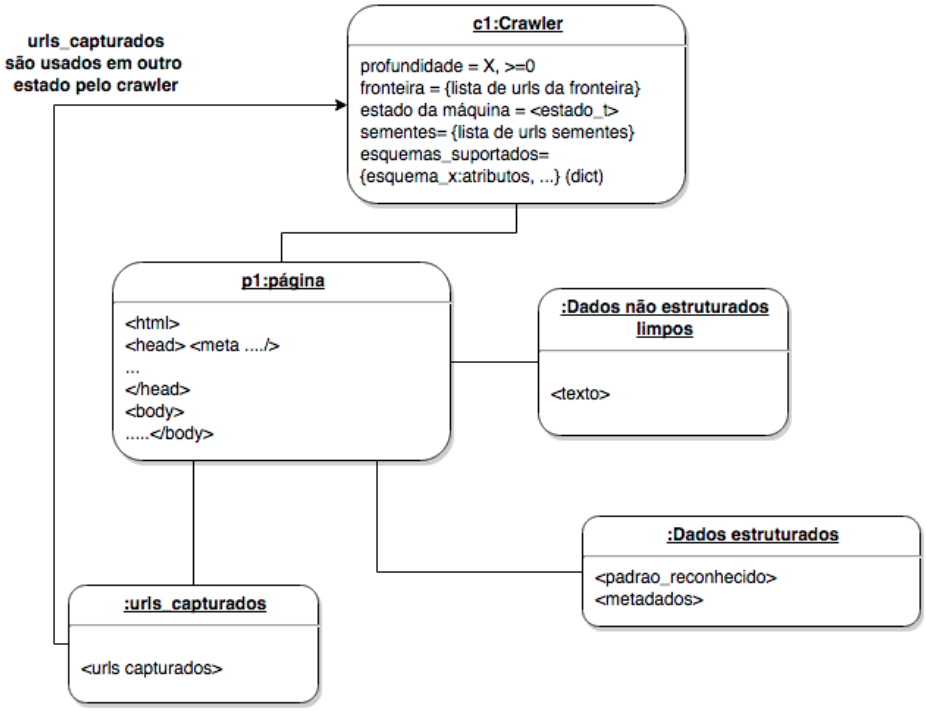


Figura 6.8: Diagrama de objetos do *Crawler*

</head>

</html> | $\alpha.Atributo_n \subset \alpha.Schema \wedge \psi \geqslant 1$

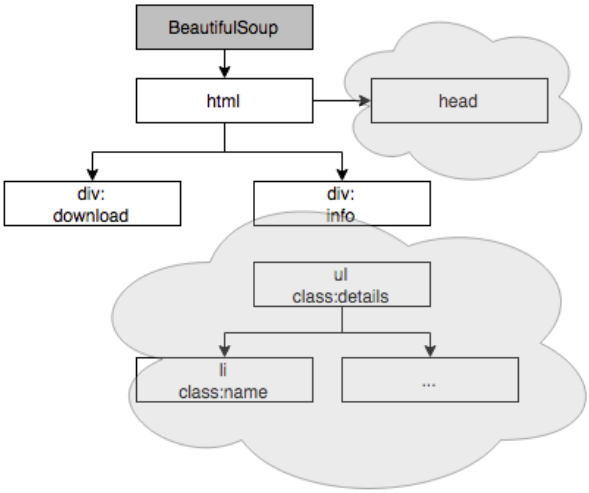


Figura 6.9: Grafo HTML possível

Para regra ξ_1, o $\alpha.Schema$ pode estar instanciado no código, como também pode não estar presente. Quando instanciado, ele aparece no código como <*link rel=Schema...*> . O ψ é a quantidade de instanciação em que é possível ser 1 ou mais.
As próximas regras ($\xi_2, \xi_3, \xi_4, \xi_5, \xi_6$) não possuem *Schema* instanciado. Apenas padrões possíveis e semânticas implícitas, como seguem abaixo.

ξ_2 Regra DivLi

$\xi_2 = \forall$ <html>
<body>
<div (class $\vee$ id $\vee$ name) =“$info_1$”>$^{\mu}$
<ul (class $\vee$ id $\vee$ name)=“$details_1$”>$^{\psi}$
<li (class $\vee$ id $\vee$ name)=“$name_1$”>$^{\psi}$ *Content*
<li (class $\vee$ id $\vee$ name)=“$name_n$”>$^{\psi}$ *Content*

...
</body></html> | $\psi \geq 1 \wedge \mu \geq 1$

Em todas as regras são considerados $\mu = n$ em que n representa os valores de descidas na árvore DOM do HTML até chegar na *class* ou *id* ou *name* possíveis. O ψ é a quantidade de instanciação em que é possível ser 1 ou mais. Nas regras ξ_2, ξ_3 e ξ_5 os atributos *Info, details e name* estarão implicitamente na semântica do repositório, não

existe um padrão exato para o nome desses três atributos.

A **Regra DivThree** é baseada em 3 tag DIV principais. Em que os atributos são descritos na DIV com (class $\vee$ id $\vee$ name) = $name_1$, ..., $name_n$ no campo *Content*.

ξ_3 Regra DivThree

$\xi_3 = \forall$ <html>
<body>
<div (class $\vee$ id $\vee$ name) =“$info_1$”>$^{\mu}$
<div (class $\vee$ id $\vee$ name)=“$details_1$”>$^{\psi}$
<div (class $\vee$ id $\vee$ name)=“$name_1$”>$^{\psi}$ *Content*
<div (class $\vee$ id $\vee$ name)=“$name_n$”>$^{\psi}$ *Content*
...
</body></html> | $\psi \geq 1 \wedge \mu \geq 1$

Considerando as regras ξ_4 e ξ_6 os atributos são diferentes das demais. Na regra **Regra TwoDiv-Simplificado** somente os atributos *details e name* estarão expressos. E, na **Regra DivOne** somente o atributo *name* estará expresso. Nessas duas regras consideram uma *organização simples* dos atributos do repositório.

ξ_4 Regra TwoDiv-Simplificado

$\xi_4 = \forall$ <html>
<body>
<div (class $\vee$ id $\vee$ name) =“$details_1$”>$^{\mu}$
<div (class $\vee$ id $\vee$ name)=“$nome_1$”>$^{\psi}$ *Content*
<div (class $\vee$ id $\vee$ name)=“$name_n$”>$^{\psi}$ *Content*

...
</body></html> | $\psi \geq 1 \wedge \mu \geq 1$

A **Regra TwoDivSpan** é a única regra considerar o uso da tag <*span*>. Contudo, é uma regra pouco vista na maioria dos repositórios. E, quando instanciada, estará entre DIV.

ξ_5 **Regra TwoDivSpan**

$\xi_5 = \forall$ <html>
<body>
<div (class $\vee$ id $\vee$ name) =“$info_1$”>$^{\mu}$
<div (class $\vee$ id $\vee$ name)=“$details_1$”>$^{\psi}$
<span (class $\vee$ id $\vee$ name)=“$name_1$”>$^{\psi}$ *Content*
<span (class $\vee$ id $\vee$ name)=“$name_n$”>$^{\psi}$*Content*

...
</body></html> | $\psi \geq 1 \wedge \mu \geq 1$

ξ_6 **Regra DivOne**

$\xi_6 = \forall$ <html>
<body>
<div (class $\vee$ id $\vee$ name) =“$name_1$”>$^{\mu}$ *Content*
<div (class $\vee$ id $\vee$ name) =“$name_n$”>$^{\mu}$*Content*
...
</body></html> | $\psi \geq 1 \wedge \mu \geq 1$

As regras ξ_1 **Regra InstaMeta**, ξ_2 **Regra DivLi**, ξ_3 **Regra DivThree**, ξ_4 **Regra TwoDiv-Simplificado**,ξ_5 **Regra TwoDivSpan**, ξ_6 **Regra DivOne** suprem as principais lacunas de instanciação de metadados REA. Mas, não excluem a necessidade de estender possíveis novas regras. Essas regras são de grande valia para um *Crawler* que é um dos componentes chaves do mecanismo de busca na Web. Por isso, a importância de sua eficácia. Mas, também pode ser usado por outros mecanismos ou componentes.

Para exemplificar tais regras a Figura 6.10 mostra a regra ξ_1 **InstaMeta** usado pelos repositórios da USP. E, a Figura 6.11 mostra a regra ξ_3 **DivThree** bastante utilizada por diversos repositórios.

```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="pt-br" lang="pt-br">
 <head>
  <title>Abordagens educacionais baseadas em dinâmicas colaborativas on line.</title>
  <meta http-equiv="content-type" content="text/html; charset=utf-8">
  <meta name="robots" content="index, follow">
  <meta name="description" content="Ao oferecer condições, tanto para que o indivíduo possa enfrentar situações no
  comunicação, como para que ele possa gerar descobertas...">
  <meta name="keywords" content="Abordagens educacionais, Collaborative dynamics, Dinâmicas colaborativas, Educati
  <meta name="generator" content="Biblioteca Digital de Teses e Dissertações da USP">
  <meta name="date" content="2015-01-27T18:58:12-02:00">
  <link rel="stylesheet" href="/estante/site/lib/jquery/css/flick/jquery-ui.css" type="text/css">
  <link rel="schema.DCTERMS" href="http://purl.org/dc/terms/">
  <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/">
  <meta name="DC.language" content="por" xml:lang="pt-br" scheme="DCTERMS.RFC1766">
  <meta name="DC.creator" content="Barbosa, Ana Cristina Lima Santos" xml:lang="pt-br">
  <meta name="DC.contributor" content="Siqueira, Idmea Semeghini Prospero Machado" xml:lang="pt-br">
  <meta name="DC.title" content="Abordagens educacionais baseadas em dinâmicas colaborativas on line." xml:lang="p
  <meta name="DC.title" content="Educational approaches based on collaborative dynamics on line." xml:lang="en">
  <meta name="DCTERMS.abstract" content="Ao oferecer condições, tanto para que o indivíduo possa enfrentar situaçõ
  e comunicação, como para que ele possa gerar descobertas e inovações, novas políticas de formação são requeridas
  sistemas " integrados", que oferecem oportunidades diversas de formação, organizáveis modo flexível, com atividad
  estudantes. as dinÂmicas colaborativas apresentam-se, então, como uma estratégia educativa para formação do futu
  dinâmica colaboração on line em processos educacionais desencadeados cursos via internet. isso, o estudo apresen
  desenvolvida disciplina pósgraduação modalidade internet, quatro edições consecutivas. encaminhamento natureza q
  coletados, foram delimitadas cinco categorias, orientaram ordenaram construção texto da pesquisa: infraestrutura
  gestão. resultado dessa investigação comprovou hipótese dinâmicas possibilitam criação comunidade aprendizagem r
  constituem cerne atual, tanto presencial quanto distância. coletados demonstrou tais consistem num processo comp
  afetivos individuais social conhecimento, onde ocorre identificação pessoal por meio outras pessoas. assim, dever
  objetivos comuns aprendizagem. foco não deve tecnologia, mas atividade humana realização." xml:lang="pt-br">
  <meta name="DCTERMS.abstract" content="Once conditions are offered, not only to the individual can face the new
  technologies, but also to he/she can generate new discoveries and innovations, new policies of formation are req
  that of " integrated" systems, that offer several opportunities of formation, organized by a flexible way, with
  interaction. thus, collaborative dynamics present themselves as an educational strategy to formation future profe
  processes aroused from courses via internet. this end, study presented analysis pedagogical proposal designed re
  consecutive editions. qualitative nature orientation, case method was adopted strategy. collected data regroup,
  technological infra-structure, individual postures, collective methodological strategies management. result inve
  different ways acting quality education, since group activities constitute core current both distance. demonstra
  inter-relation affective cognitive aspects social construction knowledge, personal identification occurs means i
  reaching common objectives learning. focus should not technology, but human activity progress." xml:lang="en">
  <meta name="DC.subject" content="Abordagens educacionais" xml:lang="pt-br">
  <meta name="DC.subject" content="Dinâmicas colaborativas" xml:lang="pt-br">
  <meta name="DC.subject" content="Ensino on line." xml:lang="pt-br">
  <meta name="DC.subject" content="Collaborative dynamics" xml:lang="en">
  <meta name="DC.subject" content="Educational approaches" xml:lang="en">
  <meta name="DC.subject" content="On line teaching" xml:lang="en">
  <meta name="DC.description" content="Tese de Doutorado" xml:lang="pt-br">
  <meta name="DC.description" content="Doctoral Thesis" xml:lang="en">
```


Figura 6.10: Exemplo de instância ξ_1

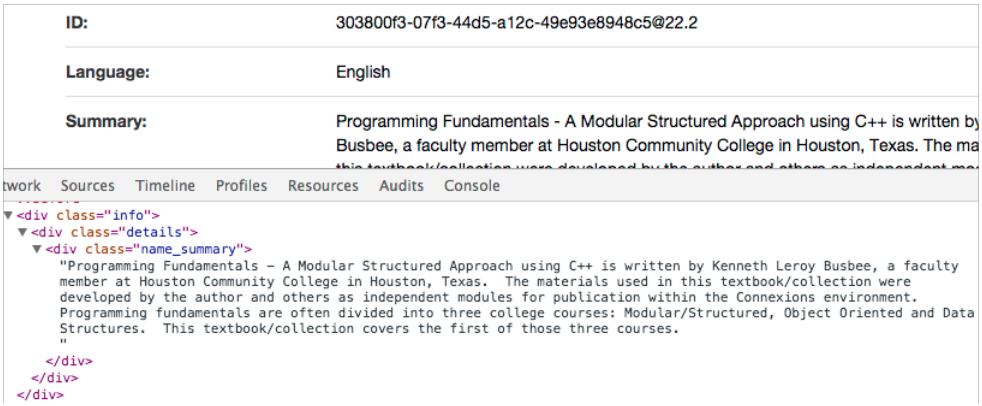


Figura 6.11: Exemplo de instância ξ_3

6.4 Integração de Esquemas

A *integração de esquema* é um componente da arquitetura do SeeOER. A integração de esquemas de metadados no contexto de REA foi divida em três esquemas: Esquema Padrão (EP), Esquema Extensivo (EE) e Esquema Divergente (ED).

O **EP** é o padrão de metadados sem nenhuma modificação. O **EE** é considerado uma extensão do padrão de metadados com suporte pelo **Integrador**, mas que possui modificações e não é idêntico aos padrões suportados atendidos. Por fim, o **ED** que é um padrão ainda não suportado pelo integrador, o qual diverge do EP e do EE, ou que possui um grau de semelhança inferior ao necessário.

entrada: P_S = Padrões Suportados, E_C = Esquema *Crosswalk*, P_{Vx} = Padrão Verificado, G_s=0.9

1 **if** *(*$P_{Vx} \in P_S$*)* $\vee$ *(grau-semelhança(*P_{Vx}*,*P_S*)* $\geq G_s$*)* **then**
2 armazenar-metadados(integrar(E_C,P_{Vx}))
3 **else**
4 armazenar-padrao-divergente(P_{Vx})
5 **end if**

Figura 6.12: Algoritmo para Integrar Esquemas de Padrões Metadados

Com o algoritmo apresentado na Figura 6.12 é possível iterar os padrões de metadados considerando um *padrão crosswalk*. O *padrão crosswalk* é um padrão representativo e inicial. Desta forma, é possível realizar a integração dos padrões heterogêneos considerados por **EE** e **EP**. Inicialmente é feito a entrada dos padrões suportados representado por P_s, o *esquema crosswalk* por E_c, o padrão que deseja verificar por P_{vx}, e G_s é um valor que o usuário deverá escolher e que representa o grau de semelhança flexível. Na linha 1 é verificado se o padrão é suportado pelo mecanismo de busca na Web. Caso o padrão verificado for considerado um **EE** e **EP** o padrão é incrementado no padrão *crosswalk* (integrar ou incrementar) e os metadados são armazenados (armazenar-metadados) na linha 2. Na linha 4 ele armazena o padrão divergente, os quais podem ser considerados no futuro. Na Figura 6.13 é apresentado o *grau-semelhança* usado na linha 1.

No algoritmo *grau-semelhança* da Figura 6.13 é empilhado todos os atributos do metadados a ser verificado e calculado de forma normalizada considerando os atributos suportados, a seguir é detalhado o seu funcionamento. Na linha 3 é empilhado todos os atributos do metadados a ser verificado. Da linha 5 até linha 10 é feito até que nenhum atributo do metadados a ser verificado não esteja na pilha. Na linha 6 até 9 é adicionado

entrada: M_n = Metadados n a ser verificado, PM_y = Conjunto de Atributos Suportados

1 ***Definition*** grau−semelhança(M_n, PM_y)
2 **begin**
3 $stack_{atributos}$=empilhar($\forall n \in M$)
4 $G_{grau-semelhana}$=0
5 **while** !*empty*($stack_{atributos}$) **do**
6 $atributo_x$=desempilhar($stack_{atributos}$))
7 **if** $atributo_x \in PM_y$ **then**
8 $G_{grau-semelhana}$=$G_{grau-semelhana}$+$\frac{1}{atributos_{quantidade}(PM_y)}$
9 **end if**
10 **end while**
11 **return** $G_{grau-semelhana}$
12 **end**

Figura 6.13: Algoritmo grau de semelhança usado para metadados estendidos ou personalizados, mas com semelhanças

um valor incremental no $G_{grau-semelhanca}$ caso o atributo pertença ao padrão suportado. Por fim, na linha 11 é retornado o $G_{grau-semelhanca}$.

Um exemplo simples e explicativo do algoritmo *grau-semelhança* é $PM_y = name, url, ..x_\delta$ e $len(name, url, ..x_\delta) = 12$, $M_n = name, url, ..., x_\gamma$ e $len(name, url, ..., x_\gamma = 11)$ em que $x_\delta \neq x_\gamma$. Neste caso o grau de semelhança é 0.91, e x_γ é considerado não integrante e o padrão é **EE**.

A Figura 6.14 representa um diagrama de interação geral. Diagramas de interação geral são considerados diagramas comportamentais ou dinâmicos. O termo **RD** significa um objeto instanciado para o *repositório de dados*, o termo **RM** do *repositório de metadados integrados*, o termo *RMDiv* é um repositório para uso futuro para armazenamento de *esquemas divergentes*. Após o início, o diagrama pode considerar dois casos. O primeiro é os padrões de metadados sendo EP e o segundo caso é os padrões de metadados sendo EE ou ED. Caso ele considere como um esquema reconhecido, quer dizer que os atributos mínimos do padrão de metadados *verificado* e o *suportado* são satisfeitos. Neste caso, o *crawler* irá armazenar os dados estruturados (armazenarDE) e os metadados serão integrados sendo considerado como um EP. No segundo caso, o *crawler* armazenará os dados estruturados e se tem apenas duas opções possíveis as quais são EE ou ED. Será verificado por meio do *grau de semelhança*. Caso o grau de semelhança for menor do que o grau de semelhança passado, o esquema será considerado divergente. Caso o grau

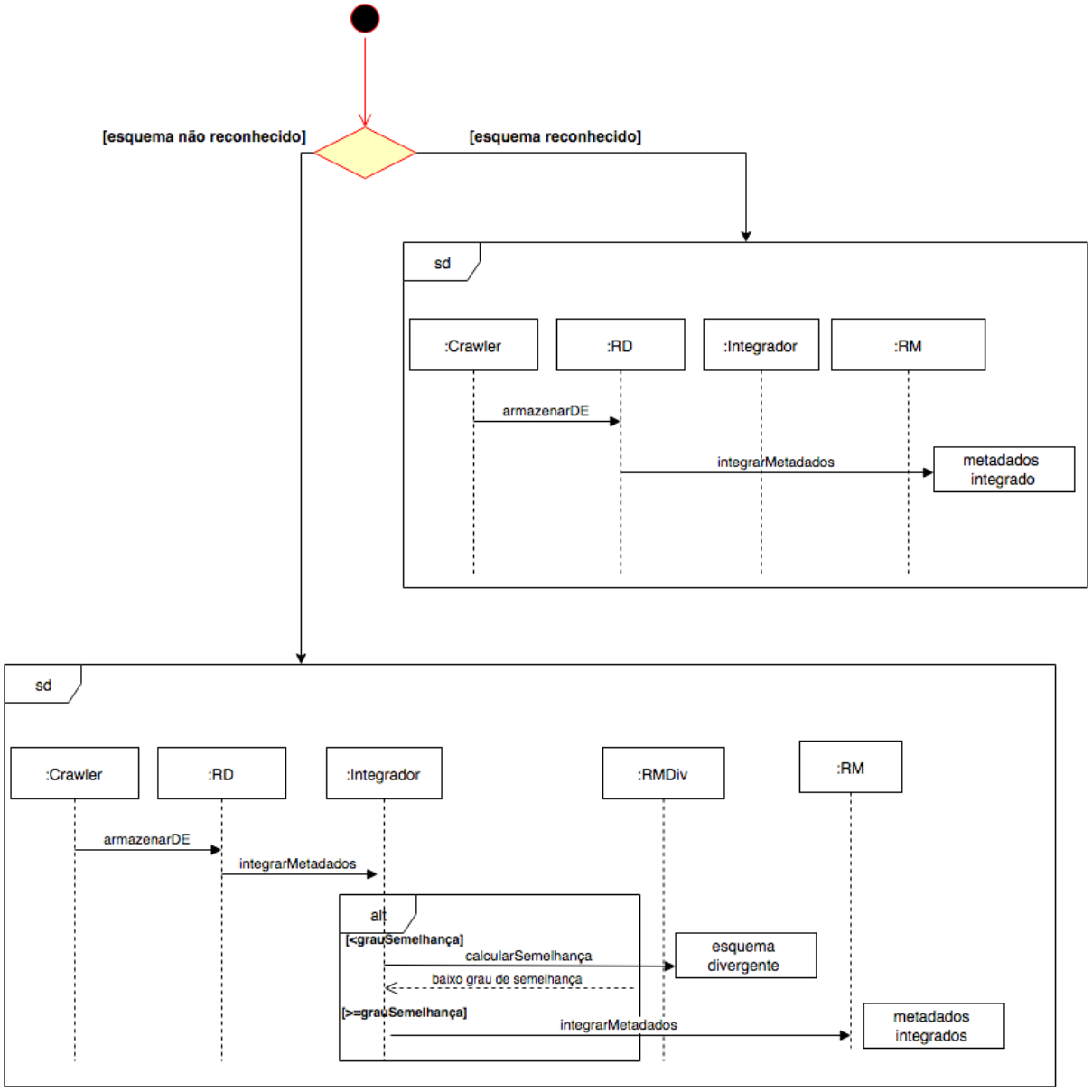


Figura 6.14: Diagrama de interação geral do componente de integração de esquemas da arquitetura do SeeOER

de semelhança for maior ou igual ao instanciado, os metadados serão integrados e será considerado um EE.

6.5 Indexador

Indexador é um componente da arquitetura do SeeOER. Com a execução do *Crawler*, os dados não estruturados são enviadas ao *compactador*, o qual comprime e armazena-os

no repositório, de forma que cada URL é associado a um número de identificação único (Figura 6.15). Os dados estruturados são armazenados no repositório, mas de forma que possam ser uniformizados futuramente.

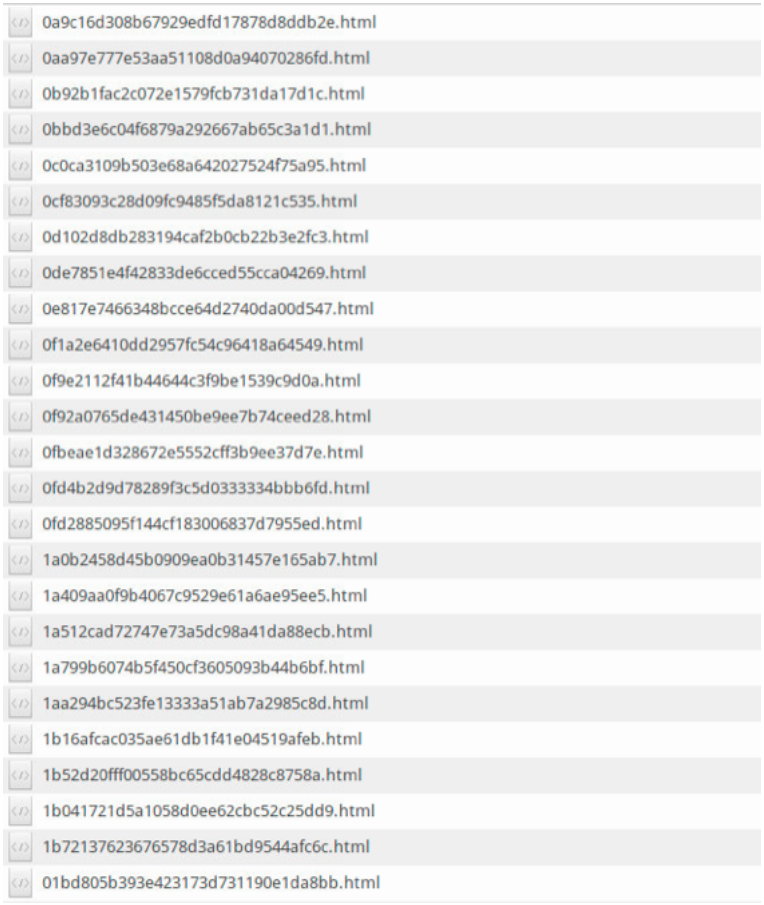


Figura 6.15: Dados não estruturados

O **repositório de dados** armazena os dados não estruturados e os dados estruturados de forma temporária. Os dados não estruturados são comprimidos por meio da biblioteca *zLib*, nativa do Python (LibraryPy, 2013). O repositório não armazena os dados de forma permanente, ele é considerado um armazenamento temporário de dados estruturados e não estruturados. Isso, devido sua capacidade de armazenamento ser inferior a quantidade necessária para armazenar todos os documentos capturados pelo *Crawler* na Web. A temporariedade do repositório é feita por meio do Indexador que remove dados estruturados e não estruturados do repositório de dados, após eles serem indexados no índice.

O **Indexador** tem várias funções. Ele lê o repositório de dados, descompacta os documentos e os analisa. Cada documento é convertido em um conjunto de ocorrências de texto chamado de *hits*. Cada *hit* possui a palavra, a posição no documento, uma

aproximação do tamanho da fonte e a capitalização (maiúsculo ou minúsculo). Esses *hits* são distribuídos na forma de um índice para frente (*forward index*) e armazenados. O classificador gera um índice invertido a partir do índice para frente. Os dados das páginas, como origem do servidor, última atualização, entre outros dados, são armazenados no *Repositório de procedência de dados*. O índice temporal é usado para próximas iteração do *crawler* e pode ser usado pelo RankSee. O ID dos documentos, status de funcionamento, quantidade de iterações e última atualização são armazenados no índice temporal que usa o identificador único das páginas.

A Figura 6.16 mostra uma matriz representando um índice invertido. Nesta figura as linhas são os termos pré-processados e as colunas são os documentos. O termo $P(T_n, D_n)$ é uma função posição que entra com o termo Tn e o documento D_n. É uma matriz simples e representativa que mostra para cada termo versus documento pode ser retornado um ou mais posições no mesmo documento.

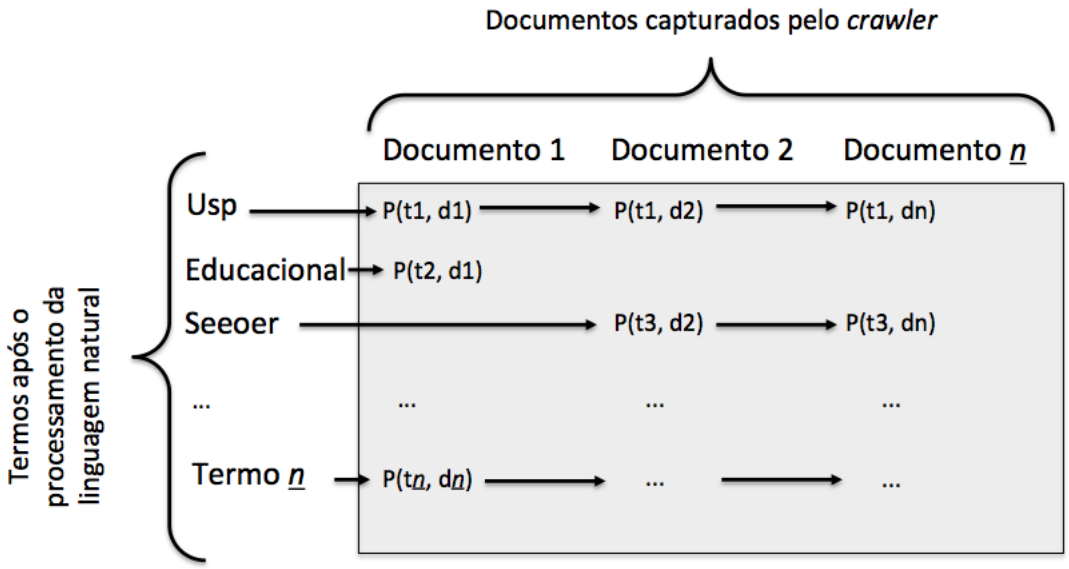


Figura 6.16: Matriz de índice invertido

Para exemplificar o índice invertido será considerados dois documentos, os dois documentos foram retirados de um fragmento de um repositório de REA [36]. Nessa proposta foi considerado um índice invertido completo (como geralmente é chamado), que é um índice que considera as posições no documento que simplificam a busca por frases e consultas por proximidade. No índice invertido sem ser completo é usado a frequência do termo em invés da sua posição.

[36] https://oli.cmu.edu/

Documento 1 = "*The OPEN Learning Initiative offers online courses to anyone who wants to learn or teach. Our aim is to combine open, high-quality courses, continuous feedback, and research to improve learning and transform higher education.*"

Documento 2 = "*Higher education is faced with some big, big challenges from resource issues to compressed teacher/student interaction times. Watch this video to get a glimpse of what doing to help the higher education system respond.* "

A Figura 6.17 mostra um índice invertido dos documentos 1 e 2. Considere que as *stop words* foram eliminadas desses documentos. O termo "open" aparece apenas no documento 1 e nas posições 4 e 116 do documento 1. O termo "higher" está presente no documento 2 na posição 0 e 186 e no documento 1 na posição 212.

```
respond {'doc2': [210]}
help {'doc2': [177]}
doing {'doc2': [168]}
challenges {'doc2': [45]}
video {'doc2': [137]}
learning {'doc1': [9, 189]}
education {'doc2': [7, 193], 'doc1': [219]}
quality {'doc1': [127]}
issues {'doc2': [70]}
feedback {'doc1': [155]}
glimpse {'doc2': [152]}
open {'doc1': [4, 116]}
transform {'doc1': [202]}
research {'doc1': [169]}
offers {'doc1': [29]}
combine {'doc1': [108]}
online {'doc1': [36]}
higher {'doc2': [0, 186], 'doc1': [212]}
wants {'doc1': [65]}
big {'doc2': [36, 41]}
continuous {'doc1': [144]}
watch {'doc2': [126]}
compressed {'doc2': [80]}
student {'doc2': [99]}
teach {'doc1': [83]}
teacher {'doc2': [91]}
improve {'doc1': [181]}
faced {'doc2': [20]}
high {'doc1': [122]}
interaction {'doc2': [107]}
resource {'doc2': [61]}
times {'doc2': [119]}
aim {'doc1': [98]}
courses {'doc1': [43, 135]}
learn {'doc1': [74]}
initiative {'doc1': [18]}
```

Figura 6.17: Exemplo de índice invertido

A Figura 6.18 mostra o índice invertido completo com os *hits*. A capitalização foi representada de forma binária. Por exemplo, o termo "open" está em maiúsculo na primeira posição do documento 1, enquanto na segunda aparição está em minúsculo. Neste caso, a *fonte size* dos termos são todas 10.

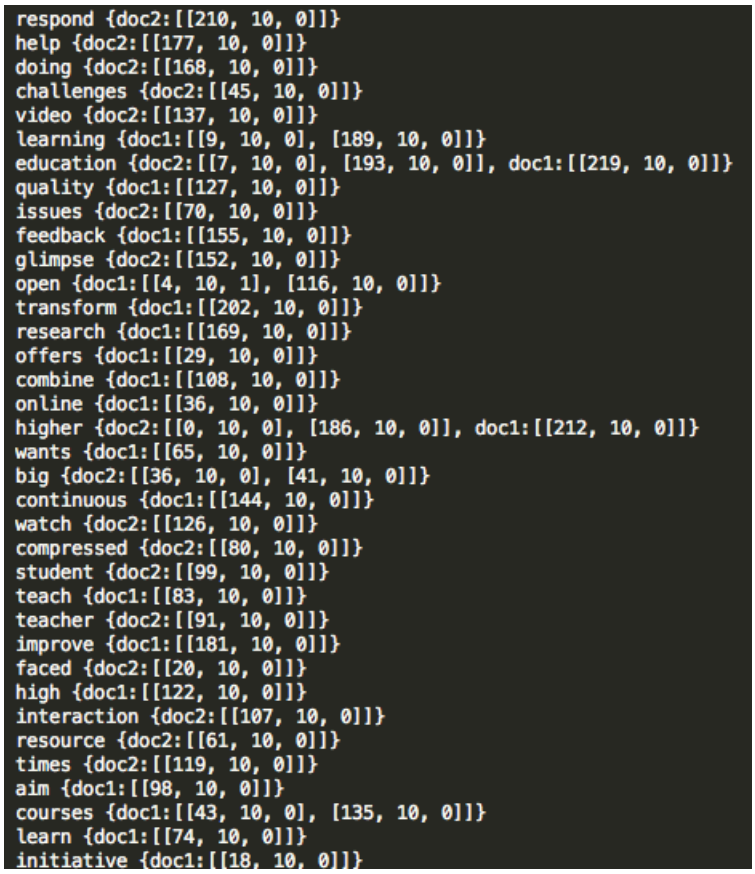


Figura 6.18: Exemplo de índice invertido com hits

O índice invertido é usado pela *Interface de consulta* conjuntamente com o Analisador Léxico e o RankSee. Para o componente *RankSee* foi considerado o ranking modelo vetorial. E, o analisador léxico tradicional.

6.6 Consulta ao SeeOER

Foi desenvolvido uma **consulta tradicional e uma consulta por meio de uma *Application Programming Interface* (API)**. Além disso, nenhum dos trabalhos correlatos consideraram a consulta por meio de uma API. O SeeOER considera que o uso de uma API possibilita o avanço de pesquisas na área e a consulta tradicional como uma forma para disseminação dos REA. A API desenvolvida permite a consulta por funções acessíveis diretamente por programação. A Figura 6.19 mostra a interface de consulta construída para o SeeOER, considerando a arquitetura já mencionada.

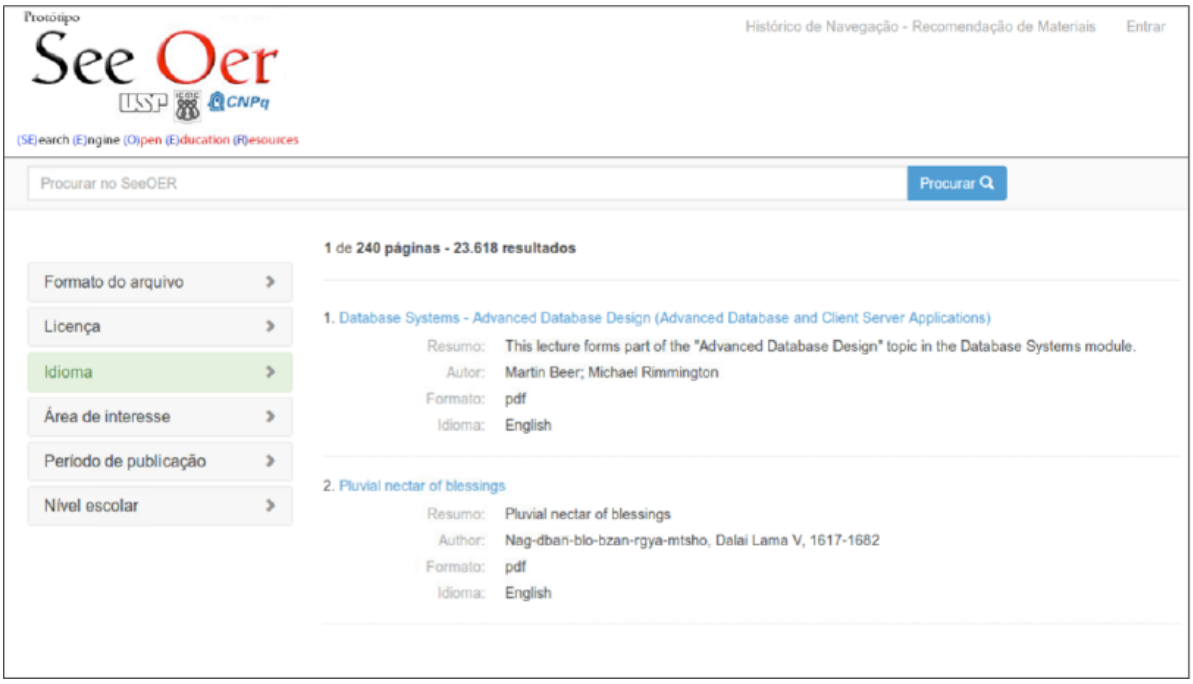


Figura 6.19: Consulta ao SeeOER

Em relação a API desenvolvida, a técnica Representational State Transfer (REST) é a mais usada para o uso da API. Com o uso da API é possível obter acesso diretamente aos resultados indexados. O acesso pode ser feito por meio do navegador ou integrado na programação de outros aplicativos. O retorno da API possui diversos formatos de respostas, as quais são: i) Ruby; ii) JSON; iii) Python (py); e iv) PHP. O retorno já vem sintaticamente e semanticamente construído na opção de formato pedido ao SeeOER. Por exemplo, na Figura 6.20 mostra o retorno do SeeOER para linguagem PHP.

Neste exemplo, da Figura 6.20, mostra os resultados da expressão de busca *"ciência"*. São retornados para esta expressão 685 documentos, por padrão são retornados de 10 documentos por intervalo, começando pelo zero (start=0). É possível alterar este intervalo

```
array(
  'responseHeader'=>array(
    'status'=>0,
    'QTime'=>1,
    'params'=>array(
      'indent'=>'true',
      'q'=>'ciencia',
      'wt'=>'php')),
  'response'=>array('numFound'=>695,'start'=>0,'docs'=>array(
      array(
        'content'=>'Technical Certification criteria of the educational content of the OSR repository | OpenScienceResources   Home OSR Repository Help
Survey Log in Technical Certification criteria of the educational content of the OSR repository Guidelines for the OSR content certification The OSR content
was reviewed by the OSR quality team consisting of partners from the OSR consortium, according to the following certification guidelines: Certification
criteria Certification guidelines Open educational resources (OER) Educational pathways (EP) Metadata The Metadata attached to the resource is complete and
it describes the resource appropriately The Metadata attached to the EP is complete and it describes the resource appropriately Relevance OER content needs
to be relevant with the topic of the OER. Educational pathways content should be relevant to the corresponding OSR and its domain.  EPs need to be have all
the phases filled in. Links All links in OER are functioning and relevant All links in EPs are functioning and relevant. Semantics The titles and
descriptions of the OER are meaningful in rtelation to the content The titles and descriptions of educational pathways are meaningful in relation to the
content. Spam filtering OER should not contain any obvious spam words or inappropriate material. EPs should not contain any obvious spam words or
inappropriate material. These guidelines represent the content certification guidelines for the reviewed content for the duration of OSR project. The
criteria can be changed by the administrator of the portal in the future. The certification of the content concerns the technical aspects of the content.
Contributors of OERs and EPs are responsible themselves for the content they provide with regards to the scientific quality. The quality team in 2009-2012
included the following organizations (in parentheses the languages they were responsible for): Ellinogermaniki Agogi (Greek, English), Palace of Miracles
(Hungarian), HEUREKA (Finnish), University of Jyväskylä (Finnish), Eugenides Foundation (Greek), National Museum of Science and Technology Leonardo Da Vinci
(Italian), Pavilion of Knowledge - Ciencia Viva (Portugese), Bundesministerium fur Unterricht, Kunst und Kultur (German). Imprint    Google Disclaimer
About OpenScienceResources Project     Contact Technical Support',
        'title'=>'Technical Certification criteria of the educational content of the OSR repository | OpenScienceResou',
        'segment'=>'20140516105048',
        'boost'=>0.12507501,
        'digest'=>'941f4c93603b65a391b120d8a78e88b0',
        'tstamp'=>'2014-05-16T14:12:16.043Z',
        'id'=>'http://www.osrportal.eu/node/96431',
        'url'=>'http://www.osrportal.eu/node/96431',
        '_version_'=>1468295259373961216),
      array(
        'content'=>'Eduteka - Artículos > Contenido General > Políticas Públicas > pag: 1 Ingresar Registrarse Recordar Contraseña Quiénes Somos Inicio
Artículos Proyectos Módulos Recursos REDuteka Artículos en Eduteka: Accesibilidad Actividades Alfabetismo en Medios Aprendizaje por Proyectos Aprendizaje en
```


Figura 6.20: Retorno da API do SeeOER no formato de saída PHP

de 10 documentos, para isso é necessário incluir na consulta à API a quantidade de linhas de retorno usando o parâmetro *rows*. Por exemplo, rows=100 retornariam o intervalo de 100 documentos por consulta. Quanto maior o valor do parâmetro *rows* da quantidade de linhas de retorno passadas, maior será o tempo de latência de resposta entre a consulta e a reposta do SeeOER.

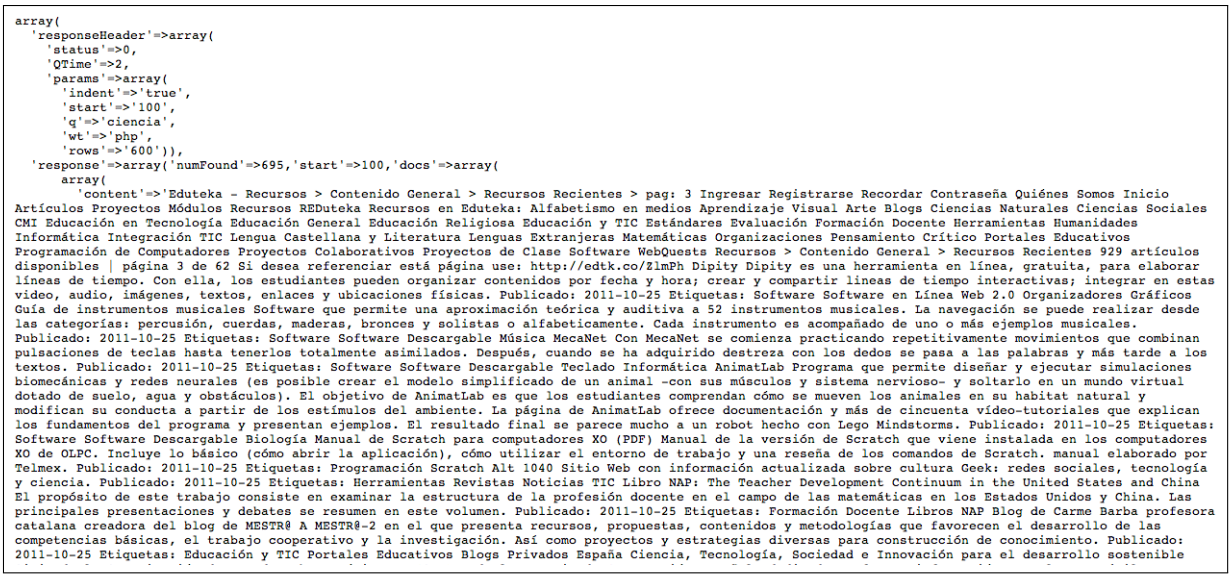
A consulta é feita por meio da seguinte URL *http://usp.seeoer.com/* adicionando os parâmetros de busca, os quais são: expressão de busca (q), inicio (*start*), intervalo (*rows*), indentação (*indent*) e formato. Todos os parâmetros são opcionais, exceto o $\boldsymbol{q}$. É possível incluir os seguintes parâmetros na API, as linhas de retorno *rows*. Por exemplo: `http://usp.seeoer.com/?q=ciencia&wt=php&rows=600`. A API retornaria os 600 resultados da consulta *"ciência"* no formato PHP.

Também é possível incluir a indentação usando o parâmetro *indent*. Por exemplo: `http://usp.seeoer.com/?q=ciencia&wt=php&rows=600&indent=true`, como mostrado na Figura 6.20, a resposta esta indentada para facilitação de identificação do retorno ao usuário.

Além disso, é possível alterar o parâmetro do *start*, podendo iniciá-lo em outro valor. Por exemplo: `http://usp.seeoer.com/?q=ciencia&wt=php&rows=600&indent=true&start=100`, como é mostrado na Figura 6.21. Neste caso, os resultados de 99 documentos anteriores não são retornados.

Os **parâmetros válidos** para formatos de respostas são:

- Formato Ruby ⇒ wt=ruby



```
array(
  'responseHeader'=>array(
    'status'=>0,
    'QTime'=>2,
    'params'=>array(
      'indent'=>'true',
      'start'=>'100',
      'q'=>'ciencia',
      'wt'=>'php',
      'rows'=>'600')),
  'response'=>array('numFound'=>695,'start'=>100,'docs'=>array(
      array(
        'content'=>'Eduteka - Recursos > Contenido General > Recursos Recientes > pag: 3 Ingresar Registrarse Recordar Contraseña Quiénes Somos Inicio
Artículos Proyectos Módulos Recursos REDuteka Recursos en Eduteka: Alfabetismo en medios Aprendizaje Visual Arte Blogs Ciencias Naturales Ciencias Sociales
CMI Educación en Tecnología Educación General Educación Religiosa Educación y TIC Estándares Evaluación Formación Docente Herramientas Humanidades
Informática Integración TIC Lengua Castellana y Literatura Lenguas Extranjeras Matemáticas Organizaciones Pensamiento Crítico Portales Educativos
Programación de Computadores Proyectos Colaborativos Proyectos de Clase Software WebQuests Recursos > Contenido General > Recursos Recientes 929 artículos
disponibles | página 3 de 62 Si desea referenciar está página use: http://edtk.co/Z1mPh Dipity Dipity es una herramienta en línea, gratuita, para elaborar
líneas de tiempo. Con ella, los estudiantes pueden organizar contenidos por fecha y hora; crear y compartir líneas de tiempo interactivas; integrar en estas
video, audio, imágenes, textos, enlaces y ubicaciones físicas. Publicado: 2011-10-25 Etiquetas: Software Software en Línea Web 2.0 Organizadores Gráficos
Guía de instrumentos musicales Software que permite una aproximación teórica y auditiva a 52 instrumentos musicales. La navegación se puede realizar desde
las categorías: percusión, cuerdas, maderas, bronces y solistas o alfabeticamente. Cada instrumento es acompañado de uno o más ejemplos musicales.
Publicado: 2011-10-25 Etiquetas: Software Software Descargable Música MecaNet Con MecaNet se comienza practicando repetitivamente movimientos que combinan
pulsaciones de teclas hasta tenerlos totalmente asimilados. Después, cuando se ha adquirido destreza con los dedos se pasa a las palabras y más tarde a los
textos. Publicado: 2011-10-25 Etiquetas: Software Software Descargable Teclado Informática AnimatLab Programa que permite diseñar y ejecutar simulaciones
biomecánicas y redes neurales (es posible crear el modelo simplificado de un animal -con sus músculos y sistema nervioso- y soltarlo en un mundo virtual
dotado de suelo, agua y obstáculos). El objetivo de AnimatLab es que los estudiantes comprendan cómo se mueven los animales en su habitat natural y
modifican su conducta a partir de los estímulos del ambiente. La página de AnimatLab ofrece documentación y más de cincuenta vídeo-tutoriales que explican
los fundamentos del programa y presentan ejemplos. El resultado final se parece mucho a un robot hecho con Lego Mindstorms. Publicado: 2011-10-25 Etiquetas:
Software Software Descargable Biología Manual de Scratch para computadores XO (PDF) Manual de la versión de Scratch que viene instalada en los computadores
XO de OLPC. Incluye lo básico (cómo abrir la aplicación), cómo utilizar el entorno de trabajo y una reseña de los comandos de Scratch. manual elaborado por
Telmex. Publicado: 2011-10-25 Etiquetas: Programación Scratch Alt 1040 Sitio Web con información actualizada sobre cultura Geek: redes sociales, tecnología
y ciencia. Publicado: 2011-10-25 Etiquetas: Herramientas Revistas Noticias TIC Libro NAP: The Teacher Development Continuum in the United States and China
El propósito de este trabajo consiste en examinar la estructura de la profesión docente en el campo de las matemáticas en los Estados Unidos y China. Las
principales presentaciones y debates se resumen en este volumen. Publicado: 2011-10-25 Etiquetas: Formación Docente Libros NAP Blog de Carme Barba profesora
catalana creadora del blog de MESTR@ A MESTR@-2 en el que presenta recursos, propuestas, contenidos y metodologías que favorecen el desarrollo de las
competencias básicas, el trabajo cooperativo y la investigación. Así como proyectos y estrategias diversas para construcción de conocimiento. Publicado:
2011-10-25 Etiquetas: Educación y TIC Portales Educativos Blogs Privados España Ciencia, Tecnología, Sociedad e Innovación para el desarrollo sostenible
```


Figura 6.21: Retorno da API do SeeOER no formato de saída PHP com os 600 primeiros resultados iniciados em 100

- Formato Python ⇒ wt=python
- Formato JSON ⇒ wt=json
- Formato PHP ⇒ wt=php

Qualquer outro formato que não estejam contido nessa lista acima e que forem colocadas neste parâmetro, a saída será mal formatada.

6.7 Considerações Finais

Neste capítulo foi abordado o funcionamento do SeeOER, um mecanismo de busca na Web por REA. Foram descritas as diretrizes de desenvolvimento do projeto. Além disso, é fundamental salientar que muitos mecanismos de busca na Web são poucos publicados em revistas, periódicos, entre outros meios científicos. Muitas vezes, eles são restritos a grandes empresas como Microsoft (Mecanismo de Busca na Web Bing), Alphabet (Mecanismo de Busca na Web Google), entre outros. O detalhamento de cada componente do mecanismo de busca na Web especificamente por REA foram feitas nas seções seguintes a diretrizes, as quais são a arquitetura em uma visão geral e abstrata, o *crawler*, o uso da procedência, a integração de esquemas, a indexação e a consulta. No próximo capítulo são descritos os experimentos realizados e os resultados.

Capítulo

7

Análise Experimental

No Capítulo 5 foram descritos trabalhos correlatos voltados à proposta de mecanismos de busca genéricos na Web, de mecanismos de busca na Web especificamente por REA, de componentes específicos para o tratamento de recuperação de REA na Web e de procedência de dados e sistemas de integração. Para cada trabalho correlato, foram destacadas as suas limitações, as quais motivaram o desenvolvimento deste trabalho de mestrado.

Neste capítulo são descritos os experimentos e os resultados obtidos para validação da arquitetura de um mecanismo de busca na Web por recursos educacionais abertos.

Foram realizados experimentos comparativos com os trabalhos correlatos e o mecanismo de busca na Web por REA desenvolvido. Também foram realizados experimentos limitados por fronteira para estimar indexação de REA, metadados identificados e indexados, alterações no REA usando procedência informativa, e estimativa de localidades de *remix* de materiais por autores usando também procedência. Foi considerado como **procedência informativa** aquele dado pré-informado pelo autor ou pelo repositório que é um dado de procedência e que faz parte do metadados.

A forma de validação em mecanismos de busca na Web, como um todo, e de forma quantitativa e específica por componentes são feitas em grau de especialidade e podem ser fundamentadas pelos trabalhos de Bissell et al. (2009); Hogan et al. (2011); de Santiago e Raabe (2010), em que cita suas particularidades e trata sua validação a partir disto. Os experimentos quantitativos do SeeOER se basearam nesses fundamentados de validação.

7.1 Crawler

O rastreamento de REA na Web foi feito por meio do *Crawler* desenvolvido especificamente para esse fim, como descrito na Seção 6.3.

Para o experimento foi usado o *Crawler* Nutch e o *Crawler* do SeeOER. Os outros trabalhos correlatos não disponibilizam seus *Crawlers* para comparação com o SeeOER. Foi utilizado o *crawler* Nutch [37] sabendo-se de que ele é *open-source* e bastante citado por diversos autores.

Para validação do experimento foi usado uma fronteira e considerados os REA indexados e seus metadados. A Figura 7.1 mostra a quantidade de metadados recuperados do repositório *http://cnx.org*, também conhecido como Connexions. Todos os experimentos foram realizados na mesma máquina, ou seja, com configurações idênticas. É possível observar que o Nutch não conseguiu obter nenhum padrão de metadados e o *Crawler* do SeeOER obteve 1568 REA. Um estudo mais aprofundado do Nutch foi identificado que é possível obter apenas metadados associados ao arquivo (nome, tamanho, etc) por meio do Tika juntamente com o Nutch. Mas, não foi identificado uma forma para obtenção dos padrões de metadados. Os trabalhos correlatos citado nesta dissertação não disponibilizam seus *Crawlers ou rôbos* para comparação.

Na Figura 7.2 são mostrados por meio de um gráfico a quantidade de documentos obtidos pelo *Crawler* Nutch na primeira coluna, e na segunda coluna o *Crawler* SeeOER. Foram considerados o mesmo ambiente e fronteira citado no experimento anterior. Enquanto, o Nutch conseguiu indexar 153 documentos, o *Crawler* SeeOER conseguiu indexar 1568 documentos. Foram analisados os resultados indexados pelo Nutch, e observou que em sua maioria estavam apenas páginas de informações com pouca ou nenhuma informação sobre os REA. Enquanto, o *Crawler* SeeOER conseguiu obter REA e seus metadados.

7.2 Reutilização de REA

Foi realizado um experimento controlado de reutilização de REA com os dados armazenados. Foi desenvolvido um *script* para o experimento. Esse *script* obtêm dos metadados integrados dos REA, a data de publicação e a última atualização feita. Com isso, foi possível obter a diferença entre as duas datas. Essa diferença foi convertida em dias e foi considerada como *quantidade de dias* entre a publicação do documento e a atualização

[37] *Open source* - versão 1.7

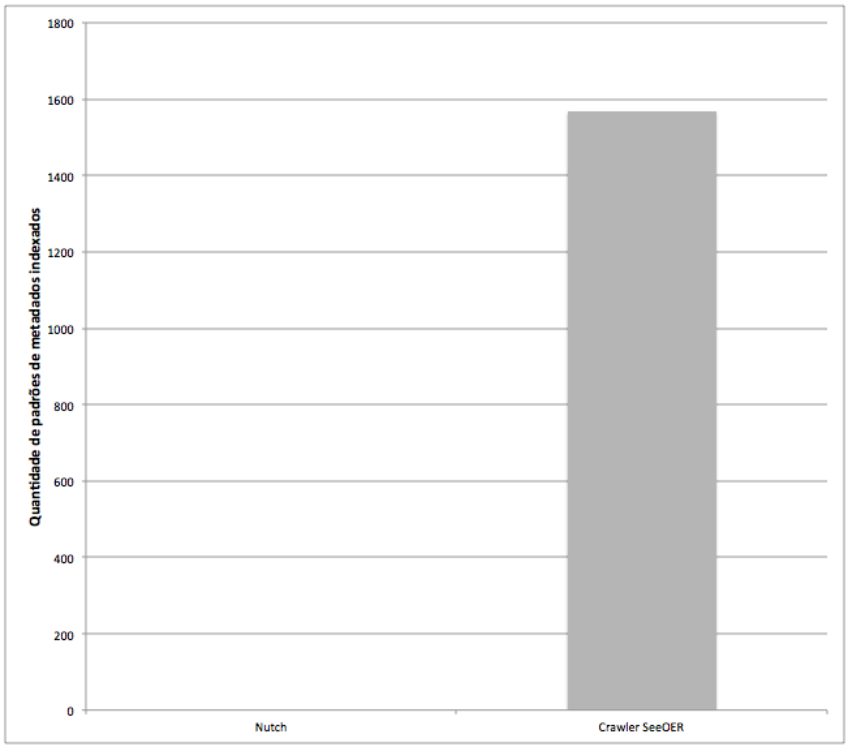


Figura 7.1: Resultado comparativo entre o Crawler proposto e o Nutch

do mesmo REA. Após isso, foi realizado uma contagem da quantidade de REA e foram associados com a diferença de dias por meio de uma tabela hash.

As datas não seguem um padrão de data compatível. Então, foi necessário converter *string* para um formato de data compatível. Segue na Figura 7.3 o formato de data instanciado e o formato convertido.

Na Figura 7.5 seguem alguns metadados de procedência pré-tratados para data e utilizados antes de serem usados no *script*. Pré-tratados quer dizer que os metadados foram obtidos do repositório com os identificadores dos REA e demais informações que não eram necessárias para esse experimento.

O *script* desenvolvido para o experimento segue na Figura 7.4. O *script* fez o tratamento dessas datas e foram tratados os dados que seriam necessário para o experimento.

A Figura 7.6 exibe os resultados obtidos por meio do experimento realizado. Ou seja, foi possível observar com esse experimento o uso e reuso com modificações nos REA ao passar do tempo. É possível observar no gráfico uma curva exponencial da quantidade de dias paralela com a quantidade de REA.

Na Figura 7.7 é possível observar o comportamento do experimento de outra forma. Os pontos da quantidade de dias pelos pontos da quantidade de REA. Os valores variam de 0, no centro, até a extremidade 5000. É possível observar nesse gráfico a variação da

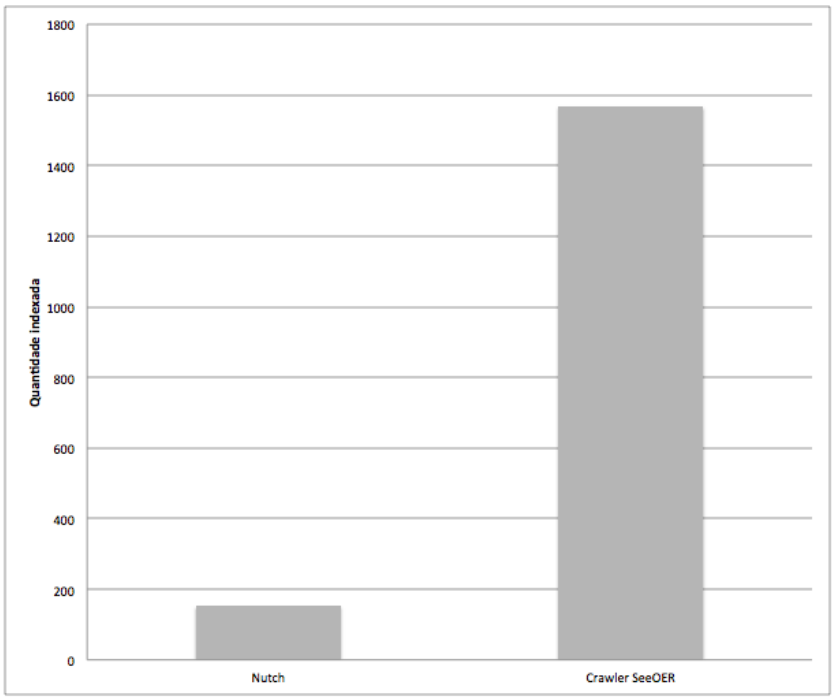


Figura 7.2: Resultado de indexação comparativo entre o SeeOER e os outros mecanismos de busca

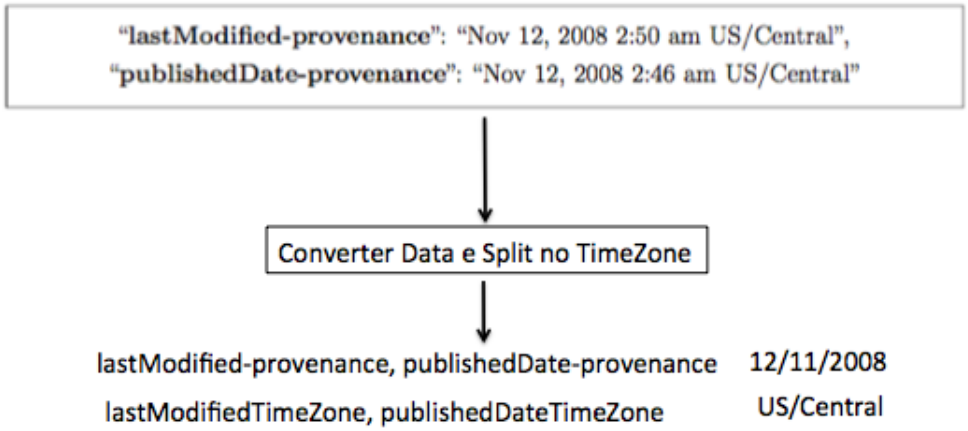


Figura 7.3: Conversão de datas

quantidade de dias entre a publicação do documento e sua atualização com a quantidade de REA.

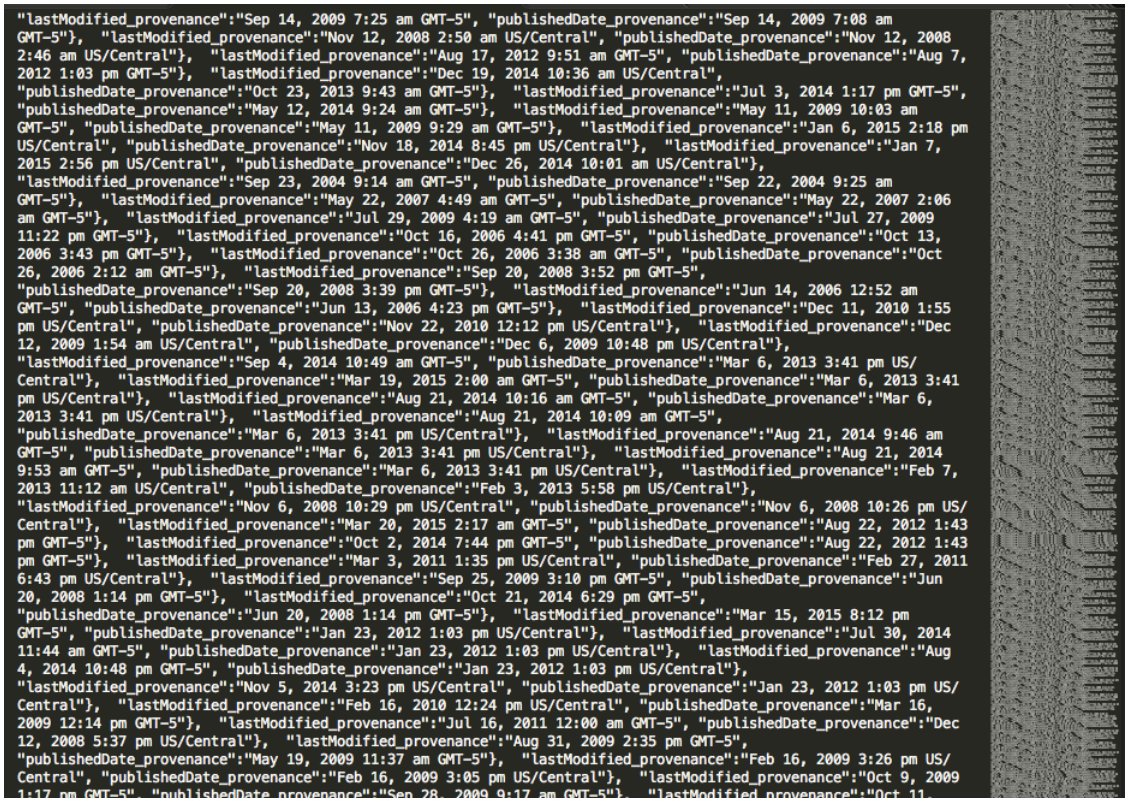


Figura 7.4: Procedência de dados com uso e reuso de REA

7.3 Experimento com Procedência

O experimento de procedência baseou-se na integração de instâncias. Além de integrar os esquemas dos padrões de metadados no início da arquitetura, foi integrado para esse experimento as instâncias do *TimeZone* para verificar se os REA foram criados por uns e reutilizados com outros que não pertenciam ao seu *TimeZone*. Neste caso, o *TimeZone* está representando de forma abstrata e continental do que regional e específico.

Na Figura 7.8 são mostrados os metadados de *TimeZone* dos REA publicados. O gráfico também considera o *TimeZone* dos republicados. Foram encontrados 3 *TimeZone* diferentes. No US/Central 35%, o maior com 65,74% do GMT-5 e por fim 0,26% do GMT-0.

A Figura 7.9 mostra o resultado do experimento. Além disso, mostra que 19% dos REA foram alterados por *TimeZone* diferentes, quer dizer que provavelmente não foi o primeiro publicador do REA que fez a alteração e republicou, e sim, outro usuário. Enquanto que 81% dos REA republicados tiveram origem e edição com republicação no mesmo *TimeZone*.

```
from datetime import date
from seeoer import list_proc
#analisa as datas de procedencia

def diff_datas(d1,d2):
        return abs(d2-d1).days

meses=["Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"]
fusos_horarios_acc=[]
cont=0
cont_f_diff=0
diff_list=[]
calculo=[]
for item_duplo in list_proc:
        item_0_split=item_duplo[0].replace(",","").split(" ")
        item_1_split = item_duplo[1].replace(",","").split(" ")

        #antes
        mes=item_0_split[0]
        mes_number=meses.index(mes)+1 #comeca em zero
        dia=item_0_split[1]
        ano=item_0_split[2]
        fuso_horario=item_0_split[5]

        #depois
        mes2=item_1_split[0]
        mes2_number=meses.index(mes2)+1 #comeca em zero
        dia2=item_1_split[1]
        ano2=item_1_split[2]
        fuso_horario2=item_1_split[5]


        if fuso_horario!=fuso_horario2:
                cont_f_diff+=1
        fusos_horarios_acc.append(fuso_horario)
        fusos_horarios_acc.append(fuso_horario2)

        d1=date(int(ano),int(mes_number),int(dia))
        d2=date(int(ano2),int(mes2_number),int(dia2))
        res=diff_datas(d2,d1)
        diff_list.append(res)
```

Figura 7.5: Script do experimento

7.4 Resultados do SeeOER

Nesta seção são apresentados os resultados do SeeOER. Na Seção 7.4.1 os resultados gerais dos experimentos. E, na Seção 7.4.2 os resultados de busca utilizando *strings* de busca para medição da qualidade.

7.4.1 Resultados Gerais

Os resultados gerais mostram uma comparação entre o SeeOER e os trabalhos correlatos a partir dos dados publicados por meio de seus trabalhos. A Figura 7.10 exibe os resultados obtidos pelo SeeOER. Pode ser observado que o SeeOER obteve 23.618 que

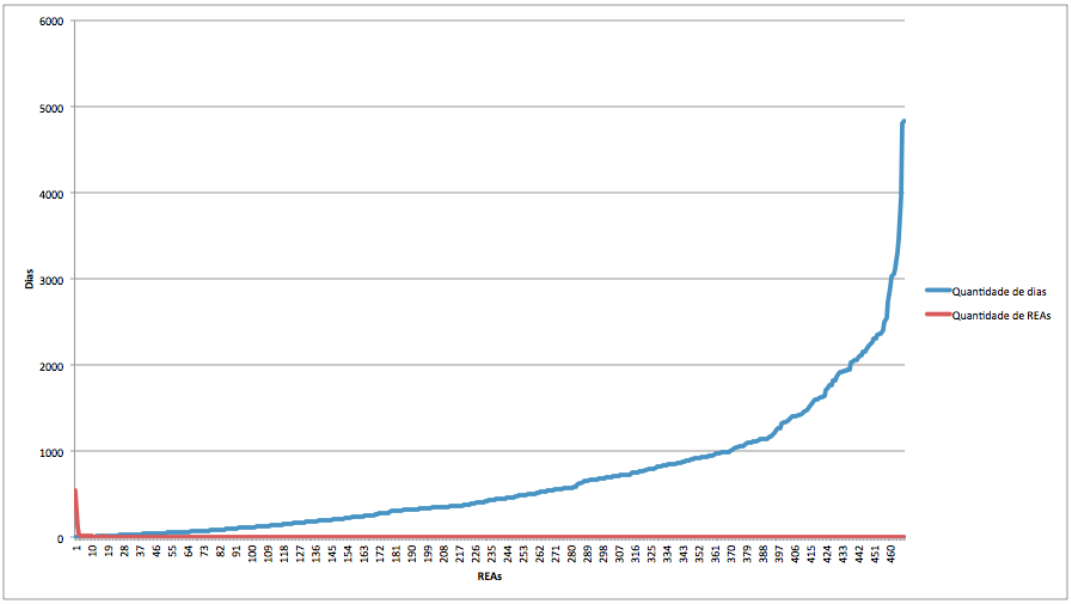


Figura 7.6: Resultado da reutilização com modificações de REA

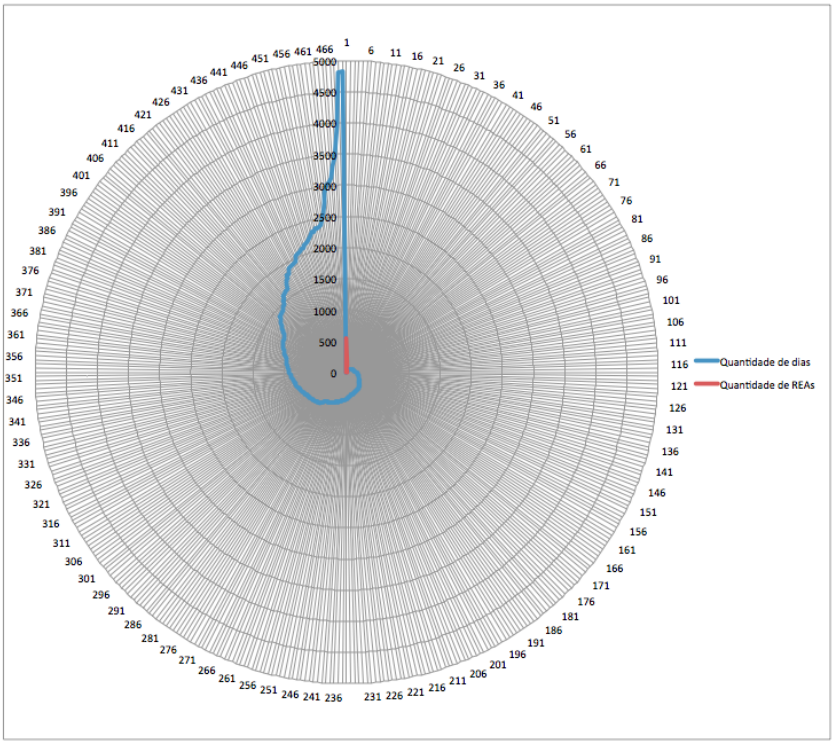


Figura 7.7: Resultado da reutilização com modificações de REA de forma espiral

representa a maior quantidade de REA indexados em comparação com os outros trabalhos correlatos. O OERScout obteve a menor quantidade de REA indexados, com 1.999 REA. Em seguida, o Jorum obteve a segunda maior quantidade de REA indexados, com

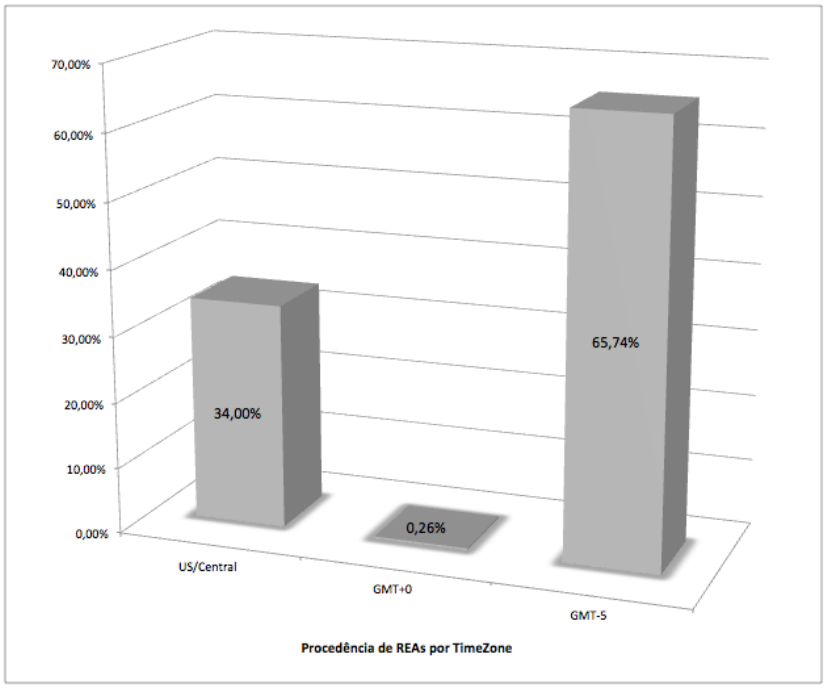


Figura 7.8: Resultado de localidade

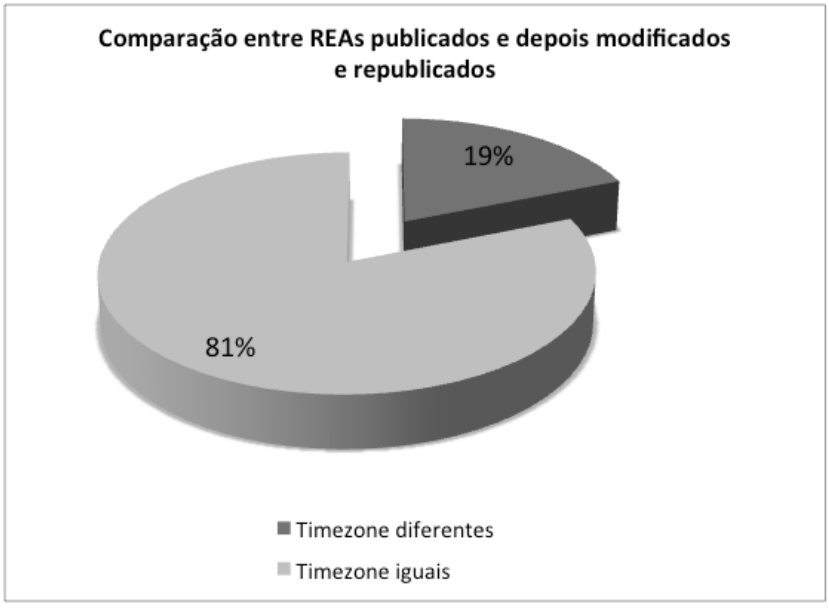


Figura 7.9: Resultado do *TimeZone* dos REA considerando a procedência de dados

15.955 REA. Atualmente, o Jorum é um mecanismo de busca na Web disponível para o público em geral, além de publicar como foi desenvolvido seu mecanismo de busca na Web (Gazzola et al., 2014).

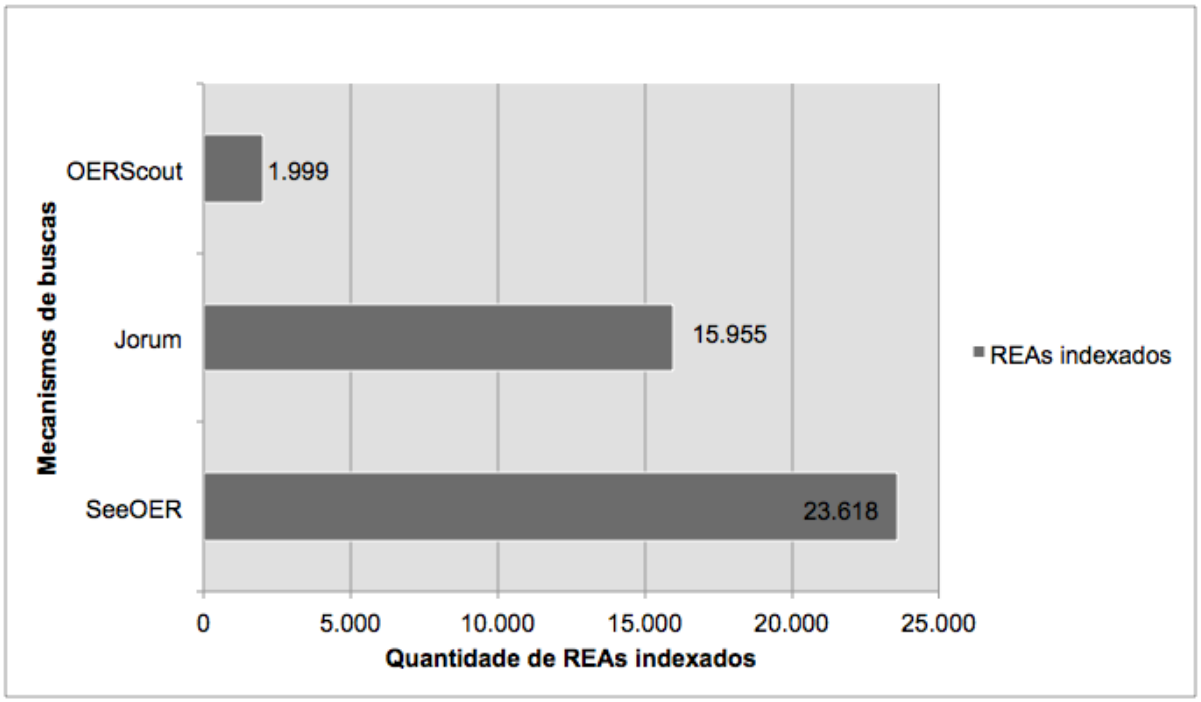


Figura 7.10: Resultado de indexação comparativo entre o SeeOER e os outros mecanismos de busca

7.4.2 Qualidade inicial dos resultados do SeeOER

Para validação a metodologia usada foi baseada em Boudreau et al. (2001); Chae et al. (2002); Fogarty et al. (2001). O **grupo de participantes** foram 10 participantes de 3 universidades distintas. Um grupo de exatas da Universidade de São Paulo, que está na pós-graduação. E, o outro grupo de exatas da Universidade Federal de São Carlos. O grupo de biológicas, composto por alunos da graduação da Universidade Estadual de Maringá. O último grupo, do ensino médio, foram participantes que ainda não realizaram o ensino superior. Os **materiais** utilizados foram um questionário de 10 perguntas essenciais e 3 perguntas informativas. As 10 perguntas foram divididas em 5 classificações. A primeira classificação é sobre conhecimentos básicos de informática, relativos ao cotidiano e inclusos em qualquer área de conhecimento. A segunda classificação é sobre conhecimentos da matemática, inclusa nos tópicos requeridos de conhecimento dos cursos de exatas. A terceira classificação é sobre conhecimentos avançados de matemática e física, inclusa em tópicos avançados na exatas. A quarta classificação é um tópico exploratório, podendo ser incluso na área de biológicas, mas não excludente. A quinta classificação é um tópico específico da área de biológicas, e inclusa em tópicos na área de biológica e não incidentes em área de exatas.

A medição da qualidade de resultados de consultas em um mecanismo de busca na Web por REA pode ser muito subjetivo, por isso foram realizados experimentos usando expressões de consultas e os resultados foram comparados com o principal trabalho correlato e o qual está disponível na Web. Para cada documento buscado m_x e m_y (m = mecanismo de busca na Web por REA) foi realizados a $f(m_x, m_y)$. O documento vencedor é aquele que possui a $f(m_x, m_y) > 0$ e $\Sigma m_\lambda > 0$.

Se $\Sigma m_1 > 0$, então:
$f(m_1, m_2) = \Sigma m_1 + \Sigma m_2 * -1$
Senão, se $\Sigma m_2 > 0$, então:
$f(m_1, m_2) = \Sigma m_1 * -1 + \Sigma m_2$

As notas foram baseadas na escala que segue na Tabela 7.1. Essa escala penaliza mecanismos de busca que não foram considerados em nível de satisfação mínima pelo usuário. O mínimo é o regular, onde o valor é nulo, e o mecanismo de busca não perde e também não ganha nenhum ponto. Abaixo do mínimo o mecanismo de busca é penalizado. E, superior ao mínimo o mecanismo é favorecido por pontos dados pelo usuário. Numa base comparativa, os pontos negativos são somados para o outro mecanismo de busca. A base comparativa leva em consideração que o usuário poderia estar usando o outro mecanismo de busca em invés do mecanismo desfavorecido. Os testes levam em consideração a base sem comparação e com comparação.

Notas representativas	Notas numéricas
Ótimo	2
Bom	1
Regular	0
Ruim	-1
Muito ruim	-2

Tabela 7.1: Escala usada na tabela de notas

A Tabela 7.2 mostra a organização realizada por grupos. Também foram considerados o nível do inglês dos participantes. O mínimo para realização do experimento foi o inglês básico. Também, para que não houve uma discrepância dos resultados, os participantes selecionados foram baseados em um nível de normalização simplificada, onde houvesse uma intercalação entre básico e intermediário ou avançado, e intermediário ou avançado.

Esse experimento foi elaborado considerando a efetividade do mecanismo de busca sem modificações de configurações posteriores, a facilidade para responder o questionário

Grupo	Quantidade de Pessoas	Nível de inglês	Quantidade de Pessoas
Pós-graduação	4	Intermediário	2
Exatas		Avançado	2

Grupo	Quantidade de Pessoas	Nível de inglês	Quantidade de Pessoas
Graduação	2	Básico	1
Biológicas		Intermediário	1

Grupo	Quantidade de Pessoas	Nível de inglês	Quantidade de Pessoas
Graduação	2	Básico	1
Exatas		Intermediário	1

Grupo	Quantidade de Pessoas	Nível de inglês	Quantidade de Pessoas
Ensino Médio	2	Básico	1
Nenhuma		Intermediário	1

Tabela 7.2: Grupo de pessoas distintas organizadas

pelos grupos de pessoas abrangendo várias áreas do conhecimento e não apenas uma área, o fato que mais de 78% das pessoas após o retorno da consulta apenas clicam no primeiro resultado da busca (Eysenbach e Köhler, 2002), a qualidade dos REA e não a análise de ranqueamento dos resultados. Por esses motivos, o experimento consistiu em consultar a *string de busca* no SeeOER e no Jorum e usar essas características.

Abaixo seguem os documentos e suas análises.

Documento 1 - Introdução à Internet

A expressão de busca usada foi *Intro to Internet* para os dois mecanismos de busca na Web por REA, o SeeOER e o Jorum. A expressão de busca usada leva em consideração as palavras *Intro* e *Internet* e uma *stop word* (*to*). Além disso, a busca pretende buscar pela história da Internet, seu surgimento, relevância, funcionamento de forma simplificada. No questionário foi usado o termo em português para facilitar o entendimento das pessoas que leriam o questionário e por não afetar seu entendimento com o resultado de busca. Porém, o resultado foi mantido o original obtido pelo mecanismo de busca. Nessa expressão de busca, *Intro to Internet*, o **Jorum** encontrou o seguinte resultado:

Identifying opportunities and using an interactive whiteboard with learners. Authors: Geoff Foot. Created: 3 January 2008, by Geoff Foot This collection of resources offers a guide to using an interactive whiteboard in teaching. It includes instructions on switching on the whiteboard, suggested ways of using the whiteboard with learners and example resources in early years, key skills and construction education. This resource was funded by the LSN as part of the Q projects programme. The resource was designed and produced by South Devon College. Published: 23 February 2011, by Learning and Skills Network Ltd, Geoff Foot Keywords: digital, presentations, whiteboards, learning, teaching, interactive Licence: Attribution-Noncommercial-Share Alike 2.0 UK: England and Wales

Abaixo são detalhados os metadados os metadados obtidos pelo Jorum. O título do material (*title*), resumo (*summary*) e os outros metadados.

Documento 1 - Jorum

Title: *Identifying opportunities and using an interactive whiteboard with learners*
Authors: *Geoff Foot*
Created: *3 January 2008*
Last Published: 23 February 2011
Keywords: *digital, presentations, whiteboards, learning, teaching, interactive*
Licence: *Attribution-Noncommercial-Share Alike 2.0 UK: England and Wales*
Summary: This collection of resources offers a guide to using an interactive whiteboard in teaching. It includes instructions on switching on the whiteboard, suggested ways of using the whiteboard with learners and example resources in early years, key skills and construction education. This resource was funded by the LSN as part of the Q projects programme. The resource was designed and produced by South Devon College.

No **SeeOER**, usando a mesma expressão de busca o primeiro resultado obtido segue abaixo. É possível fazer uma correlação a expressão de busca com o *Title* em ***Intro to Internet*** e as *Keywords* em ***internet, technology***.

Documento 1 - SeeOER

Title: *This is a basic intro to the Internet.*
Published: *12 Oct 2007.*
University: *University Psu.*
Author: *Cole Camplese.*
Keywords: *internet, technology.*
Type: *Course.*
License: *Creative Commons Attribution Licence CC-BY 2.0*

O resultado, por meio do questionário, segue na Tabela 7.3. E, na Figura 7.11 segue uma contagem comparativa dos resultados. É possível verificar que no *Documento 1* o Jorum obteve um ponto positivo e todos negativos. Enquanto, o SeeOER obteve nenhum valor negativo, três regulares (nulos) e o restantes positivos. O resultado foram 12 pontos positivos para o SeeOER e 14 pontos negativos para o Jorum. O SeeOER foi classificado com **26 pontos** superior ao Jorum, neste **Documento 1**, considerando a escala do experimento.

Nível escolaridade	Área	Inglês	Jorum		SeeOER	
Pós-graduação	Exatas	Avançado	Muito Ruim	-2	Ótimo	2
Pós-graduação	Exatas	Intermediário	Ruim	-1	Regular	0
Pós-graduação	Exatas	Avançado	Muito Ruim	-2	Ótimo	2
Pós-graduação	Exatas	Intermediário	Muito Ruim	-2	Ótimo	2
Graduação	Biológicas	Básico	Bom	1	Regular	0
Ensino Médio	Nenhuma	Básico	Muito Ruim	-2	Ótimo	2
Graduação	Exatas	Básico	Ruim	-1	Bom	1
Graduação	Exatas	Intermediário	Muito Ruim	-2	Bom	1
Ensino Médio	Nenhuma	Intermediário	Muito Ruim	-2	Ótimo	2
Graduação	Biológicas	Intermediário	Ruim	-1	Regular	0

Tabela 7.3: Documento 1 - Questionário

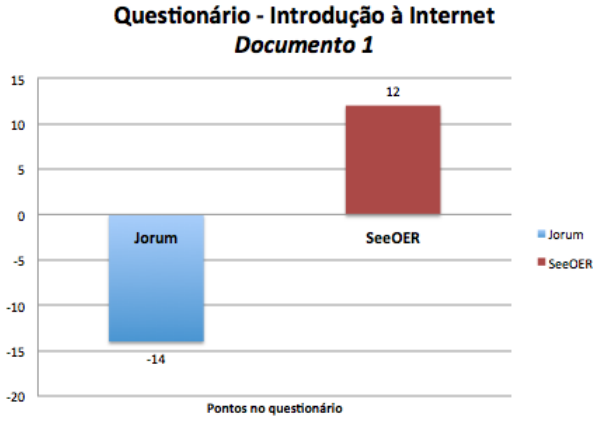


Figura 7.11: Documento 1 - Resultado do questionário

Documento 2 - Álgebra I

No Documento 2, foi usada a expressão de busca *Algebra I* para os dois mecanismos de busca na Web por REA. A consulta leva em consideração uma palavra (*Algebra*) e um numeral (*I*). É possível que alguns mecanismos de buscas na Web desconsiderem esse numeral. Além disso, a busca pretende buscar por REA relacionados ao estudo de Álgebra, podendo encontrar em Álgebra que inclui princípio da indução matemática, MDC, entre outros tópicos. Apenas no questionário foi usado o termo em português Álgebra I. Nessa expressão de busca o **Jorum** encontrou o seguinte resultado:

> *The First Law of Thermodynamics and Enthalpy In this video, I shall present the first law of thermodynamics, and show how it governs the state function internal energy, and how considerations of the first law for isochoric and isobaric processes leads to the definition of the further state function called enthalpy. Published: 26 February 2015, by University of Manchester. Authors: Dr. Jonathan Agger. Keywords: virtual experiment, virtual lab, chemistry, Thermodynamics, physical chemistry, School of Chemistry, Faculty of Engineering and Physical Sciences, University of Manchester.Created: 1 June 2014, by Dr. Jonathan Agger. Contributor(s): Stephen Wheeler, Ian Hutt, Prof. Michael Anderson, Dr. Patrick O'Malley. Licence: Attribution-NonCommercial-ShareAlike 4.0 International*

O resultado do Jorum, segue com seus metadados. No *title*, inicialmente, não é possível inferir nenhuma relação com a expressão de busca. No *summary* e *keywords*, também, não existem nenhum *matching* com a *string* de busca.

Documento 2 - Jorum

Title: The First Law of Thermodynamics and Enthalpy
Summary: In this video, I shall present the first law of thermodynamics, and show how it governs the state function internal energy, and how considerations of the first law for isochoric and isobaric processes leads to the definition of the further state function called enthalpy.
Published: 26 February 2015.
University: University of Manchester.
Author: Dr. Jonathan Agger.
Keywords: virtual experiment, virtual lab, chemistry, Thermodynamics, physical chemistry, School of Chemistry, Faculty of Engineering and Physical Sciences, University of Manchester.
Created: 1 June 2014.
Licence: Attribution-NonCommercial-ShareAlike 4.0 International

Em relação ao **SeeOER**, segue o resultado obtido. É possível observar que o metadado do resumo (*summary*) não foi atribuído pelo criador do REA. Porém, o título (*title*) e a sub-coleção (*subcollection*) existem, e há uma possível relação com a expressão de busca em **Algebra I** no *title* e na sub-coleção em ***Mathematics***.

Documento 2 - SeeOER

Title: Algebra I for the Community College.
Subcollection: Mathematics and Statistics.
Language: English (en).
Type: Textbook.
Created: Oct 23, 2013 9:43 am GMT-5.
Author: Ann Simao (afsimao@stcc.edu).
Licence: Creative Commons Attribution CC-BY 4.0.

No questionário segue o resultado na Tabela 7.4. E, na Figura 7.12 segue o resultado comparativo. É possível observar no *Documento 2* que Jorum obteve nenhum ponto

positivo, um nulo e todos negativos. Enquanto, o SeeOER obteve nenhum valor negativo, dois regulares (nulos) e o restante positivos. O resultado foram 12 pontos positivos para o SeeOER e 15 pontos negativos para o Jorum. O SeeOER foi classificado com **27 pontos** a mais do que o Jorum, neste **Documento 2**.

Nível escolaridade	Área	Inglês	Jorum		SeeOER	
Pós-graduação	Exatas	Avançado	Muito Ruim	-2	Bom	1
Pós-graduação	Exatas	Intermediário	Ruim	-1	Regular	0
Pós-graduação	Exatas	Avançado	Muito Ruim	-2	Ótimo	2
Pós-graduação	Exatas	Intermediário	Muito Ruim	-2	Ótimo	2
Graduação	Biológicas	Básico	Ruim	-1	Bom	1
Ensino Médio	Nenhuma	Básico	Muito Ruim	-2	Ótimo	2
Graduação	Exatas	Básico	Muito Ruim	-2	Ótimo	2
Graduação	Exatas	Intermediário	Ruim	-1	Regular	0
Ensino Médio	Nenhuma	Intermediário	Muito Ruim	-2	Bom	1
Graduação	Biológicas	Intermediário	Regular	0	Bom	1

Tabela 7.4: Documento 2 - Questionário

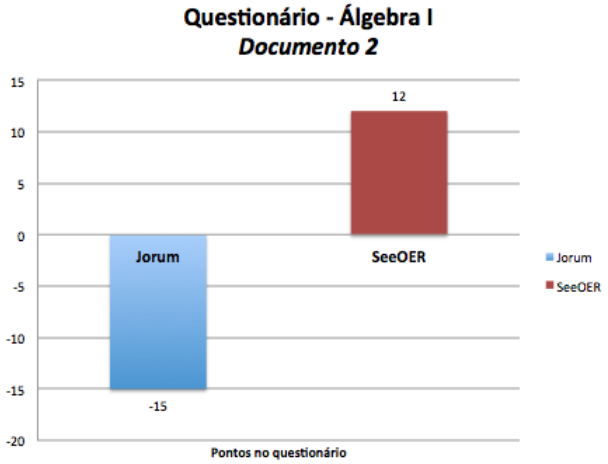


Figura 7.12: Documento 2 - Resultado do questionário

Documento 3 - Álgebra e Física

Foi usada a expressão de busca *Algebra and Physics* para os dois mecanismos de busca na Web por REA. A consulta leva em consideração duas palavras (*Algebra, Physics*) e uma *stop-words* (*and*). Além disso, com a expressão de busca se pretende encontrar REA relacionados a Álgebra e Física, como princípios da cinemática, mecânica Newtonian, impulso e momento linear, entre outros. Apenas no questionário foi usado o termo em português Álgebra e Física. Nessa expressão de busca o **Jorum** encontrou o seguinte resultado:

> *Observation, measurement and the recording of data are central activities in science. Speculation and the development of new theories are crucial as well, but ultimately the predictions resulting from those theories have to be tested against what actually happens and this can only be done by making further measurements. Whether measurements are made using simple instruments such as rulers and thermometers, or involve sophisticated devices such as electron microscopes or lasers, there are decisions to be made about how the results are to be represented, what units of measurements will be used and the precision to which the measurements will be made. In this unit we will consider these points in turn. Published: 15 January 2010, by Open University Keywords: statistics, standard deviation, significant figure, scientific notation, science, sample, probability, normal distribution, measurement, maths, logarithm, integer, decimal, data, average, science and nature Licence: Attribution-Noncommercial-Share Alike 2.0 UK: England and Wales*

O resultado do Jorum, segue de forma especificada com os metadados. É possível observar que existem poucas relações com a expressão de busca. As palavras relacionadas estão nas palavras chaves as quais são: *average*, *decimal*, *science*, *integer*, *normal distribution* e *scientific notation*. Porém, é baixo a relação com que se estava buscando. Os outros metadados, como *title*, *summary* não fazem *matching* ou pouca relação com a expressão de busca.

Documento 3 - Jorum

Title: Observation, measurement and the recording of data are central activities in science.

Summary: Speculation and the development of new theories are crucial as well, but ultimately the predictions resulting from those theories have to be tested against what actually happens and this can only be done by making further measurements. Whether measurements are made using simple instruments such as rulers and thermometers, or involve sophisticated devices such as electron microscopes or lasers, there are decisions to be made about how the results are to be represented, what units of measurements will be used and the precision to which the measurements will be made. In this unit we will consider these points in turn.
Published: 15 January 2010
Keywords: statistics, standard deviation, significant figure, scientific notation, science, sample, probability, normal distribution, measurement, maths, logarithm, integer, decimal, data, average, science and nature
Licence: Attribution-Noncommercial-Share Alike 2.0 UK: England Wales

Segue o resultado obtido no **SeeOER**. É possível observar uma relação com os metadados *Title:* em ***Applied Math and Physics***, *Summary* em ***algebra-based, physics e physics application***, e *Keywords* em ***collisions, elasticity, energy, forces, friction, kinematics, linear momentum, physics, rotational motion and angular momentum, torque, uniform circular motion and gravitation***.

Documento 3 - SeeOER

Title: Introduction to Applied Math and Physics.
Summary: This introductory, algebra-based, one-semester physics book is based on OpenStax College Physics. This online, fully editable and customizable title includes learning objectives, concept questions, links to labs and simulations, and ample practice opportunities to solve traditional physics application problems.
Language: English (en).
Collection: Mathematics and Statistics, Science and Technology.
Keywords: college physics, collisions, elasticity, energy, forces, friction, kinematics, linear momentum, Newton Laws of Motion, physics, rotational motion and angular momentum, statics and torque, uniform circular motion and gravitation, work.
License: Creative Commons Attribution License CC-BY 3.0. Contributor (s): Oka Kurniawan.

Na Tabela 7.5 segue o resultado do questionário. A Figura 7.13 segue o resultado comparativo. É possível observar no *Documento 3* que Jorum obteve apenas um ponto positivo, dois nulos e todos negativos. Enquanto, o SeeOER obteve apenas um valor negativo, um regular (nulo) e o restante positivos. O resultado foram 16 pontos positivos para o SeeOER e 11 pontos negativos para o Jorum. O SeeOER foi classificado com **27 pontos** a mais do que o Jorum, na escala do experimento.

Nível escolaridade	Área	Inglês	Jorum		SeeOER	
Pós-graduação	Exatas	Avançado	Muito Ruim	-2	Ótimo	2
Pós-graduação	Exatas	Intermediário	Regular	0	Bom	1
Pós-graduação	Exatas	Avançado	Regular	0	Ótimo	2
Pós-graduação	Exatas	Intermediário	Muito Ruim	-2	Ótimo	2
Graduação	Biológicas	Básico	Bom	1	Ruim	-1
Ensino Médio	Nenhuma	Básico	Ruim	-1	Ótimo	2
Graduação	Exatas	Básico	Ruim	-1	Ótimo	2
Graduação	Exatas	Intermediário	Muito Ruim	-2	Ótimo	2
Ensino Médio	Nenhuma	Intermediário	Muito Ruim	-2	Ótimo	2
Graduação	Biológicas	Intermediário	Muito Ruim	-2	Ótimo	2

Tabela 7.5: Documento 3 - Questionário

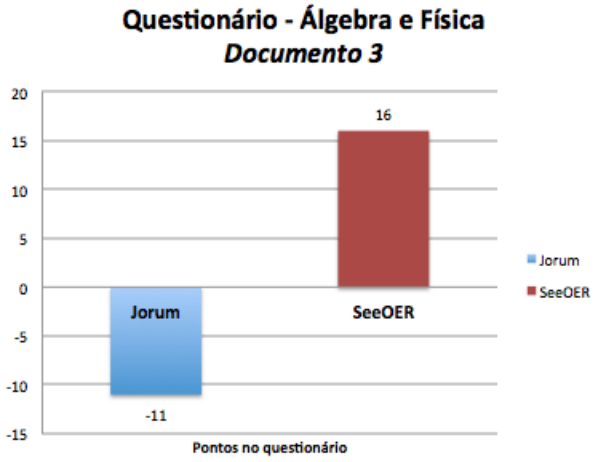


Figura 7.13: Documento 3 - Resultado do questionário

Documento 4 - Biotecnologia

Foi usada a expressão de busca *biotechnology* para os dois mecanismos de busca na Web por REA. A consulta leva em consideração apenas uma palavra (*biotechnology*) e nenhum dado adicional. Se pretende encontrar REA relacionados com biotecnologia, afim de realizar uma busca exploratória sobre o assunto. A busca exploratória é usada como um estudo inicial para que se possa em seguida ter uma maior compreensão sobre o assunto. No questionário foi usado o termo em português, mas no mecanismo de busca foi usado o termo em inglês. Nesta expressão de busca o **Jorum** encontrou de forma idêntica ao resultado seguinte:

> *Final assessment SCORM format Authors: Janet Fyfe Created: 26 March 2008, by Janet Fyfe This resource comprises of interactive material created as part of the CeLLs Project covering the topic of labelled antibodies. This section is a revision self-assessment with multiple-choice questions. Published: 23 February 2011, by The Adam Smith College Fife Keywords: immunoassays, antigens, antibodies Licence: Attribution-Noncommercial-Share Alike 2.0 UK: England and Wales*

Segue abaixo, o resultado do Jorum, de forma estruturada manualmente com os metadados. É possível observar a baixa relação com a expressão de busca. Mesmo o REA possuindo diversos metadados representativos, como *Title*, *Summary* e *Keywords*, não é possível inferir diretamente uma relação com a *string* de busca e o metadados.

Documento 4 - Jorum

Title: *Final assessment SCORM format*
Authors: *Janet Fyfe*
Created: *26 March 2008*
Summary: *This resource comprises of interactive material created as part of the CeLLs Project covering the topic of labelled antibodies. This section is a revision self-assessment with multiple-choice questions.*
Published: *23 February 2011*
Keywords: *immunoassays, antigens, antibodies*
Licence: *Attribution-Noncommercial-Share Alike 2.0 UK: England and Wales*

Segue abaixo, o resultado obtido no **SeeOER**. Os metadados são poucos representativos, o criador do REA não adicionou palavras chaves e o resumo.

Documento 4 - SeeOER

Title: *Biotechnology.*
Language: *English (en).*
Created:*Mar 13, 2014 5:06 pm GMT-5.*
Licence: *Creative Commons Attribution*
License *CC-BY 4.0.*
Author: *Jeffrey Mahr.*

Na Tabela 7.6 segue o resultado do questionário. A Figura 7.14 segue o resultado comparativo. É possível observar no *Documento 4* que Jorum obteve um ponto positivo, dois nulo e todos negativos. Enquanto, o SeeOER obteve um valor negativo, nenhum regular (nulo) e o restante positivos. O resultado foram 8 pontos positivos para o SeeOER e 15 pontos negativos para o Jorum. O SeeOER foi classificado com **23 pontos** superior ao Jorum, neste **Documento 4**, na escala utilizada. Foi o pior resultado do SeeOER em todo experimento.

Nível escolaridade	Área	Inglês	Jorum		SeeOER	
Pós-graduação	Exatas	Avançado	Muito Ruim	-2	Regular	0
Pós-graduação	Exatas	Intermediário	Ruim	-1	Regular	0
Pós-graduação	Exatas	Avançado	Regular	0	Ótimo	2
Pós-graduação	Exatas	Intermediário	Muito Ruim	-2	Ótimo	2
Graduação	Biológicas	Básico	Ruim	-1	Bom	1
Ensino Médio	Nenhuma	Básico	Muito Ruim	-2	Ótimo	2
Graduação	Exatas	Básico	Muito Ruim	-2	Regular	0
Graduação	Exatas	Intermediário	Muito Ruim	-2	Bom	1
Ensino Médio	Nenhuma	Intermediário	Muito Ruim	-2	Regular	0
Graduação	Biológicas	Intermediário	Ruim	-1	Regular	0

Tabela 7.6: Documento 4 - Questionário

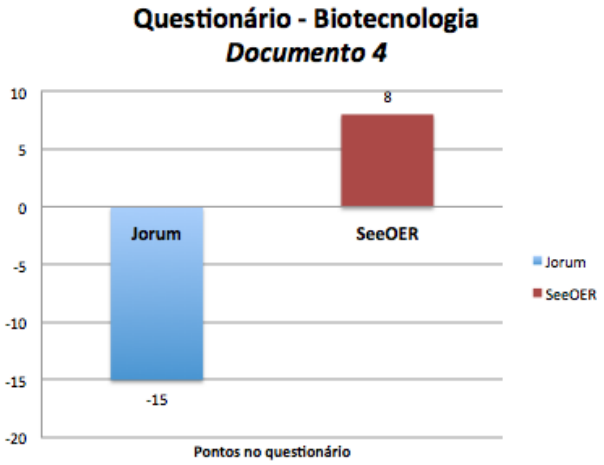


Figura 7.14: Documento 4 - Resultado do questionário

Documento 5 - Fisiologia e Anatomia Humana

A expressão de busca usada no Jorum e no SeeOER foi *Physiology and Anatomy*. A consulta leva em consideração apenas duas palavras possivelmente relevantes (*Physiology, Anatomy*) e uma *stop-word* (*and*). A necessidade de informação é encontrar REA relacionados com anatomia humana e fisiologia. No questionário foi usado o termo em português Fisiologia e Anatomia Humana. Na expressão de busca em inglês o **Jorum** encontrou de forma idêntica ao resultado seguinte:

> *The use of 3D technology to deliver science curriculum Authors: Dumfries and Galloway College, JISC RSC Scotland Created: 20 January 2012, by Dumfries and Galloway College, JISC RSC Scotland 3-D technology can be used to add innovative and dynamically interesting dimensions to the teaching and learning process. The use of 3-d technology in a range of subjects can be used to emphasise, highlight and trigger attention processes to key learning concepts in the classroom. Using a portable active 3-d projector and specific 3-d curriculum packs, the college has begun to integrate 3-d technology to offer the learners a new and exciting way to combine both visual and kinaesthetic preferences to learning with a current focus on trialling its use in plant biology, human anatomy and physiology and sports science. Published: 7 November 2012, by JISC RSC Scotland Keywords: science, 3D technology, Dumfries and Galloway College, case studies Licence: Attribution 3.0 Unported*

Abaixo, segue o resultado do Jorum, com a estrutura dos metadados. É possível observar uma fuga de tema. Porém, uma possível relação seria no *title* em *3D technology*

e *science*, no *Summary* em *human anatomy*. Mesmo assim, são possíveis palavras com possível grau de aproximação.

Documento 4 - Jorum

Title: *The use of 3D technology to deliver science curriculum*
Authors: *Dumfries and Galloway College, JISC RSC Scotland*
Created: *20 January 2012*
Summary: *The 3-D technology can be used to add innovative and dynamically interesting dimensions to the teaching and learning process. The use of 3-d technology in a range of subjects can be used to emphasise, highlight and trigger attention processes to key learning concepts in the classroom. Using a portable active 3-d projector and specific 3-d curriculum packs, the college has begun to integrate 3-d technology to offer the learners a new and exciting way to combine both visual and kinaesthetic preferences to learning with a current focus on trialling its use in plant biology, human anatomy and physiology and sports science.*
Published: *7 November 2012, by JISC RSC Scotland Keywords: science, 3D technology, Dumfries and Galloway College, case studies*
Licence: *Attribution 3.0 Unported*

Segue abaixo, o resultado obtido no **SeeOER**. No *Documento 5* do SeeOER é possível observar uma possível relação visual com a expressão de busca e o *Title* em ***Anatomy, Physiology***, o *Summary*, e as *Keywords*.

Documento 5 - SeeOER

Title: *Anatomy & Physiology: Energy, Maintenance and Environmental Exchange.*
Summary: *Human Anatomy and Physiology is designed for the two-semester anatomy and physiology course taken by life science and allied health students. The textbook follows the scope and sequence of most Human Anatomy and Physiology courses, and its coverage and organization were informed by hundreds of instructors who teach the course. Instructors can customize the book, adapting it to the approach that works best in their classroom. The artwork for this textbook is aimed focusing student learning through a powerful blend of traditional depictions and instructional innovations.*

Color is used sparingly, to emphasize the most important aspects of any given illustration. Significant use of micrographs from the University of Michigan complement the illustrations, and provide the students with a meaningful alternate depiction of each concept. Finally, enrichment elements provide relevance and deeper context for students, particularly in the areas of health, disease, and information relevant to their intended careers.

Language:*English (en).*

Collection:*Science and Technology.*

Keywords: *acid-base balance, action potential, adrenal, anatomy, antibody, appendicular skeleton, atom, autonomic nervous system, axial skeleton, blood, body fluid, bone, capillary, cardiac, cardiovascular system, cell, central nervous system, chemical, circulatory pathway, connective tissue, cytoplasm, development, diet, digestion, digestive system, DNA, electrical activity, electrolyte, element, endocrine, energy, epithelial, erythrocyte, esophagus, exam, fascicle, fertilization, fetal, fluid balance, gallbladder, gland, heart, heat balance, hemostasis, homeostasis, hormone, hypothalamus, immune system, immunity, inheritance, inorganic compound, integumentary system, intestine, joint, kidney, lactation, leukocyte, liver, lung, lymphatic system, lymphocyte, metabolism, motor response, muscle, nervous system, neurological, neuron, nucleus, nutrition, organic compound, pancreas, pathogen, pectoral girdle, pelvis, peripheral nervous system, pharynx, physiology, placental hormone, platelet, protein, renal, reproduction, reproductive system, respiration, respiratory system, sensory perception, skeletal system, skeleton, skin, skull, stomach, thoracic cage, thorax, tissue, urinary system, urine, vascular system, vertebrae, water balance. Type: Textbook. Contributor (s): Wade Hedegard.*

Licence: *Creative Commons Attribution CC-BY 4.0.*

Na Tabela 7.7 segue o resultado do questionário do *Documento 5 - SeeOER*. A Figura 7.15 segue o resultado comparativo entre o SeeOER e Jorum. É possível observar que o Jorum obteve nenhum ponto positivo, três nulos e todos negativos. Enquanto, o SeeOER obteve nenhum valor negativo, nenhum regular (nulo) e o restante positivos. O resultado foram 18 pontos positivos para o SeeOER e 8 pontos negativos para o Jorum. O SeeOER foi classificado com **26 pontos** superior ao Jorum, neste **Documento 5**, na escala utilizada.

Nível escolaridade	Área	Inglês	Jorum		SeeOER	
Pós-graduação	Exatas	Avançado	Muito Ruim	-2	Ótimo	2
Pós-graduação	Exatas	Intermediário	Regular	0	Bom	1
Pós-graduação	Exatas	Avançado	Ruim	-1	Ótimo	2
Pós-graduação	Exatas	Intermediário	Regular	0	Ótimo	2
Graduação	Biológicas	Básico	Ruim	-1	Bom	1
Ensino Médio	Nenhuma	Básico	Regular	0	Ótimo	2
Graduação	Exatas	Básico	Regular	0	Ótimo	2
Graduação	Exatas	Intermediário	Ruim	-1	Ótimo	2
Ensino Médio	Nenhuma	Intermediário	Ruim	-1	Ótimo	2
Graduação	Biológicas	Intermediário	Muito Ruim	-2	Ótimo	2

Tabela 7.7: Documento 5 - Questionário

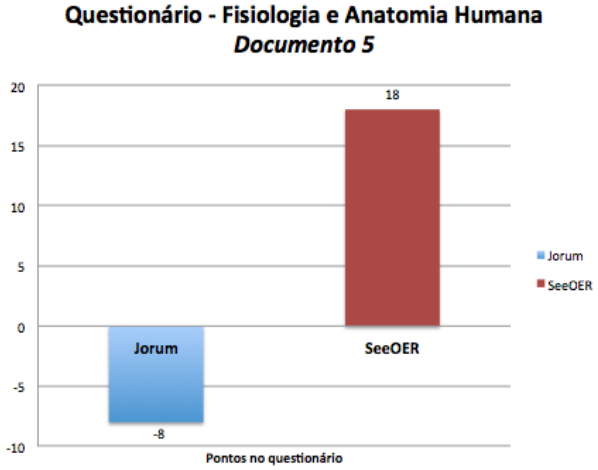


Figura 7.15: Documento 5 - Resultado do questionário

7.4.3 Usando a função score e a função penalizado de modo comparativa

Os resultados dos documentos também foram avaliados e comparados considerando a função score, apresentada nesta seção, e a função penalizado (Seção 7.4.2).

A função score é uma função simples baseada na escala de 0.15 até 0.75 considerando os seguintes atributos e pontuações associados:

- Muito Ruim (0.15)
- Ruim (0.3)
- Regular (0.45)

- Bom (0.6)
- Ótimo (0.75)

A função penalizada, descrita na Seção 7.4.2, é uma função que penaliza os documentos que tiveram respostas ruins e pontua os documentos que tiveram bons resultados. Com isso, foram realizados experimentos comparativos entre a função penalizada e a função score.

A Tabela 7.8 compara os resultados das funções score e penalizada com o mecanismo de busca Jorum. Foram realizados os experimentos em todos os documentos.

String	Título	Meta dados	Keywords	Score	F(mx, my)
Introdução à Internet	Identificando oportunidades e usando uma lousa iterativa com alunos	7	digital, apresentações, lousas, aprendizagem, ensino, iterativo	2.4	-14
Álgebra I	A Primeira Lei da Termodinâmica e entalpia	8	experimento virtual, laboratório virtual, química, termodinâmica, química física	2.25	-15
Álgebra e Física	Observação, medição e registro de dados são atividades centrais em ciências.	5	estatísticas, desvio padrão, figura significativa, notação científica, ciência, probabilidade	2.85	-11
Biotecnologia	Verificação final do formato SCORM	7	imunoensaios, antigénios, anticorpos	2.25	-15
Fisiologia e Anatomia	O uso da tecnologia 3D para integrar o currículo científico	6	sem keywords	3.3	-8

Tabela 7.8: Resultados usando função score e função penalizado com o mecanismo Jorum

A Tabela 7.9 usa os mesmos documentos da análise experimental qualitativa, mas com o mecanismo de busca SeeOER. Os resultados são comparados com as funções score e penalizada. Também foram realizados os experimentos considerando todas características descritas na Seção 7.4.2.

String	Título	Meta dados	Keywords	Score	F(mx, my)
Introdução à Internet	Esta é uma introdução básica à Internet.	7	intenet, tecnologia	6.3	12
Álgebra I	Álgebra I para a Community College	8	matemática e estatística	6.3	12
Álgebra e Física	Introdução à Matemática Aplicada e Física	6	física, faculdade, colisões, elasticidade, energia, forças, fricção cinemática, dinâmica, linear	6.9	16
Biotecnologia	Biotecnologia	6	nenhum keywords	5.7	8
Fisiologia e Anatomia	Anatomia e Fisiologia: energia, manutenção e intercâmbio com o ambiente	6	equilíbrio ácido-base, potencial de ação, adrenal, anatomia, anticorpo, caixa torácica, tórax, tecido	7.2	18

Tabela 7.9: Resultados usando função score e função penalizado com o mecanismo SeeOER

A Figura 7.16 considera de forma comparativa as Tabelas 7.9 e 7.8 dos mecanismos de busca na Web por REA Jorum e SeeOER. É possível observar que o SeeOER obteve uma margem superior em todos os documentos de questionários obtidos pelo experimento realizado.

Considere a função penalizada descrita na Seção 7.4.2 de modo comparativo entre documentos de questionário para a Figura 7.17. É possível observar que o menor resultado obtido pelo SeeOER foi no documento 4 totalizando apenas 8 pontos. Enquanto, o Jorum não obteve nenhum valor maior que zero usando a função penalizada.

A Figura 7.18 considera a função score de modo comparativo entre documentos de questionário. No documento 4 o SeeOER obteve o menor resultado (5,7). Enquanto, o Jorum não obteve nenhum resultado superior a 3,3.

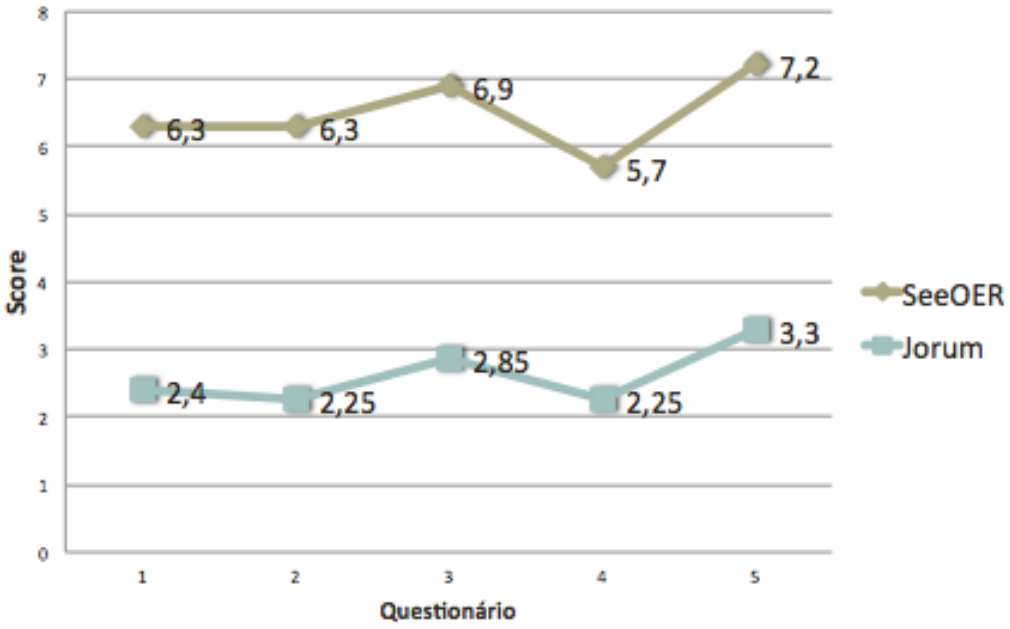


Figura 7.16: Resultados usando a função score

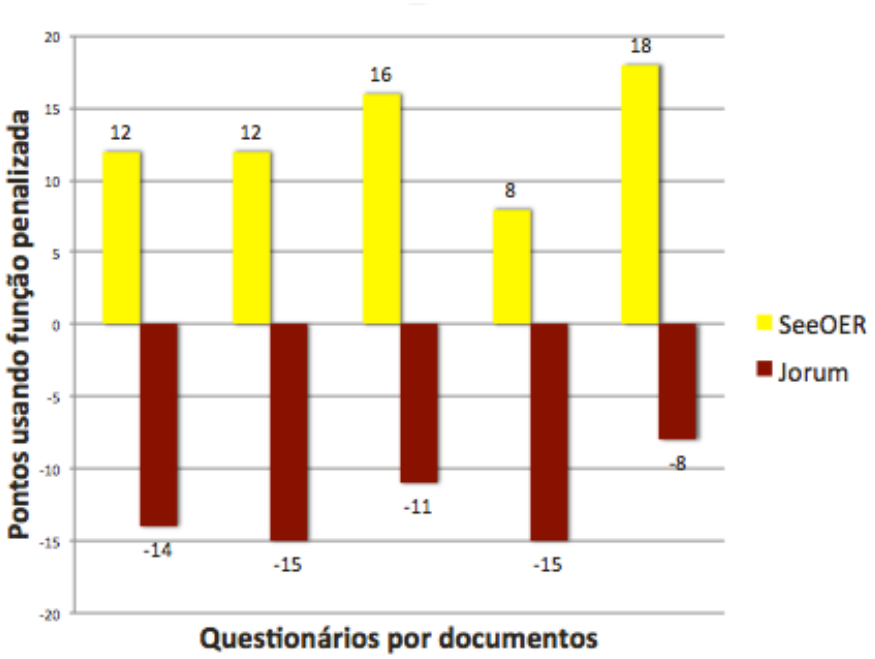


Figura 7.17: Resultado dos questionários de documentos usando a função penalizada de forma comparativa entre os conjuntos

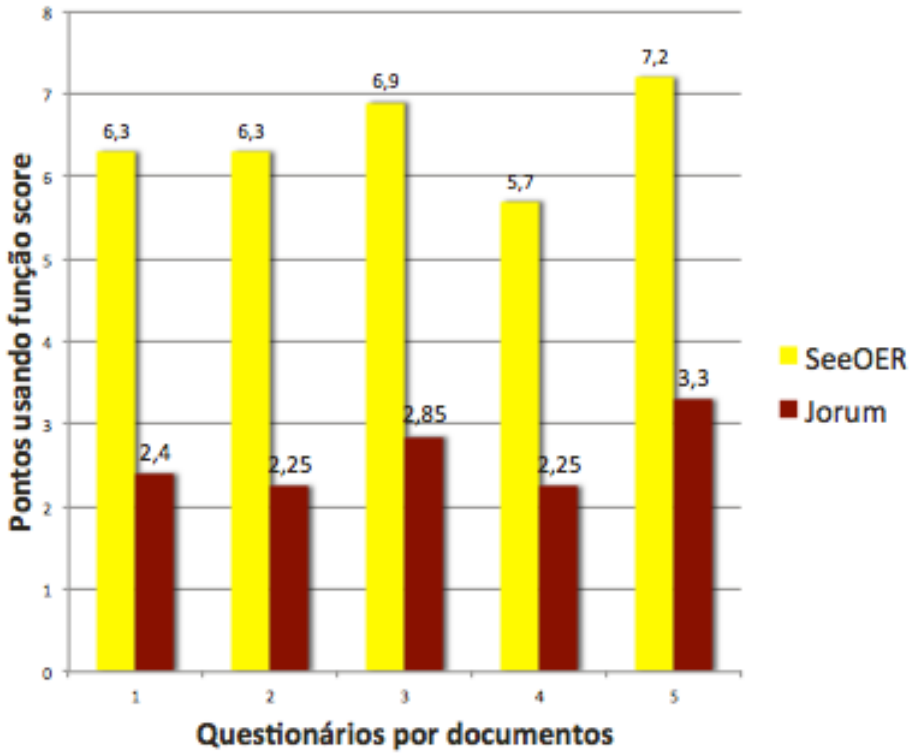


Figura 7.18: Resultado dos questionários de documentos usando a função score de forma comparativa entre os conjuntos

7.5 Considerações Finais

Neste capítulo foram abordados os resultados do mecanismo de busca na Web desenvolvido comparados com os trabalhos correlatos. Ademais, análise experimental foi dividido em crawler, reutilização de REA, experimento com procedência e resultados quantitativos e qualitativos do SeeOER. Além disso, os resultados qualitativos se dividiram em documentos de questionário e resultados comparativos usando função penalizada e função score. No próximo capítulo são descritos as conclusões, trabalhos publicados e trabalhos futuros.

Capítulo

8

Conclusões

De um ponto de vista mais abrangente, a pesquisa desenvolvida neste projeto de mestrado visa incentivar as práticas de utilização e produção de REA na educação, pois propõe um mecanismo de busca especializado. Os mecanismos de busca na Web genéricos não consideram as especificidades de REA e assim prejudicam a sua disseminação e incorporação em práticas educacionais. Vislumbra-se que os resultados deste projeto de mestrado tenha um alto impacto social.

Os REA possuem padrões de metadados, repositórios e plataformas com características particulares, assim introduzem heterogeneidades específicas. Os atuais mecanismos de busca na Web apresentam dois problemas importantes. Primeiro, eles são genéricos, assim buscam informação em qualquer lugar, desde páginas comerciais até definições escritas por pessoas anônimas (por exemplo, Wikipédia). Os mecanismos de busca na Web específicos por REA, como citado nos trabalhos correlatos, possuem diversas dificuldades para recuperar REA na Web. Além disso, muitos repositórios REA brasileiros não estão sendo identificados facilmente nos mecanismos de buscas genéricos e nem indexados nesses mecanismos de busca na Web específicos por REA.

Portanto, o SeeOER foi desenvolvido para suprir as lacunas deixadas em aberto para recuperação de REA na Web. As principais contribuições são:

- Avanço da publicação de REA no Brasil
- Mapeamento dos padrões de metadados que os mecanismos de busca devem usar em seu desenvolvimento
- Formas de instanciação e recuperação de REA por meio dos padrões de metadados e repositórios.

- Publicação de como desenvolver um mecanismo de busca na Web.
- Criação de um mecanismo de busca na Web focado em REA e considerando os padrões de metadados e repositórios brasileiros e estrangeiros.

Além disso, o projeto foi hospedado em uma máquina da *Amazon* e serviu como API para outros trabalhos relacionados na área. Pode citar um trabalho em conjunto com a Universidade Estadual de Maringá (UEM) o qual a API foi disponibilizada e usada para apoiar um outro projeto da área de REA.

8.1 Trabalhos publicados durante o mestrado

Durante o período do mestrado foram publicados pôster, artigos e relatório técnico como seguem:

- Publicação nos anais do Workshop of PhD and MSc Research (WTD)- *Proposal of a Web Search Engine for Open Educational Resources*;
- Apresentação no Workshop of PhD and MSc Research (WTD) - *Proposal of a Web Search Engine for Open Educational Resources*;
- Publicação "*SeeOER: Uma Arquitetura para Mecanismo de Busca na Web por Recursos Educacionais Abertos*" nos anais do 25º Simpósio Brasileiro de Informática na Educação (SBIE) - Principal publicação - com apresentação oral e como **artigo completo**;
- Publicação nos anais do 3º Congresso Brasileiro de Informática na Educação (CBIE) - "*SeeOER: Uma Arquitetura para Mecanismo de Busca na Web por Recursos Educacionais Abertos*" - artigo completo; e
- Relatório Técnico Publicado - *Metasearch unified shows the problems of Web search engines for OERs* - DOI: 10.13140/RG.2.1.2546.9920.

8.2 Trabalhos futuros

Os trabalhos futuros a serem realizados são:

- Usar algoritmos de aprendizagem de máquina para tratamento eficaz e real para qualidade de REA;

- Criação de novos algoritmos que utilizem inteligência artificial para tratar a qualidade efetiva de REA em pequena e grande escala;
- Tratar a qualidade dos REA;
- Usar técnicas computacionais para tratar o resultado obtido pelo SeeOER;
- Incrementar novos REA no SeeOER;
- Usar técnicas simples e avançadas para avaliação da qualidade de REA; e
- Novos experimentos quantitativos e qualitativos sobre a base.

Tendo em vista a diversidade e complexidade da efetivada para tratarem qualidades de REA, esse trabalho será abordado no futuro *doutorado* do próprio aluno. O projeto já foi escrito e aprovado na mesma instituição e terá inicio em breve.

Appendices

Apêndice

$\mathcal{A}$

Experimento usando o BigHand

Foi desenvolvido, inicialmente, um protótipo chamado *BigHand*, o qual consiste de um mecanismo de busca que trata a Web de forma geral e que é semelhante aos mecanismos de busca tradicionais. A tela inicial do *BigHand* é ilustrada na Figura A.1.

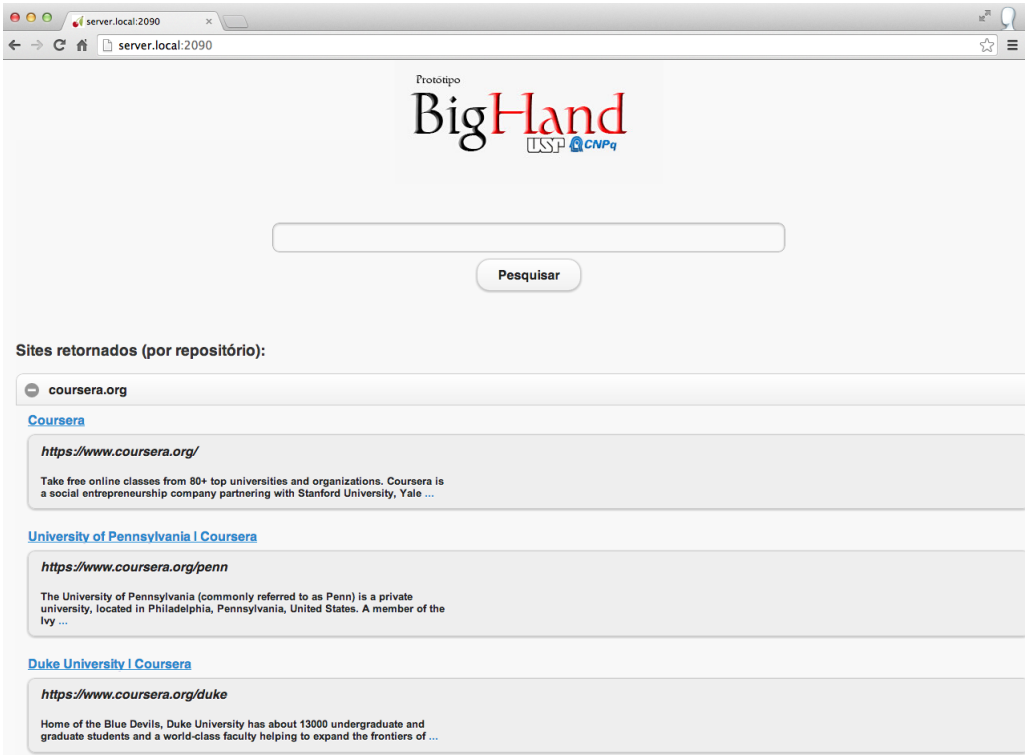


Figura A.1: *BigHand*: Tela inicial de busca.

A arquitetura do *BigHand* é ilustrada na Figura A.2. O retângulo tracejado representa o módulo da saída temporária, os outros módulos são representados por retângulos

ovulados, exceto o módulo externo que é representado por retângulo. A partir de uma expressão de busca informada pelo usuário, o *BigHand* percorre uma lista predefinida de repositórios, faz uma consulta individual a cada repositório por meio do Google e obtém os resultados retornados pelo Google. Esses resultados são retornados ao usuário categorizados por repositório, como ilustrado na Figura A.3. No exemplo dessa figura, a expressão de busca informada pelo usuário é *database*.

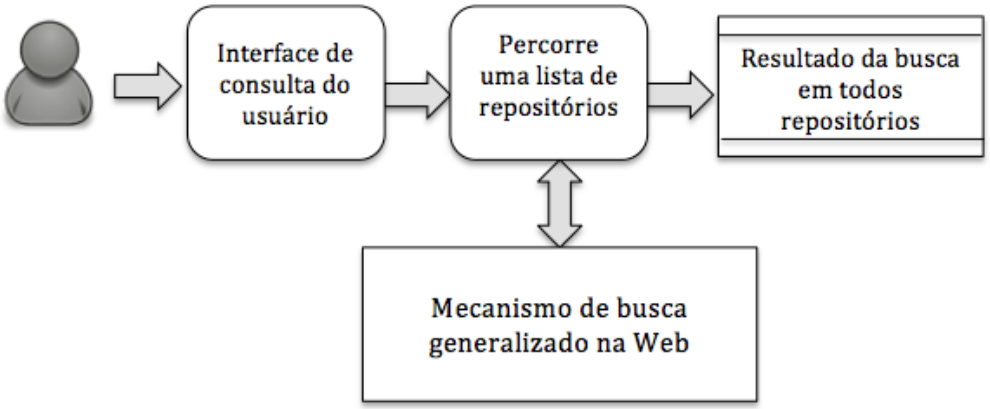


Figura A.2: Arquitetura do *BigHand*.

Paralelamente ao desenvolvimento do *BigHand*, foi criado um Google Personalizado[38] com uma lista de repositórios. Também foi feito um experimento utilizando o *BigHand* e o Google Personalizado. Nesse experimento, utilizou-se uma lista de repositórios idênticos a serem consultados e a mesma expressão de busca (ou seja, *database*). A diferença entre o *BigHand* e o Google Personalizado refere-se ao fato de que o ranqueamento do Google Personalizado é feito em relação ao conjunto de todos os sites, enquanto que o ranqueamento do *BigHand* é feito em termos dos documentos disponíveis em cada repositório, excluindo a influência de outros sites. Por exemplo, o *BigHand* retornou diversas páginas relacionadas ao Connexions, como a primeira página com o título *Database* e o URL `http://cnx.org/content/col10465/latest/`. Porém, no Google Personalizado, nenhuma página do Connexions foi ranqueada. No exemplo corrente, o Google Personalizado retornou 10 páginas, com 10 resultados cada uma. Foram verificadas todas as páginas de resultados, sendo que nenhuma delas referenciou o Connexions.

A Tabela A.1 mostra os resultados do experimento. Na primeira primeira coluna é indicado o nome do repositório retornado. Na segunda coluna são exibidos os títulos retornados pelo *BigHand*, considerando a primeira posição nas quais eles aparecem, enquanto

[38]Google Personalizado criado para experimento (URL público): `https://www.google.com/cse/publicurl?cx=007035912091441787493:ayywccjgon4`

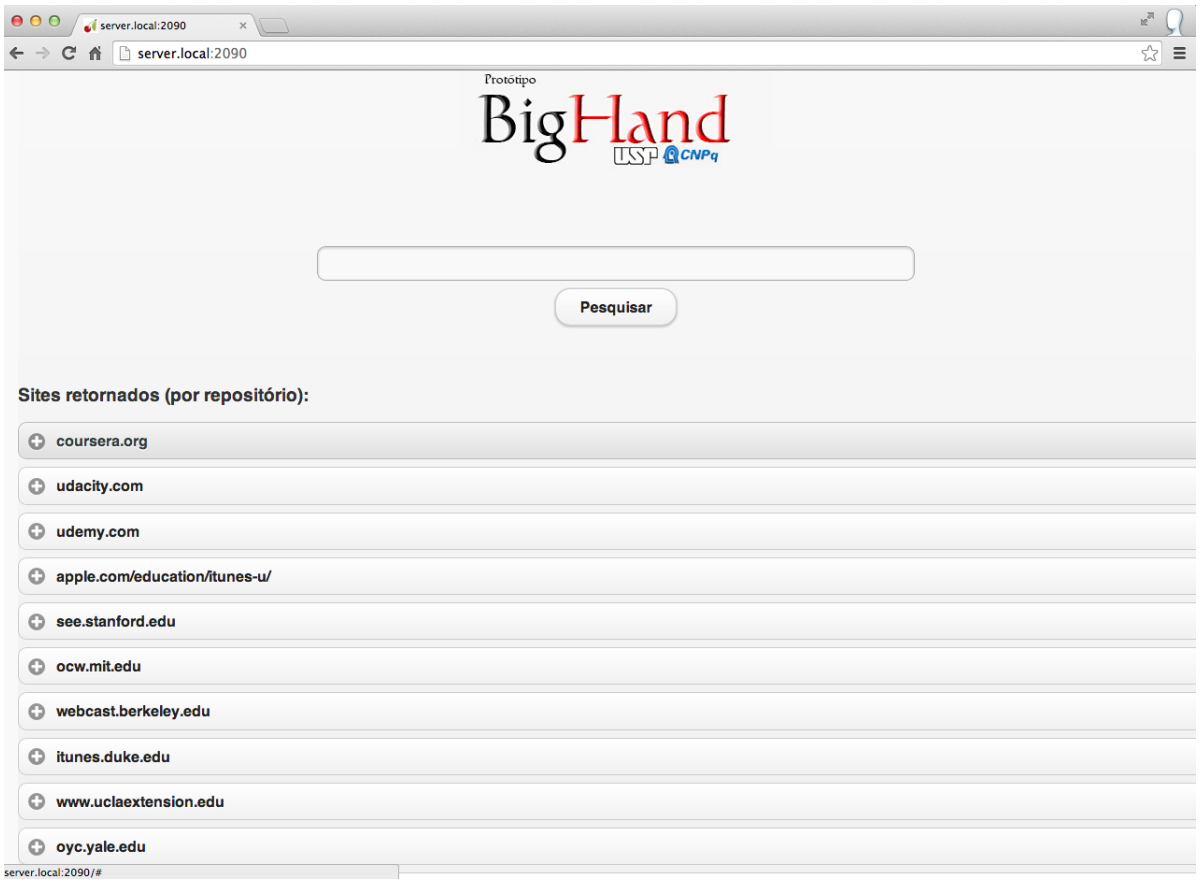


Figura A.3: *BigHand*: Tela com os resultados retornados por repositório.

que na terceira coluna são mostradas as posições nas quais são exibidos os resultados retornados pelo Google Personalizado. Para os repositórios que não retornaram nenhum resultado nos ambos mecanismos de busca, foram simbolizados por "-". Pode-se notar, na terceira coluna, que diversos repositórios não são retornados pelo Google Personalizado.

Com a realização desse experimento, foi possível observar que a forma de exibição dos resultados obtidos ao usuário é importante no contexto de REA, principalmente no que tange à posição dos resultados obtidos nas páginas retornadas ao usuário. Por exemplo, pode ser que o Google Personalizado tenha encontrado resultados relacionados ao Connexions, porém esses resultados não foram ranqueados adequadamente e, portanto, não foram retornados ao usuário.

Visando tornar o experimento mais geral, também foi feita uma busca usando a palavra-chave *database* no mecanismo de busca da Google disponível na Web, ou seja, usando-se o URL `http://www.google.com/search?num=100&start=0&q=database`, sem considerar o uso do Google Personalizado. Essa busca teve como objetivo encontrar nos resultados o site do Connexions. Foram retornadas 613 páginas, porém nenhuma página

Repositório	1° Posição usando o *BigHand*	Posição usando Google Personalizado
cousera.org	Introduction to Databases	1
udacity.com	CS253 - Lesson 3: Databases	16
udemy.com	MySQL Database For Beginners	2
apple.com/education/itunes-u/	-	-
see.stanford.edu	Introduction to Databases	Não apareceu
ocw.mit.edu	Syllabus | Database	24
webcast.berkeley.edu	View - UC Berkeley Webcasts - Video and Podcasts	Não apareceu
itunes.duke.edu	-	-
uclaextension.edu	UCLA Extension: Database	7
oyc.yale.edu	Open Yale Courses | Lecture 35 ...	Não apareceu
oli.web.cmu.edu	-	-
open.ac.uk	Databases	3
youtube.com/education	Disaster Epidemiology - YouTube	Não apareceu
youtube.com/course	CS403 Database	Não apareceu
cnx.org	Database	Não apareceu

Tabela A.1: Resultados do experimento usando o BigHand e o Google Personalizado.

referenciou o site desejado. Pode-se observar, novamente, que o ranqueamento que está sendo feito pelo principal motor de busca na Web atualmente disponível não garante bons resultados para o contexto de REA. Por exemplo, existem materiais muito importantes no site do Connexions, relacionados à busca, como o *Database Storage and Indexing*[39], os quais não podem ser usados pelo usuário porque não são retornados.

Com base nos experimentos descritos nesta seção, o mecanismo de busca desenvolvido nesta dissertação também investigou e implementou uma forma mais eficiente para recuperar os REA na Web, levando em consideração o contexto de REA e seus metadados.

[39] `http://cnx.org/content/m28159/latest/`

Apêndice

$\mathcal{B}$

Crawler: Lista de sementes

Na Tabela B.1 mostra uma lista inicial de sementes que será usada no *crawler* proposto.

Nome	Lista de sementes	País
Open Educational Resources for Typography	www.oert.org/	Argentina
The Le@rning Federation	www.ndlrn.edu.au/	Austrália
Repositório de teses da USP	teses.usp.br	Brasil
E-Aulas USP	eaulas.usp.br	Brasil
Banco de Metadados Geospaciais	metadados.geo.ibge.gov.br	Brasil
Banco Internacional de Objetos Educacionais	objetoseducacionais2.mec.gov.br	Brasil
Matematica Mutimidia	www.m3.mat.br/	Brasil
Pearson Copyleft	www.copyleftpearson.com.br	Brasil
Recursos Educacionais Abertos Brasil	rea.net.br/site/	Brasil
Commonwealth of Learning	www.col.org/	Canadá
RRU Open Educational Resources	oer.royalroads.ca/moodle/	Canadá
Educar Chile	www.educarchile.cl	Chile
Banco de Objetos de Aprendizaje	aplicaciones.virtual.unal.edu.co	Colômbia
Eduteka	www.eduteka.org/	Colômbia
RVP Metodicky Portal	dum.rvp.cz	Checa
Materialeplatformen	materialeplatform.emu.dk/	Dinamarca
Ariadne	www.ariadne-eu.org/	União Europeia

Open Science Resources	www.osrportal.eu/	União Europeia
Organic.Edunet Federation	www.organic-edunet.eu/	União Europeia
Edu Fi	www.edu.fi	Finlândia
LeMill	lemill.net/	Finlândia
FREIburger Multimedia Object Repository	freimore.uni-freiburg.de/	Alemanha
OpenLearnWare	openlearnware.hrz.tu-darmstadt.de/	Alemanha
The world lecture project	www.world-lecture-project.org/	Alemanha
UNITRACC	www.unitracc.com/	Alemanha
eGyankosh	www.egyankosh.ac.in/	India
Almae Matris Studiorum Campus	campus.unibo.it/	Itália
I Cleen	www.icleen.muse.it/	Itália
Open Educational Resources (OER) Africa	www.oerafrica.org/	Quênia
African Health OER Network	www.oerafrica.org/healthoer	Quênia
Science Attic	science-attic.org/	Coréia do Sul
VCILT	vcampus.uom.ac.mu/	Maurícia
Centro de Recursos para la Enseñanza y el Aprendizaje (CREA)	www.crea.udg.mx/	México
Desarrolla, Aprende y Reutiliza (DAR)	catedra.ruv.itesm.mx/	México
ITSON repositorio de objetos de aprendizaje	biblioteca.itson.mx/	México
Temoa	www.temoa.info	México
Wikiwijs	www.wikiwijs.nl/	Holanda
NLDA	ndla.no	Noruega
Banco de iten	bi.gave.min-edu.pt/	Portugal
Portal das Escolas	www.portaldasescolas.pt/	Portugal
Repositorio E-Learning	e-repository.tecminho.uminho.pt/	Portugal
Aljazeera creative commons repository	cc.aljazeera.net/	Catar
Maknaz	maknaz.elc.edu.sa/	Arábia Saudita
Everything Maths	everythingmaths.co.za/	África do Sul
Escuela virtual de Padres	www.web-familias.com/	Espanha
RODA	roda.culturaextremadura.com/	Espanha

Digiref	www.digiref.se/index.php	Suécia
Kursnavet	www.kursnavet.se/	Suécia
Skolresurser	skolresurser.se	Suécia
OER Online Archive	www.archive.org/	-
La Flor (Laclo)	laflor.laclo.org/	-
The Open Learning	www.open.edu/openlearn/?	Reino Unido
Economics Network Online Learning and Teaching Materials	www.economicsnetwork.ac.uk/	Reino Unido
First World War Poetry Digital Archive	www.oucs.ox.ac.uk/ww1lit/	Reino Unido
Hum Box	humbox.ac.uk/	Reino Unido
Jorum	www.jorum.ac.uk/	Reino Unido
Lab Space	labspace.open.ac.uk/	Reino Unido
MathWorld	mathworld.wolfram.com/	Reino Unido
National Learning Network	www.nln.ac.uk/	Reino Unido
OpenLearn	www.open.edu/openlearn	Reino Unido
OSTRICH	ostrich.bath.ac.uk/	Reino Unido
Restore	www.restore.ac.uk	Reino Unido
University of Leicester OER Repository	www2.le.ac.uk/projects/oer	Reino Unido
Xpert	www.nottingham.ac.uk/xpert/	Reino Unido
Academic Earth	academicearth.org	Estados Unidos
CChemCollective	www.chemcollective.org/	Estados Unidos
Connexions	cnx.org	Estados Unidos
OCW-MIT	ocw.mit.edu	Estados Unidos
Khan *Academy*	khanacademy.org	Estados Unidos
Plataforma EDX	edx.org	Estados Unidos
Fundação Saylor	saylor.org	Estados Unidos
Coursera	coursera.org	Estados Unidos
Consortium for the Advancement of Undergraduate Statistics Education	www.causeweb.org	Estados Unidos
CSTC (Computing Science Teaching Center)	www.cstc.org/	Estados Unidos
Culturally Authentic Pictorial Lexicon	capl.washjeff.edu	Estados Unidos

Curriki	welcome.curriki.org/	Estados Unidos
Digital Library for Earth System Education	www.dlese.org/library/	Estados Unidos
Federal Resources for Educational Excellence	www.free.ed.gov/index.cfm	Estados Unidos
Geoscience Data Repository	www.nrcan.gc.ca/earth-sciences/home	Estados Unidos
I-Berry	iberry.com/	Estados Unidos
Ilumina	www.ilumina-dlib.org	Estados Unidos
Maricopa Learning Exchange	www.mcli.dist.maricopa.edu/mlx/	Estados Unidos
Merlot	www.merlot.org	Estados Unidos
National Science Digital Library (NSDL)	nsdl.org/	Estados Unidos
NEEDS	www.needs.org/	Estados Unidos
OER Commons	oercommons.org/	Estados Unidos
OER Equella	oer.equella.com/access/home.do	Estados Unidos
OpenMichigan	open.umich.edu/	Estados Unidos
OTAN	www.otan.us/	Estados Unidos
Phet (Physics Education Technology)	phet.colorado.edu/en/	Estados Unidos
The Gateway	www.thegateway.org/	Estados Unidos
Wisconsin Online Resource Center	www.wisc-online.com/	Estados Unidos
World History Sources	chnm.gmu.edu/	Estados Unidos
Repositorio de objetos de aprendizaje	roa.mppeu.gob.ve/	Venezuela

Tabela B.1: Tabela de sementes para o *crawler* proposto.

Apêndice

C

Questionário de Qualidade Respondido

Abaixo seguem os questionários respondidos pelo diversos grupos de pessoas. É importante deixar anexado na dissertação, tendo em vista a importância dos dados.

Além disso, o experimento foi um experimento controlado. As pessoas poderiam deixar de responder o questionário a qualquer momento, poderiam não responder ou desistir do experimento. Também, o experimento teve pessoas que não responderam. Os quais foram eliminados do processo.

As pessoas que responderam tiveram um agradecimento pelo auxílio e ajuda à pesquisa e ciência no Brasil.

Questionário

(Exemplar do questionário)

Abaixo são apresentadas perguntas advindas de um trabalho de mestrado, orientado pela profa. Cristina A. D. Ciferri, co-orientado pela profa. Itana M. Gimenes e pelo aluno de mestrado Murilo G. Gazzola. Qualquer dúvida deve ser mencionada para que possa ser retirada antes de responder este questionário.

1. Qual seu grau de escolaridade? *

( ) Não possuo ( ) Ensino Fundamental ( ) Ensino Médio ( ) Graduação ( ) Pós-graduação

2. Qual sua área de atuação? *

( ) Nenhuma ( ) Exatas ( ) Humanas ( ) Biológicas

3. Como considera seu nível de inglês? *

( ) Básico ( ) Intermediário ( ) Avançado

4. Você está buscando por "Introdução à Internet". Que nota você daria para o resultado abaixo ? *

Identifying opportunities and using an interactive whiteboard with learners Authors: Geoff Foot Created: 3 January 2008, by Geoff Foot This collection of resources offers a guide to using an interactive whiteboard in teaching. It includes instructions on switching on the whiteboard, suggested ways of using the whiteboard with learners and example resources in early years, key skills and construction education. This resource was funded by the LSN as part of the Q projects programme. The resource was designed & produced by South Devon College. Published: 23 February 2011, by Learning and Skills Network Ltd, Geoff Foot Keywords: digital, presentations, whiteboards, learning, teaching, interactive Licence: Attribution-Noncommercial-Share Alike 2.0 UK: England & Wales

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

5. Você está buscando por "Introdução à Internet". Que nota você daria para o resultado abaixo ? *

This is a basic intro to the Internet. Published: 12 Oct 2007. University Psu by Cole Camplese. Keywords: internet, technology. Type: Course. License: Creative Commons Attribution Licence CC-BY 2.0

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

6. Você está buscando por "Álgebra I". Que nota você daria para o resultado abaixo ? *

The First Law of Thermodynamics and Enthalpy In this video, I shall present the first law of thermodynamics, and show how it governs the state function internal energy, and how considerations of the first law for isochoric and isobaric processes leads to the definition of the further state function called enthalpy. Published: 26 February 2015, by University of Manchester. Authors: Dr. Jonathan Agger. Keywords: virtual experiment, virtual lab, chemistry, Thermodynamics, physical chemistry, School of Chemistry, Faculty of Engineering and Physical Sciences, University of Manchester.Created: 1 June 2014, by Dr. Jonathan Agger. Contributor(s): Stephen Wheeler, Ian Hutt, Prof. Michael Anderson, Dr. Patrick O'Malley. Licence: Attribution-NonCommercial-ShareAlike 4.0 International

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

7. Você está buscando por "Álgebra I". Que nota você daria para o resultado abaixo ? *

Algebra I for the Community College. Subcollection: Mathematics and Statistics. Language: English (en).Type: Textbook. Created: Oct 23, 2013 9:43 am GMT-5. Author: Ann Simao (afsimao@stcc.edu). Licence: Creative Commons Attribution CC-BY 4.0.

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

8. Você está buscando por "Álgebra e Física". Que nota você daria para o resultado abaixo ? *

Observation, measurement and the recording of data are central activities in science. Speculation and the development of new theories are crucial as well, but ultimately the predictions resulting from those theories have to be tested against what actually happens and this can only be done by making further measurements. Whether measurements are made using simple instruments such as rulers and thermometers, or involve sophisticated devices such as electron microscopes or lasers, there are decisions to be made about how the results are to be represented, what units of measurements will be used and the precision to which the measurements will be made. In this unit we will consider these points in turn. Published: 15 January 2010, by Open University Keywords: statistics, standard_deviation, significant_figure, scientific_notation, science, sample, probability, normal_distribution, measurement, maths, logarithm, integer, decimal, data, average, science and nature Licence: Attribution-Noncommercial-Share Alike 2.0 UK: England & Wales

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

9. Você está buscando por "Álgebra e Física". Que nota você daria para o resultado abaixo ? *

Introduction to Applied Math and Physics. This introductory, algebra-based, one-semester physics book is based on OpenStax College Physics. This online, fully editable and customizable title includes learning objectives, concept questions, links to labs and simulations, and ample practice opportunities to solve traditional physics application problems. Language: English (en). Collection: Mathematics and Statistics, Science and Technology. Keywords: college physics, collisions, elasticity, energy, forces, friction, kinematics, linear momentum, Newton's Laws of Motion, physics, rotational motion and angular momentum, statics and torque, uniform circular motion and gravitation, work. License: Creative Commons Attribution License CC-BY 3.0. Contributor (s): Oka Kurniawan.

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

10. Você está buscando por "Biotecnologia". Que nota você daria para o resultado abaixo ? *

Final assessment SCORM format Authors: Janet Fyfe Created: 26 March 2008, by Janet Fyfe This resource comprises of interactive material created as part of the CeLLs Project covering the topic of labelled antibodies. This section is a revision self-assessment with multiple-choice questions.

Published: 23 February 2011, by The Adam Smith College Fife Keywords: immunoassays, antigens, antibodies Licence: Attribution-Noncommercial-Share Alike 2.0 UK: England & Wales

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

11. Você está buscando por "Biotecnologia". Que nota você daria para o resultado abaixo ? *

Biotechnology . Language: English (en). Created: Mar 13, 2014 5:06 pm GMT-5.Licence: Creative Commons Attribution License CC-BY 4.0. Author: Jeffrey Mahr.

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

12. Você está buscando por "Fisiologia e Anatomia Humana". Que nota você daria para o resultado abaixo ? *

The use of 3D technology to deliver science curriculum Authors: Dumfries and Galloway College, JISC RSC Scotland Created: 20 January 2012, by Dumfries and Galloway College, JISC RSC Scotland 3-D technology can be used to add innovative and dynamically interesting dimensions to the teaching and learning process. The use of 3-d technology in a range of subjects can be used to emphasise, highlight and trigger attention processes to key learning concepts in the classroom. Using a portable active 3-d projector and specific 3-d curriculum packs, the college has begun to integrate 3-d technology to offer the learners a new and exciting way to combine both visual and kinaesthetic preferences to learning with a current focus on trialling its use in plant biology, human anatomy and physiology and sports science. Published: 7 November 2012, by JISC RSC Scotland Keywords: science, 3D technology, Dumfries and Galloway College, case studies Licence: Attribution 3.0 Unported

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

13. Você está buscando por "Anatomia Humana e Fisiologia". Que nota você daria para o resultado abaixo ? *

Anatomy & Physiology: Energy, Maintenance and Environmental Exchange. Human Anatomy and Physiology is designed for the two-semester anatomy and physiology course taken by life science and allied health students. The textbook follows the scope and sequence of most Human Anatomy and Physiology courses, and its coverage and organization were informed by hundreds of instructors who teach the course. Instructors can customize the book, adapting it to the approach that works best in their classroom. The artwork for this textbook is aimed focusing student learning through a powerful blend of traditional depictions and instructional innovations. Color is used sparingly, to emphasize the most important aspects of any given illustration. Significant use of micrographs from the University of Michigan complement the illustrations, and provide the students with a meaningful alternate depiction of each concept. Finally, enrichment elements provide relevance and deeper context for students, particularly in the areas of health, disease, and information relevant to their intended careers. Language:English (en). Collection:Science and Technology. Keywords: "acid-base balance, action potential, adrenal, anatomy, antibody, appendicular skeleton, atom, autonomic nervous system, axial skeleton, blood, body fluid, bone, capillary, cardiac, cardiovascular system, cell, central nervous system, chemical, circulatory pathway, connective tissue, cytoplasm, development, diet, digestion, digestive system, DNA, electrical activity, electrolyte, element, endocrine, energy, epithelial, erythrocyte, esophagus, exam, fascicle, fertilization, fetal, fluid balance, gallbladder, gland, heart, heat balance, hemostasis, homeostasis, hormone, hypothalamus, immune system, immunity, inheritance, inorganic compound, integumentary system, intestine, joint, kidney, lactation, leukocyte, liver, lung, lymphatic system, lymphocyte, metabolism, motor response, muscle, nervous system, neurological, neuron, nucleus, nutrition, organic compound, pancreas, pathogen, pectoral girdle, pelvis, peripheral nervous system, pharynx, physiology, placental hormone, platelet, protein, renal, reproduction, reproductive system, respiration, respiratory system, sensory perception, skeletal system, skeleton, skin, skull, stomach, thoracic cage, thorax, tissue, urinary system, urine, vascular system, vertebrae, water balance. Type: Textbook. Contributor (s): Wade Hedegard. Licence: Creative Commons Attribution CC-BY 4.0.

( ) Ótimo
( ) Bom
( ) Regular
( ) Ruim
( ) Muito Ruim

Referências Bibliográficas

Abeywardena, I.; Chan, C.; Tham, C. Oerscout technology framework: A novel approach to open educational resources search. *The International Review of Research in Open and Distance Learning*, v. 14, n. 4, 2013. Disponível em: `http://www.irrodl.org/index.php/irrodl/article/view/1505`

Abeywardena, I. S.; Chan, C. S. Review of the current oer search dilemma. 2013.

Anand, M. K.; Bowers, S.; McPhillips, T.; Ludäscher, B. Efficient provenance storage over nested data collections. In: *EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology*, 2009. New York, NY, USA: ACM, 2009. p. 958–969.

Archer, D. W.; Delcambre, L. M. L.; Maier, D. A framework for fine-grained data integration and curation, with provenance, in a dataspace. In: *Workshop on the Theory and Practice of Provenance*, 2009.

Atkins, D. E.; Brown, J. S.; Hammond, A. L. *A review of the open educational resources (oer) movement: Achievements, challenges, and new opportunities.* Creative common, 2007.

Baeza-Yates, R.; Ribeiro-Neto, B. Modern information retrieval: The concepts and technology behind search. 2011. Disponível em: `http://books.google.com.br/books?id=HbyAAAAACAAJ`

Benjelloun, O.; Das Sarma, A.; Halevy, A.; Theobald, M.; Widom, J. Databases with uncertainty and lineage. *The VLDB Journal*, v. 17, n. 2, p. 243–264, 2008.

Bhattacharjee, A.; Jamil, H. A schema matching system for on-the-fly autonomous data integration. *International Journal of Information and Decision Sciences*, v. 4, n. 2/3, p. 167–181, 2012.

Bissell, A.; Park, J.; Yergler, N.; Linksvayer, M. *Enhanced search for educational resources - a perspective and a prototype from cclearn.* Relatório Técnico, ccLearn, 2009.

Borba, J. L. Elementos do núcleo de metadata "dublin core", versão 1.1: Descrição de referência. `http://purl.pt/201/1/`, acessado em: 2013-09-07, 2000.

Boudreau, M.-C.; Gefen, D.; Straub, D. W. Validation in information systems research: A state-of-the-art assessment. *MIS Quarterly*, v. 25, n. 1, p. pp. 1–16, 2001. Disponível em: `http://www.jstor.org/stable/3250956`

Brin, S.; Page, L. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, v. 56, n. 18, p. 3825 – 3833, 2012. Disponível em: `http://www.sciencedirect.com/science/article/pii/S1389128612003611`

Buneman, P.; Chapman, A.; Cheney, J. Provenance management in curated databases. In: *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, 2006a. New York, NY, USA: ACM, 2006a. p. 539–550.

Buneman, P.; Chapman, A.; Cheney, J.; Vansummeren, S. A provenance model for manually curated data. In: Moreau, L.; Foster, I. T., eds. *IPAW*, 2006b. Springer, 2006b. p. 162–170 (*Lecture Notes in Computer Science*, v.4145).

Buneman, P.; Khanna, S.; Tan, W. C. Data provenance: Some basic issues. In: *FST TCS 2000: Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science*, 2000. London, UK: Springer-Verlag, 2000. p. 87–93.

Buneman, P.; Khanna, S.; Tan, W. C. Why and where: A characterization of data provenance. In: *ICDT '01: Proceedings of the 8th International Conference on Database Theory*, 2001. London, UK: Springer-Verlag, 2001. p. 316–330.

Cameron, G. Provenance and pragmatics. In: *Workshop on Data Provenance and Annotation, Edinburgh*, 2003.

Cao, Y.; Fan, W.; Yu, W. Determining the relative accuracy of attributes. In: *SIGMOD'13: Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2013. p. 565–576.

Chae, M.; Kim, J.; Kim, H.; Ryu, H. Information quality for mobile internet services: A theoretical model with empirical validation. *Electronic Markets*, v. 12, n. 1, p. 38–46, 2002.

Chapman, A. P.; Jagadish, H. V.; Ramanan, P. Efficient provenance storage. In: *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008. New York, NY, USA: ACM, 2008. p. 993–1006.

Chen, Y.-N.; Seadle, M.; Seadle, M. A rdf-based approach to metadata crosswalk for semantic interoperability at the data element level. *Library Hi Tech*, v. 33, n. 2, 2015.

Croft, B.; Metzler, D.; Strohman, T. *Search engines: Information retrieval in practice*. Pearson Education, 2011.

DCMI Dublin core metadata element set. `http://dublincore.org/documents/dces/`, acessado em: 2013-09-09, 2012.

Del Rio, N.; Silva, P. P. Probe-it! visualization support for provenance. In: *Advances in Visual Computing*, v. Volume 4842/2007 de *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2007, p. 732–741.

Demsky, B. Cross-application data provenance and policy enforcement. *ACM Transactions on Information and System Security*, v. 14, n. 1, p. 6, 2011.

Dietze, S.; Yu, H. Q.; Giordano, D.; Kaldoudi, E.; Dovrolis, N.; Taibi, D. Linked education: interlinking educational resources and the web of data. In: *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 2012. SAC '12, New York, NY, USA: ACM, 2012. p. 366–371 (*SAC '12*,). Disponível em: `http://doi.acm.org/10.1145/2245276.2245347`

Dong, X.; Berti-Equille, L.; Hu, Y.; Srivastava, D. SOLOMON: Seeking the truth via copying detection. *PVLDB*, v. 3, n. 2, p. 1617–1620, 2010.

Elmasri, R.; Navathe, S. *Fundamentals of database systems [with access code]*. ADDISON WESLEY Publishing Company Incorporated, 2011. Disponível em: `http://books.google.com.br/books?id=ZdhAQgAACAAJ`

Eysenbach, G.; Köhler, C. How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ*, v. 324, n. 7337, p. 573–577, 2002.

Facebook Facebook developer wiki. `http://fbdevwiki.com/wiki/Open_Graph_protocol`, acessado: 2013-09-10, 2013a.

Facebook The open graph protocol. 2013b. Disponível em: `http://ogp.me/`

Fan, W.; Geerts, F.; Tang, N.; Yu, W. Inferring data currency and consistency for conflict resolution. In: *ICDE'13: Proceedings of the IEEE International Conference on Data Engineering*, 2013. p. 470–481.

Ferreira, A. A.; Gonçalves, M. A.; Laender, A. H. F. A brief survey of automatic methods for name disambiguation. *Sigmod Record*, v. 41, n. 2, p. 15–26, 2012.

Ferreira, E. C. H. G. *Geração automática de metadados: uma contribuição para a web semântica*. Tese de Doutoramento, 2006.

Fogarty, G.; Cretchley, P.; Harman, C.; Ellerton, N.; Konki, N. Validation of a questionnaire to measure mathematics confidence, computer confidence, and attitudes towards the use of technology for learning mathematics. *Mathematics Education Research Journal*, v. 13, n. 2, p. 154–160, 2001. Disponível em: `http://dx.doi.org/10.1007/BF03217104`

Freire, J.; Koop, D.; Santos, E.; Silva, C. T. Provenance for computational tasks: A survey. *IEEE Computing in Science & Engineering*, v. 10, n. 3, p. 11–21, 2008.

Gazzola, M. G.; Ciferri, C. D.; Gimenes, I. M. Seeoer: Uma arquitetura para mecanismo de busca naweb por recursos educacionais abertos. In: *Anais do Simpósio Brasileiro de Informática na Educação*, 2014. p. 1013–1022.

Gimenes, I. M.; Barroca, L.; Feltrim, V. D. Tendências na educação a distância e educação aberta na computação. *CSBC*, 2012. Disponível em: `http://www.lbd.dcc.ufmg.br/colecoes/jai/2012/001.pdf`

Glavic, B.; Ditt, K. R. Data provenance: A categorization of existing approaches. In: *The German Database Conference*, 2007. p. 227–241.

GOOGLE Video sitemaps. `http://www.google.com/schemas/sitemap-video/1.1/`, acessado em: 2013-09-15, 2011.

GRAHAM, W. Facebook developer tools. In: *Beginning Facebook Game Apps Development*, Apress, 2012, p. 201–229.

HALEVY, A. Y.; ASHISH, N.; BITTON, D.; CAREY, M.; DRAPER, D.; POLLOCK, J.; ROSENTHAL, A.; SIKKA, V. Enterprise information integration: Successes, challenges and controversies. In: *International Conference on Management of Data (SIGMOD)*, 2005. p. 778–787.

HALEVY, A. Y.; RAJARAMAN, A.; ORDILLE, J. Data integration: The teenage years. In: *International Conference on Very Large Data Bases (VLDB)*, 2006. p. 9–16.

HEINIS, T.; ALONSO, G. Efficient lineage tracking for scientific workflows. In: *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008. New York, NY, USA: ACM, 2008. p. 1007–1018.

HILU, L.; TORRES, P. L.; BEHRENS, M. A. Rea (recursos educacionais abertos)–conhecimentos e (des) conhecimentos. *Revista Científica e-curriculum. ISSN 1809-3876*, v. 13, n. 1, p. 130–146, 2015.

HOGAN, A.; HARTH, A.; UMBRICH, J.; KINSELLA, S.; POLLERES, A.; DECKER, S. Searching and browsing linked data with swse: The semantic web search engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, v. 9, n. 4, p. 365 – 401, 2011. Disponível em: `http://www.sciencedirect.com/science/article/pii/S1570826811000473`

IKEDA, R.; CHO, J.; FANG, C.; SALIHOGLU, S.; TORIKAI, S.; WIDOM, J. Provenance-based debugging and drill-down in data-oriented workflows. In: *IEEE 28th International Conference on Data Engineering (ICDE)*, 2012. p. 1249–1252.

JORUM Jorum for developers. `http://www.jorum.ac.uk/developers/`, acessado em: 2013-10-21, 2013a.

JORUM Jorum stats report. `http://find.jorum.ac.uk/report`, acessado em: 2013-10-22, 2013b.

JORUM Our backgrund jorum. `http://www.jorum.ac.uk/about-us/`, acessado em: 2013-10-20, 2013c.

JORUM Find jorum. `http://find.jorum.ac.uk`, acessado em: 2014-07-10, 2014.

Kanwar, A.; Uvalić-Trumbić, S.; Butcher, N. *A basic guide to open educational resources (oer).* Vancouver: Commonwealth of Learning; Paris: UNESCO, 2011.

Koutsomitropoulos, D. A.; Solomou, G. D.; Papatheodorou, T. S.; Alexopoulos, A. D. The use of metadata for educational resources in digital repositories: Practices and perspectives. *D-Lib Magazine*, v. 16, n. 1, p. 3, 2010.

LibraryPy, P. The python standard library. `https://docs.python.org/2/library/`, acessado em: 2015-05-15, 2013.

Little, S.; Mikroyannidis, A.; Okada, A.; Scott, P. Formal metadata and shared experiences for discovering tools to adapt open educational resources. In: *Signal-Image Technology and Internet-Based Systems (SITIS), 2011 Seventh International Conference on*, 2011. p. 147–153.

Lousangfa, P.; Sahavechaphan, N.; Buranarach, M.; Vannrat, S.; Kawtrakul, A.; Nagatsuka, T.; Ninomiya, S.; et al. Survey and modeling metadata schema relationship in agriculture domain for better metadata schema service. In: *Proceeding of World Conference on Agriculture Information and IT: IAALD AFITA and WCCA. Tokyo, Japan*, 2008.

Matkin, G. W. Open educational resources in the post mooc era. *eLearn*, v. 2013, n. 4, 2013. Disponível em: `http://dl.acm.org/citation.cfm?id=2460459.2460460`

McClelland, M. Metadata standards for educational resources. *Computer*, v. 36, n. 11, p. 107–109, 2003.

Munroe, S.; Miles, S.; Moreau, L.; Vázquez-Salceda, J. Prime: a software engineering methodology for developing provenance-aware applications. In: *SEM '06: Proceedings of the 6th international workshop on Software engineering and middleware*, 2006. New York, NY, USA: ACM, 2006. p. 39–46.

Nguyen, H.-Q.; Taniar, D.; Rahayu, J.; Nguyen, W. K. Double-layered schema integration of heterogeneous XML sources. *Journal of Systems and Software*, v. 1, n. 1, p. 63–76, 2011.

Okada, A. Knowledge media technologies for open learning in online communities. *The International Journal of Technology, knowledge & Society*, v. 3, 2007.

Pearson, D. Presentation on grid data requirements scoping metadata & provenance. In: *Workshop on Data Derivation and Provenance, Chicago*, 2002.

Phil Barker, Lorna M. Campbell, A. R. C. S. Ims meta-data best practice guide for ieee 1484.12.1-2002. `https://www.imsglobal.org/metadata/mdv1p3/imsmd_bestv1p3.html#1621610`, acessado em: 2015-09-11, 2006.

Powell, A.; Johnston, P. Guidelines for implementing dublin core in xml. `http://dublincore.org/documents/dc-xml-guidelines`, acessado em: 2013-09-13, 2003.

Prabhakar, S.; Richardson, J.; Srivastava, J.; Lim, E.-P. Instance-level integration in federated autonomous databases. In: *Hawaiian Conference for System Science*, 1993. p. 62–69.

Rathod, N.; Cassel, L. Building a search engine for computer science course syllabi. In: *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2013. JCDL '13, New York, NY, USA: ACM, 2013. p. 77–86 (*JCDL '13*,). Disponível em: `http://doi.acm.org/10.1145/2467696.2467723`

Reitsma, R.; Marshall, B.; Chart, T. Can intermediary-based science standards crosswalking work? some evidence from mining the standard alignment tool (sat). *Journal of the American Society for Information Science and Technology*, v. 63, n. 9, p. 1843–1858, 2012. Disponível em: `http://dx.doi.org/10.1002/asi.22712`

Richardson, L. Beautiful soup documentation. 2015.

Salton, G.; Harman, D. Information retrieval. In: *Encyclopedia of Computer Science*, Chichester, UK: John Wiley and Sons Ltd., 2003, p. 858–863. Disponível em: `http://dl.acm.org/citation.cfm?id=1074100.1074478`

Sanderson, M.; Croft, W. The history of information retrieval research. *Proceedings of the IEEE*, v. 100, n. Special Centennial Issue, p. 1444–1451, 2012.

de Santiago, R.; Raabe, A. L. Architecture for learning objects sharing among learning institutions. *IEEE Transactions on Learning Technologies*, v. 3, n. 2, p. 91–95, 2010.

Shu, L.; Chen, A.; Xiong, M.; Meng, W. Efficient SPectrAl neighborhood blocking for entity resolution. In: *Proceedings of the 27th International Conference on Data Engineering*, 2011. p. 1067–1078.

Simmhan, Y. L.; Plale, B.; Gannon, D. A survey of data provenance in e-science. *ACM Sigmod Record*, v. 34, n. 3, p. 31–36, 2005.

Siqueira, I. C. P. Mecanismos de busca na web: passado, presente e futuro. *Ponto-deAcesso*, v. 7, n. 2, p. 47–67, 2013.

Sitemaps What are sitemaps? `http://www.sitemaps.org/`, acessado em: 2013-09-15, 2008.

Smith, M. S.; Casserly, C. M. The promise of open educational resources. *Change: The Magazine of Higher Learning*, v. 38, n. 5, p. 8–17, 2006.

Spector, M.; Merrill, D.; van Merrienboer, J.; Driscoll, M. *Handbook of research on educational communications and technology: A project of the association for educational communications and technology.* AECT Series. Taylor & Francis, 345 - 352 p., 2007.

Tan, W.-C. Research problems in data provenance. *IEEE Data Engineering Bulletin*, v. 27, n. 4, p. 45–52, 2004.

Unal, O.; Afsarmanesh, H. Semi-automated schema integration with SASMINT. *Knowledge and Information Systems*, v. 21, n. 1, p. 99–128, 2010.

UNESCO Survey on governments' open educational resources (oer) policies. *ONU - Commonwealth of Learning*, 2009.

Warpechowski, M. Recuperação de metadados de objetos de aprendizagem no AdaptWeb. 2005. Disponível em: `http://hdl.handle.net/10183/6925`

Whang, S. E.; Garcia-Molina, H. Joint entity resolution. In: *Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE)*, 2012. p. 294–305.

Widom, J. Trio: A system for integrated management of data, accuracy, and lineage. In: *Proceedings of the Second Biennial Conference on Innovative Data Systems Research (CIDR '05)*, 2005.

Wiley, D. Openness as catalyst for an educational reformation. *EDUCAUSE Review*, v. 45, 2010.

Wiley, D.; Bliss, T.; McEwen, M. Open educational resources: A review of the literature, p. 781–789. 2014.

Wiley, D. A. *The learning objects literature.* 2002. Disponível em: `www.opencontent.org/docs/wiley-lo-review-final.pdf?`

Zhao, B.; Xu, S.; Lin, S.; Luo, X.; Duan, L. A new visual navigation system for exploring biomedical open educational resource (oer) videos. *Journal of the American Medical Informatics Association*, 2015.

Zhao, Y.; Wilde, M.; Foster, I. T. Applying the virtual data provenance model. In: *International Provenance and Annotation Workshop*, v. 4145, Chicago, IL, USA, 2006, p. 148–161.

Zhu, J.; Fung, G.; Zhou, X. Efficient web pages identification for entity resolution. In: *Proceedings of the International Conference on World Wide Web*, 2010. p. 1223–1224.