PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Programa de Pós-Graduação em Informática

Daniel Eugênio Neves

RECOMENDAÇÃO E AGREGAÇÃO DE CONTEÚDOS RELACIONADOS EM CONFORMIDADE COM O PADRÃO SCORM

Belo Horizonte
2014

Daniel Eugênio Neves

RECOMENDAÇÃO E AGREGAÇÃO DE CONTEÚDOS RELACIONADOS EM CONFORMIDADE COM O PADRÃO SCORM

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Orientadora: Lucila Ishitani

Belo Horizonte

2014

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Neves, Daniel Eugênio
N513r Recomendação e agregação de conteúdos relacionados em conformidade com o padrão SCORM / Daniel Eugênio Neves. Belo Horizonte, 2014.
94 f. : il.

Orientadora: Lucila Ishitani
Dissertação (Mestrado) - Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática.

1. Tecnologia educacional. 2. Educação a distância. 3. Recuperação da informação. 4. Ensino auxiliado por computador. 5. Mineração de dados (Computação). I. Ishitani, Lucila. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.

CDU: 37:681.3

Daniel Eugênio Neves

RECOMENDAÇÃO E AGREGAÇÃO DE CONTEÚDOS RELACIONADOS EM CONFORMIDADE COM O PADRÃO SCORM

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Lucila Ishitani (Orientadora) – PUC Minas

Mark Alan Junho Song – PUC Minas

Márcia Gorett Ribeiro Grossi – CEFET MG

Belo Horizonte, 01 de outubro de 2014.

Aos meus pais, por que estou aqui. Por seu amor e dedicação. Pelas lições primeiras, essência de tudo.

À minha esposa amada. Companheira sempre. Pela presença em todas as suas formas. Vela e horizonte, motivo e impulso.

Aos verdadeiros professores que tive, de hoje até quando me lembro. Os tenho em conta, em algumas de minhas melhores lembranças. Seria capaz de citá-los, talvez conseguisse descrever o que me deixaram de bom.

AGRADECIMENTOS

A todos que contribuíram para a realização deste trabalho, que me auxiliaram no dia a dia, que me instruíram e me ensinaram. Aos colegas de estudo, pelas ideias trocadas, pela cumplicidade na superação dos desafios. Aos amigos, pelo apoio e incentivo, pela paciência durante as longas conversas sobre sonhos, receios e expectativas.

À Professora Lucila Ishitani, por sua dedicação e generosidade, compromisso e competência, presença infalível como orientadora e professora. Por tudo o que me ensinou.

Ao Professor Wladmir Cardoso Brandão, pelo apoio, incentivo e fundamental contribuição.

Às amigas Léa Cavedagne, Elizabeth Junqueira e Maria do Carmo, pelo apoio, pelas longas conversas e pela inestimável contribuição.

À FAPEMIG, por ter me auxiliado na viabilização deste projeto.

RESUMO

Esta dissertação realizou, inicialmente, um estudo em caráter exploratório, com foco na pesquisa bibliográfica, para identificação do estado da arte em relação à composição de conteúdos educacionais para *e-Learning* em conformidade com o padrão SCORM. A partir deste estudo, percebeu-se que o modelo de metadados para agregação de conteúdos definido pelo padrão, apesar de amplamente utilizado, ainda se apresenta como elemento complexo e de difícil utilização por parte de pedagogos, desenvolvedores de conteúdos e *designers instrucionais*. Em particular, a identificação de conteúdos relacionados entre si, a partir de grandes repositórios, e sua agregação empregando os metadados da categoria *relation*, tal como definida no SCORM, têm sido alvo de consideráveis esforços por parte de pesquisadores da área da computação em busca da automatização deste processo. Neste sentido, diferentes abordagens têm sido propostas. Ao se verificar que tais abordagens acabaram por estender ou mesmo alterar os metadados definidos pelo padrão, foi conduzida, como parte deste trabalho de dissertação, uma pesquisa experimental com o objetivo de se propor uma metodologia que emprega ontologias, anotação automática de metadados, recuperação de informação e mineração de textos para identificação e agregação de conteúdos relacionados, utilizando-se os metadados da categoria *relation* tal como definidos em suas especificações. Foi implementado o protótipo de um sistema computacional, que aplica a metodologia proposta sobre uma amostragem de objetos de aprendizagem e gera os resultados necessários à avaliação de sua eficácia frente ao problema apresentado. Os resultados obtidos foram analisados e avaliados com o apoio de profissionais da pedagogia, que atuam no desenvolvimento de conteúdos para *e-Learning*, demonstrando que a metodologia proposta é viável e eficaz, produzindo os resultados esperados.

Palavras-chave: SCORM. Recomendação automática de conteúdos. Agregação de conteúdos. Objetos de aprendizagem. Recuperação de informação. Mineração de textos.

ABSTRACT

In this dissertation, initially, an exploratory study was conducted, focusing on literature search, to identify the state of the art in relation to the composition of educational content for e-Learning in accordance with the SCORM standard. From this study, it was noticed that the metadata model for aggregating content defined by SCORM, although widely used, is still considered complex and difficult to be used by educators, content developers and instructional designers. Particularly, the identification of contents related with each other, from large repositories, and their aggregation using the relation metadata category, as defined in SCORM, have been the focus of considerable efforts by researchers in the field of computing in pursuit of automation of this process. In this regard, various approaches have been proposed. As such approaches have extended or altered the metadata defined by SCORM standard, an experimental study was conducted as part of this dissertation, in order to propose a methodology which employs ontologies, automatic annotation of metadata, information retrieval and text mining for identification and aggregation of related content, using metadata category "relation" as defined in their specifications. We developed a prototype of a computer system that apply the proposed methodology on a sample of learning objects and generates the necessary results to evaluate their efficacy faced with the problem presented. The results were analyzed and evaluated with the support of professionals in the field of pedagogy, who works on the development of content for e-Learning, demonstrating that the proposed method is feasible and effective, producing the expected results.

Keywords: SCORM. Automatic content recommendation. Content aggregation. Learning objects. Information retrieval. Text mining.

LISTA DE FIGURAS

LISTA DE TABELAS E QUADROS

LISTA DE SIGLAS

ADL - Advanced Distributed Learning
ANNIE - A Nearly-New Information Extraction system
API - Application Programming Interface
CAM - Content Aggregation Model
EAD – Educação a Distância
GATE - General Architecture for Text Engineering
GB - Gigabyte
GHz - Giga-hertz
HTML - HyperText Markup Language
hylOs - Hypermidia Learning Objects System
IDE - Integrated Development Environment
IDF - Inverse Document Frequency
IDT - Instructional Design Theory
IEEE - Instituto de Engenheiros Eletricistas e Eletrônicos
JAPE - Java Annotation Patterns Engine
KD - Knowledge Discovery
KDD - Knowledge Discovery in Databases
KDT - Knowledge Discovery from Text
LOM - Learning Object Metadata
MT - Mineração de Textos
OA - Objetos de Aprendizagem
OWL - Web Ontology Language
P2P - Peer-to-peer
PDF - Portable Document Format
RAM - Random Access Memory
RDF - Resource Description Framework
RI - Recuperação de Informação
RST - Rhetorical Structure Theory
RTE - Run-Time Environment
RTF - Rich Text Format
SCORM - Sharable Content Object Reference Model
SGA - Sistemas de Gerenciamento de Aprendizagem
SN - Sequencing and Navigation
TF - Term Frequency
UIMA - Unstructured Information Management Architecture
UML - Unified Modeling Language
XHTML - eXtensible Hypertext Markup Language
XML - eXtensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, a Educação a Distância (EAD) mediada pela Web, também conhecida como *e-Learning*, instituiu-se como modalidade de ensino reconhecida e aplicada em diferentes partes do mundo, para os mais diversos fins educacionais. Sendo assim, passou a demandar esforços por parte das instituições envolvidas, no sentido de se definir normas e padrões de *software* que lhe dessem suporte. Em resposta, foi desenvolvido pela *Advanced Distributed Learning* (ADL), o padrão denominado *Sharable Content Object Reference Model* (SCORM)[1] (ADVANCED DISTRIBUTED LEARNING, 2009a), que possibilita a publicação de Objetos de Aprendizagem (OAs) na Web, por meio de Sistemas de Gerenciamento de Aprendizagem (SGA).

Em suas especificações, o SCORM define um Modelo de Agregação de Conteúdos baseado no *Learning Object Metadata* (LOM), desenvolvido pelo Instituto de Engenheiros Eletricistas e Eletrônicos (IEEE). No entanto, lidar com seu extenso e complexo modelo de metadados pressupõe um processo de anotação dispendioso se realizado a partir de um esforço unicamente humano, o que não obstante resulta em metadados insuficientes, quando não incorretos, como observado por Edvardsen e outros (2010), comprometendo a qualidade dos OAs e limitando a utilização dos recursos oferecidos pelo padrão. Dentre estes, a capacidade de extensão de seu conteúdo principal a partir da indicação de OAs correlatos é um exemplo importante.

O presente trabalho apresenta uma metodologia que emprega técnicas de Recuperação de Informação (RI) e Mineração de Textos (MT) para recomendação automática e agregação de OAs relacionados, conforme o SCORM. Foi utilizado o vocabulário da categoria *relation* para identificar relações dos tipos *requires* e *isrequiredby*, *ispartof* e *haspart*, *references* e *isreferencedby*, *isbasedon* e *isbasisfor*, sem a necessidade de extensão de seus metadados, alterações no SCORM ou implementações específicas em SGAs, diferentemente de outros trabalhos referenciados na literatura e discutidos na Seção 3 deste trabalho. Ainda neste sentido, adotou-se uma perspectiva segundo a qual conteúdos relacionados são recomendados aos autores de conteúdos, a partir de um conteúdo de referência, auxiliando-os na composição de um dado curso ou disciplina. Para que pudesse ser testada e avaliada, esta metodologia foi implementada na forma de um Sistema de Recomendação e Agregação de Conteúdos

---

[1] A primeira versão foi finalizada em 2000, com o SCORM 1.0. O SCORM 1.2 foi finalizado em 2001. A partir de 2004 foram publicadas diferentes edições do SCORM 2004, sendo que a mais recente é a de 2009: *SCORM 4th Edition*.

Relacionados e aplicada sobre uma amostragem de OAs, extraída de um repositório de conteúdos organizado para este trabalho.

Os resultados obtidos são positivos e foram verificados com o apoio de profissionais da pedagogia que atuam no desenvolvimento de conteúdos para *e-Learning*. Diante disso, conclui-se que a metodologia proposta é viável e eficaz, produzindo os resultados esperados, além de ser aplicável à construção de conteúdos didático-pedagógicos pertencentes a diferentes áreas do conhecimento.

1.1 Motivação

O SCORM fornece mecanismos para organização e estruturação de conteúdos didático-pedagógicos na forma de OAs, que podem ser mantidos em repositórios de conteúdos e reutilizados para composição de novas unidades de aprendizagem. Segundo Su e outros (2006), dentre os padrões internacionais relacionados a conteúdos para *e-Learning*, o SCORM tornou-se o mais utilizado. Sua grande aceitação, de acordo com Rey-López e outros (2009), se deve ao fato de que este reúne diversas padronizações, de diferentes instituições, possibilitando uma grande variedade de aplicações, com conteúdos reutilizáveis, adaptáveis e facilmente portáveis entre SGAs que tenham implementadas suas especificações. Redondo, Vilas e Arias (2012) se referem ao SCORM com sendo, de fato, o padrão para *e-Learning.*

No entanto, uma das maiores dificuldades enfrentadas por desenvolvedores de conteúdo, pedagogos e *designer instrucionais*, ao utilizarem o SCORM, consiste em lidar com seu extenso e complexo modelo de metadados, pois pressupõe um processo de anotação que pode se tornar dispendioso, longo e cansativo se realizado a partir de um esforço unicamente humano (MARGARITOPOLOUS; MANITSARIS; MAVRIDS, 2007). Para Maratea, Petrozino e Manzo (2012), o SCORM apresenta como principal vantagem a reutilização de OAs, mas ao custo de uma anotação de metadados difícil, lenta e expansiva, o que, para Edvardsen e outros (2009), pode comprometer a qualidade dos OAs, além de exigir conhecimento especializado para criação dos metadados necessários.

Sendo assim, segundo Roy, Sudeshna e Sujoy (2008), muitos autores de conteúdo para *e-Learning* são relutantes frente à necessidade de lidar com anotações e metadados, por acharem a tarefa desinteressante e trabalhosa, desestimulando a utilização de padrões como o SCORM. Além disso, como observado por Edvardsen e outros (2009), a geração de metadados de qualidade requer pessoas habilidosas, além de ser um processo custoso que muitas vezes apresenta, como resultado, metadados insuficientes, quando não incorretos.

Diante disso, observa-se a importância de se utilizar padrões bem definidos para publicação de conteúdos para *e-Learning*, mas ao mesmo tempo fica evidente a dificuldade inerente à adoção de um padrão cujas especificações e modelos devem ser adotados e devidamente seguidos. Sendo assim, esforços no sentido de se encontrar estratégias e ferramentas capazes de facilitar a construção de OAs em conformidade com o SCORM, tornando mais amigável o emprego dos recursos por ele oferecidos, são de visível relevância para o avanço da educação e para a diminuição das fronteiras de tempo e espaço, por meio do compartilhamento do conhecimento através da Web.

Neste sentido, a proposição de uma metodologia para recomendação automática de conteúdos relacionados, como é feito neste trabalho, que utiliza a anotação automática de metadados aliada a um conjunto de técnicas de RI, MT e uma base de conhecimento de domínio para conceituação e classificação de uma determinada área do conhecimento pode, então, oferecer mecanismos que façam frente a algumas das principais dificuldades encontradas no desenvolvimento de conteúdos educacionais a partir de OAs relacionados. Para isso, a acurácia alcançada nos processos de anotação automática e a qualidade do conjunto de metadados obtidos, em conjunto com a eficácia das técnicas de RI e MT empregadas, são de fundamental importância, pois seus resultados são a referência primeira para que se possa identificar e estabelecer relações entre conteúdos. Além disso, a recomendação automática, por si, se apresenta como estratégia cuja perspectiva se inverte da pesquisa por OAs para a composição de conteúdos com o apoio de um sistema de recomendação, o que diminui o esforço necessário para se pesquisar conteúdos e identificar aqueles que estejam relacionados entre si.

De acordo com a literatura, a recuperação de informação sobre dados textuais, assim como a identificação de relações entre seus conteúdos, consistem em matéria para qual ainda há muito que se contribuir, no sentido de se ampliar sua aplicação, além de melhorar sua eficiência e acurácia. Assim, buscar meios que auxiliem os autores de conteúdos a se preocuparem mais com a qualidade dos mesmos e menos com sua manufatura, contribuir para com a utilização de um padrão relevante como o SCORM e identificar como as técnicas de RI e MT podem contribuir neste sentido, foram as principais motivações para o desenvolvimento deste trabalho de pesquisa.

1.2 Definição do problema

Muitos recursos para composição de OAs definidos pelo SCORM não são efetivamente utilizados devido à dificuldade inerente aos seus metadados. Dentre eles, a capacidade de extensão de um dado conteúdo a partir da indicação de OAs correlatos é um exemplo importante, como pode ser lido em Lu e outros (2010):

> Uma relação, na categoria "*relation*", é utilizada principalmente para descrever um OA e expressar relacionamentos entre eles. Quando utilizada com habilidade, uma relação torna-se um metadado muito útil, que pode elevar a efetividade do aprendizado, assim como aumentar a reusabilidade de OAs. (LU et al., 2010, tradução nossa).[2]

A categoria de metadados *relation* classifica diferentes formas de relações entre conteúdos por meio do seguinte vocabulário, proposto no LOM: *ispartof*, *haspart*, *isversionof*, *hasversion*, *isformatof*, *hasformat*, *references*, *isreferencedby*, *isbasedon*, *isbasisfor*, *requires*, *isrequiredby*. Conforme consta na documentação do SCORM, há uma definição de como a categoria de metadados *relation* está organizada, mas não há um modelo estabelecido para sua utilização, ao contrário do que ocorre com outras categorias. É indicado no documento que a utilização destes metadados, assim como a maneira com que isso será feito, seja definida pelos próprios desenvolvedores de conteúdos ou por cada SGA em que se queira lhes oferecer suporte: "organizações ou comunidades de desenvolvimento são incentivadas a identificar os casos de uso e requisitos para utilização de tal mecanismo de associação de metadados." (ADVANCED DISTRIBUTED LEARNING, 2009a, tradução nossa).[3]

Diante disso, percebe-se que há duas questões importantes a serem observadas: a dificuldade inerente à anotação manual de metadados e a falta de uma definição específica para utilização da categoria *relation* no intuito de se construir conteúdos para *e-Learning* a partir de OAs relacionados.

Sendo assim, este trabalho de pesquisa procurou apresentar uma solução para a seguinte questão: é possível estabelecer uma metodologia para recomendação automática de conteúdos correlacionados, que possibilite identificar e estabelecer relações entre OAs, por meio da categoria *relation*, sem que seja necessário alterar seus metadados, modificar o padrão SCORM ou recorrer a implementações customizadas em SGAs?

---

[2] A relation in the "RELATION" category is mainly used to describe a LO and express relationships between Los. When used skillfully, a relation is a very useful metadata that can enhance learning effectiveness as well as increase the usability of LOs.

[3] Organizations or communities of practice are encouraged to identify their use cases and requirements for using such a mechanism for associating metadata.

1.3 Hipótese

O estabelecimento de relações entre OAs, conforme definido pela categoria *relation* do SCORM, sem alterações no padrão, é possivel de ser realizado a partir um sistema de recomendação automática de conteúdos relacionados, utilizando-se de estratégias de RI e MT. Para isso, deve-se adotar uma metodologia que empregue uma base de conhecimento de domínio, que seja capaz de modelar e conceituar a área de conhecimento pertinente aos OAs, de forma que seus conteúdos possam ser caracterízados a partir dos principais conceitos neles presentes e que as relação possam ser, então, estabelecidas entre os conteúdos de um ou mais OAs.

1.4 Justificativa

O *e-Learning* surgiu em resposta à necessidade de se ampliar o acesso das pessoas ao conhecimento e das instituições às pessoas, a quem desejem transmiti-lo. Nos últimos anos, instituiu-se como modalidade de ensino reconhecida e aplicada mundialmente. Universidades, centros de treinamento profissional e instituições governamentais e privadas têm elaborado e publicado materiais didático-pedagógicos, via Web, para os mais diversos fins educacionais. Todavia, o acesso à informação, assim como a ampliação e democratização do ensino são fatores que ainda apresentam grandes lacunas, principalmente em países como o Brasil (SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 2006).

Diante disso, esforços no sentido de facilitar e tornar mais ágeis os processos de desenvolvimento e publicação de conteúdos para *e-Learning* se fazem necessários. Assim, o trabalho realizado, como parte desta pesquisa, apresenta uma metodologia que emprega recursos da computação para reduzir o tempo e o esforço necessários ao desenvolvimento e agregação de conteúdos, além de contribuir para a expansão e efetiva utilização do SCORM, que se constitui como um padrão de referência mundial para publicação de OAs via Web.

1.5 Objetivos

Esta pesquisa teve como objetivo geral a proposição de uma metodologia que emprega ontologias, anotação automática de metadados, recuperação de informação e mineração de textos, para recomendação automática de OAs relacionados, em conformidade com a categoria de metadados *relation*, tal como definida pelo padrão SCORM.

Foram objetivos específicos deste trabalho:

a) definir uma estratégia por meio da qual a categoria *relation* do SCORM possa ser efetivamente utilizada para descrever relações entre os conteúdos de diferentes OAs;
b) definir estratégias de utilização das técnicas de RI e MT, para automatização dos processos inerentes à anotação automática de metadados, classificação de conceitos e identificação de conteúdos relacionados;
c) implementar o protótipo de sistema que permita experimentar a metodologia proposta, no sentido de verificar sua eficácia diante do objetivo principal deste trabalho;
d) avaliar os resultados alcançados, buscando identificar as possibilidades de melhorias e oportunidades para a realização de novos trabalhos.

2 FUNDAMENTAÇÃO TEÓRICA

São três os elementos principais que fundamentam este trabalho de pesquisa: o padrão SCORM e seu modelo de agregação de conteúdos, os estudos e técnicas acerca da Recuperação de Informação e Mineração de Textos e o uso de ontologias como modelo de domínio para uma dada área do conhecimento. Dessa forma, este capítulo se dedica a uma breve discusão de cada um deles.

2.1 O padrão SCORM

A ADL surgiu a partir da necessidade do Departamento de Defesa dos Estados Unidos de oferecer treinamento sob demanda para suas unidades em todo o mundo, configurando-se em uma estratégia para unir processos de ensino-aprendizagem e tecnologias da informação. No *site* da ADL pode-se encontrar os principais objetivos desse consórcio de entidades:

> Identificar e recomendar as normas para *software* de treinamento e serviços associados adquiridos por órgãos federais e contratantes. Facilitar e acelerar o desenvolvimento dos principais padrões de treinamento técnico nas indústrias e nas entidades de desenvolvimento de padrões. Estabelecer diretrizes sobre o uso de padrões e fornecer um mecanismo para auxiliar o Departamento de Defesa e demais agências federais no desenvolvimento em larga escala, implementação e avaliação de sistemas de aprendizagem interoperáveis e reutilizáveis. (ADVANCED DISTRIBUTED LEARNING, 1999, tradução nossa).[4]

Para atender a esses objetivos, a ADL definiu três núcleos essenciais para o padrão SCORM: um *Modelo de Agregação de Conteúdo*, um *Modelo de Sequenciamento e Navegação* e um *Ambiente de Tempo de Execução*. Suas especificações resultaram em três documentos que compõem a 4ª edição da Suíte de Documentação do SCORM 2004.

O *Modelo de Agregação de Conteúdo* trata especificamente dos diversos elementos que irão compor o que se pode denominar como "pacote SCORM", ou seja, um conjunto de arquivos que acompanham os OAs e possibilitam sua disponibilização via SGA. Além disso, especifica e descreve elementos e metadados utilizados na organização, estruturação e sequenciamento do conteúdo. Todos os detalhes, requisitos e especificações referentes a este modelo estão descritos no *Content Aggregation Model* (CAM), ou simplesmente SCORM – CAM.

---

[4] Identify and recommend standards for training software and associated services purchased by Federal agencies and contractors. Facilitate and accelerate the development of key technical training standards in industry and in standards-development organizations. Establish guidelines on the use of standards and provide a mechanism to assist DoD and other Federal agencies in the large-scale development, implementation, and assessment of interoperable and reusable learning systems.

O *Ambiente de Tempo de Execução* trata da interface entre Objetos de Aprendizagem e os SGAs que lhes dão suporte. Desde a publicação de um pacote SCORM, passando pelo acesso do aluno ao seu conteúdo, incluindo diferentes formas de interação que podem ser registradas e monitoradas, até o término da seção de aprendizagem, há uma série de informações que podem ser trocadas entre os OAs e os SGAs. Além disso, há vários registros que podem ser coletados e armazenados de acordo com os objetivos dos diferentes interessados, tais como alunos, instituições de ensino e professores. Todos estes processos encontram suporte no O *Ambiente de Tempo de Execução*, cujas especificações e detalhes constam no documento intitulado *Run-Time Environment* (RTE), ou simplesmente SCORM – RTE.

O *Modelo de Sequenciamento e Navegação*, por sua vez, trata das transições entre os OAs de uma dada unidade de aprendizagem e da ordenação das atividades em um determinado caminho que sirva ao processo de ensino-aprendizagem pretendido. Sendo assim, este modelo define o comportamento requerido e as funcionalidades que um SGA deve implementar para processar as informações de navegação e sequenciamento de conteúdo. O documento publicado pela ADL, que traz os detalhes e requisitos deste modelo, é denominado *Sequencing and Navigation* (SN), ou simplesmente SCORM – SN.

A partir das especificações contidas nos documentos SCORM – CAM, SCORM – SN e SCORM – RTE é possível criar, agregar e entregar diversos tipos de conteúdo, organizados de inúmeras maneiras e com diferentes caminhos para navegação e interação por parte do aluno. Além disso, um mesmo pacote de conteúdo SCORM pode ser facilmente publicado, sem quaisquer alterações, em diferentes SGAs. Estas são algumas das principais características que o tornaram um padrão de referência para *e-Learning*. Todavia, algumas questões importantes têm sido levantadas com relação ao SCORM. A complexidade de sua estrutura e a dificuldade de compreensão e utilização dos metadados de seu modelo de agregação de conteúdos estão entre elas.

Quando da agregação de conteúdos em um pacote SCORM, sua estrutura e organização são definidas em um arquivo principal, denominado *imsmanifest.xml*. Este arquivo é o primeiro a ser interpretado pelo SGA, no momento em que o pacote SCORM é publicado. Com base nas informações obtidas, o sistema cria a estrutura visual que permite ao aluno acessar as unidades, componentes e OAs que compõem o conteúdo a ser estudado. A Figura 1 apresenta parte de um *imsmanifest.xml* e a Figura 2 apresenta a estrutura definida pelo SGA Moodle a partir dos dados presentes no arquivo, cujo conteúdo é mostrado na Figura 1.

Figura 1 - Parte inicial de um arquivo *imsmanifest.xml*

```
<?xml version="1.0" encoding="UTF-8"?>
<manifest xmlns="http://www.imsglobal.org/xsd/imscp_v1p1" xmlns:lom="http://ltsc.ieee.org/xsd/LOM" x
"http://www.w3.org/2001/XMLSchema-instance" xmlns:adlcp="http://www.adlnet.org/xsd/adlcp_v1p3" xmlns
"http://www.imsglobal.org/xsd/imsss" xmlns:adlnav="http://www.adlnet.org/xsd/adlnav_v1p3" identifier
"MANIFEST-FDC59F51-B25F-70CF-3923-B549F45F83F0" xsi:schemaLocation="http://www.imsglobal.org/xsd/ims
http://ltsc.ieee.org/xsd/LOM lom.xsd http://www.adlnet.org/xsd/adlcp_v1p3 adlcp_v1p3.xsd http://www.
imsss_v1p0.xsd http://www.adlnet.org/xsd/adlnav_v1p3 adlnav_v1p3.xsd" version="1.3">
  <metadata>
    <schema>ADL SCORM</schema>
    <schemaversion>2004 3rd Edition</schemaversion>
  </metadata>
  <organizations default="Teste">
    <organization identifier="Nome do curso" structure="hierarchical">
      <title>Curso de Teste Estrutura SCORM</title>

    <item identifier="MOD_1" isvisible="true">
      <title>"Unidade 1 - Título da Unidade I"</title>
        <item identifier="MOD_1_SCO_1" identifierref="RES_1" isvisible="true">
            <title>Tópico 1 - Titulo do tópico 1</title>
        </item>
        <item identifier="MOD_1_SCO_2" identifierref="RES_2" isvisible="true">
            <title>Tópico 2 - Titulo do tópico 2</title>
        </item>
```

Fonte: Elaborada pelo Autor

Figura 2 - Estrutura de um curso

Fonte: Elaborada pelo autor

Além da estrutura principal do conteúdo, também estão presentes no *imsmanifest.xml* os metadados que possibilitam definir diferentes formas de organização, sequenciamento,

navegação e agregação de conteúdos. No extenso modelo de metadados do SCORM, há nove categorias obtidas diretamente do LOM. Cada uma delas possui subcategorias e elementos específicos para apresentação dos dados necessários.

A categoria *general* contém elementos que descrevem um componente de conteúdo como um todo, através de elementos contendo aspectos tais como título, linguagem utilizada, palavras-chave, dentre outros. Informações relativas ao histórico e ao estado atual de um componente são registradas na categoria *lifecycle.* Na categoria *metametadata*, são encontrados os elementos que descrevem os próprios registros de metadados, permitindo identificar de que modo e em que contexto foi criada uma instância de metadados. A categoria *technical* contém elementos por meio dos quais são descritos os requisitos e as características técnicas de um dado componente. Consistindo em uma das maiores categorias, com o maior número de elementos, a categoria *educational* descreve as principais características didáticas e pedagógicas de um componente do conteúdo. Professores, gestores, autores e alunos são os principais interessados nas características descritas por seus elementos. Os direitos de propriedade intelectual são descritos nos elementos da categoria *rights* enquanto a categoria *relation* se apresenta, ao lado da *educational*, como uma das maiores e mais complexas. Nela estão presentes os elementos cujos metadados descrevem diferentes possibilidades de relações entre dois ou mais OAs, tais como a relação do tipo *isbasedon*, ilustrada na Figura 3. A categoria *annotation* permite a inserção de comentários acerca do uso didático-pedagógico previsto para um dado componente de conteúdo. A categoria *classification*, por sua vez, descreve como um componente de conteúdo se encaixa em um sistema de classificação específico, relacionando-o a um vocabulário controlado ou a um sistema de classificação.

Figura 3 - Relação do tipo *isbasedon*

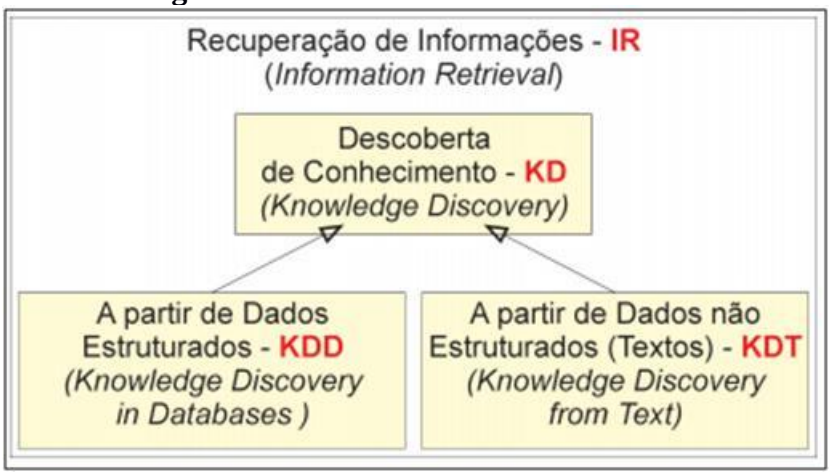
```
<lom>
   <relation>
      <kind>
         <source>LOMv1.0</source>
         <value>isbasedon</value>
      </kind>
      <resource>
         <identifier>
            <catalog>URN</catalog>
            <entry>urn:ADL:1234-45FD</entry>
         </identifier>
         <description>
            <string language="en">Microsoft MSCE</string>
         </description>
      </resource>
   </relation>
</lom>
```

Fonte: ADVANCED DISTRIBUTED LEARNING, 2009

2.2 Recuperação de infomação e mineração de textos

De acordo com Morais e Ambrósio (2007) e conforme visualizado na Figura 4, Mineração de Dados e Mineração de Textos constituem-se como importantes áreas de pesquisa relacionadas à descoberta de conhecimento, que utilizam técnicas da Recuperação de Informação em grandes volumes de dados. A primeira lida com dados estruturados cuja estrutura semântica se encontra organizada, por exemplo, em uma base de dados relacional. A segunda, mais recente, lida, por sua vez, com dados textuais, que se caracterizam como não estruturados ou semiestruturados, cuja organização está muito mais próxima da linguagem natural.

Figura 4 - Descoberta de Conhecimento



Fonte: Morais e Ambrósio, 2007

Segundo Baghela e Tripathi (2012), o termo Mineração de Dados se refere a "métodos de análise de dados com o objetivo de encontrar regras e modelos que descrevem as propriedades características dos dados." Para os autores, o crescimento exponencial do volume de dados textuais em formato eletrônico, impulsionado principalmente pela Web, traz este desafio da Mineração de Dados para o contexto da Mineração de Textos, no qual se busca uma forma eficiente de recuperar informações relevantes a partir deste tipo de dado, heterogêneo e não estruturado. Os autores destacam algumas características específicas da Mineração de Textos, tais como a análise de relações entre conceitos e a necessidade de uma etapa de pré-processamento linguístico para extração das características principais do texto em análise.

Conforme discutido em Yuan e Belkin (2010) há diferentes técnicas advindas da Recuperação de Informação que permitem identificar informações relevantes e que são

utilizadas, como apresentado por Morais e Ambrósio (2007), nos processos de Mineração de Textos. Neste sentido, a Recuperação de Informação se dedica a extrair, de um ou mais documentos, informações relevantes a algum tópico, assunto ou área de interesse, enquanto a Mineração de Textos procura identificar padrões, referências e similaridades em um conjunto de documentos.

Segundo Ebecken e outros (2003), citados por Morais e Ambrósio (2007), a análise de dados textuais pode ser realizada via análise semântica, análise estatística ou por uma combinação de ambas. Pela definição de Cordeiro (2005), citado por Morais e Ambrósio, a análise semântica visa identificar a função e a importância de um termo ou conceito na oração onde ocorre, utilizando-se de técnicas de processamento de linguagem natural. A análise estatística, por sua vez, procura avaliar a relevância de um dado termo a partir de métricas associadas à sua ocorrência em um texto ou em uma coleção de textos. A partir destes processos, é identificado então o conjunto de termos e conceitos que mais fortemente expressam o conteúdo total do texto onde se inserem. Em Morais e Ambrósio são descritos diferentes modelos para classificação e representação deste conjunto que, por final, é utilizado nos processos específicos da Mineração de Textos possibilitando, por exemplo, encontrar os documentos mais relevantes em relação aos dados de uma pesquisa, informados por um usuário, ou identificar o grau de similaridade entre dois ou mais documentos, dentre outras aplicações.

Chen, Liu e Ho (2013) desenvolveram um sistema de apoio à pesquisa por documentos jurídicos e legais, utilizando um sistema denominado CKIP para realização de uma etapa de pré-processamento dos documentos utilizados. O sistema oferece um serviço remoto para processamento de textos em Chinês, que realiza a segmentação do texto, remoção de termos irrelevantes, e análise sintática e gramatical dos demais termos e palavras presentes nos textos. Em seguida, utilizaram dois métodos clássicos para computar o peso de cada um dos termos presentes nos documentos: o cálculo da Frequência do Termo (TF, *Term Frequency*) e o cálculo da Frequência Inversa de Documentos (IDF, *Inverse Document Frequency*). Em seguida, foi gerado um *ranking* dos termos de maior peso, obtendo-se, assim, um conjunto de treinamento. Os resultados destes processos foram submetidos à análise de especialistas e reajustados com base em suas observações. Para realização das consultas por parte dos usuários, os autores utilizaram o método *Google Similarity Distance*, para transformar os termos da entrada do usuário em termos relacionados aos termos legais obtidos nas etapas anteriores e que compõem sua base de treinamento. Este passo de transformação dos termos permitiu que as pesquisas se tornassem mais amigáveis para o usuário, pois

possibilitou que buscas fossem feitas fornecendo-se como entrada termos mais populares e menos específicos do jargão do direito. Segundo os autores, os resultados dos testes realizados demonstraram que a metodologia por eles proposta apresentou-se mais eficiente que a abordagem na qual se utiliza apenas os métodos TF-IDF.

Os processos de RI em documentos de texto envolvem análises linguísticas (FOREST; SYLVA, 2011). Assim, são recorrentes os conceitos específicos desta área do conhecimento, tais como *lexicons, corpora e corpus*, ontologias, *stop words*, dentre outros. Bontcheva e outros (2004) apresentam a forma como alguns destes conceitos são comumente entendidos e aplicados. Basicamente, *lexicons* são dicionários de termos que compõem uma dada linguagem em um idioma específico, e cumprem um papel importante para análise semântica e processamento de linguagem natural. *Corpus*, singular de *corpora*, indica uma coleção de documentos sobre os quais serão aplicados os recursos de processamentos necessários à recuperação de informação. Ontologias oferecem informações sobre classificação de conceitos de um dado domínio de conhecimento, a serem empregadas em análises realizadas por recursos de processamento que busquem, por exemplo, obter a distância semântica entre dois conceitos. Por fim, *stop words* são termos essencialmente não relevantes por não trazerem informações em si, como no caso de artigos e conjunções (CUNNINGHAM, 2012).

2.3 Ontologia de domínio

Para Araújo e Ferreira (2003), sem a conceituação do conhecimento, não há um vocabulário capaz de representá-lo. A simples identificação de um termo ou conceito no texto de um documento, não significa que se tenha obtido uma informação relevante ao seu conteúdo, pois, para que o seja, é necessário que se tenha como parâmetro a área de conhecimento com a qual o termo em questão precise estar relacionado, o que requer conhecimento do domínio, ou seja, uma ontologia que o represente. Neste sentido, as ontologias representam um domínio de interesse, por meio de um conjunto de conceitos a ele relevantes e das relações entre eles, possibilitando seu entendimento de forma compartilhada. Para os autores, a utilização de ontologias permite estrutura e organizar as informações, tornando os processos de busca por conteúdo mais inteligentes e eficientes. Quando utilizadas na modelagem de OAs, para organização e estruturação de materiais de aprendizagem, propiciam maior reaproveitamento de seu conteúdo.

A conceituação presente em uma ontologia fornece uma representação abstrata de uma entidade do mundo real, enquanto sua formalização a torna compreensível para uma máquina,

permitindo o compartilhamento do conhecimento (BHOWMICK, 2010). Conforme a estratégia adotada, a ontologia de domínio cumpre um papel fundamental para a recuperação da informação, estabelecendo as relações existentes entre os diversos termos e conceitos presentes em uma base conhecimento, que são utilizados para identificação de informações relevantes. Neste sentido, segundo Hernández e outros (2009), o usa de ontologias permite que se identifique relações semânticas entre os diferentes conceitos de uma área de domínio, possibilitando identificar, através destes conceitos, textos ou partes de um texto que possam ser associados a objetos de aprendizagem.

Para Hernández e outros (2009), ao se tentar recuperar informações em arquivos que contenham determinados conteúdos, uma das principais questões consiste em como se identificar nos mesmos a informação que se está buscando. Para os autores, a utilização de buscas tradicionais, realizadas apenas por palavras-chave, não leva em conta os sinônimos dos termos da busca e não apresenta suporte a diferentes idiomas, tornando os resultados menos precisos. Segundo eles, para que se possa tornar o resultado de uma busca mais preciso e capaz de identificar corretamente os conceitos presentes em um documento de interesse, fazendo com que seja encontrada a informação requerida, é necessário que se utilize processos que realizem buscas semânticas. Estas, por sua vez, segundo Hernández e outros, se baseiam na utilização de dicionários de termos, ontologias e modelos de domínio. Para eles, um modelo de domínio se refere a uma área de interesse e tem como objetivo definir objetos e entidades que a ela pertençam e que se deseje representar. Dessa forma, têm-se a definição do tema sobre o qual se deseje realizar uma busca.

Para Borges e Barbosa (2009), a modelagem de conteúdos é parte fundamental do processo de construção de conteúdos educacionais, onde o uso de ontologias no nível conceitual permite compreender o domínio de dada área de conhecimento para a qual se deseje construir módulos de aprendizagem. Segundo eles, uma ontologia é uma representação formal e declarativa do conhecimento, incluindo um vocabulário que se refere os conceitos pertencentes ao domínio, sua organização lógica e como eles estão relacionados. Sendo assim, utilizaram em seu trabalho uma ontologia de domínio para a área de conhecimento relacionada a teste de software, buscando melhor compreensão do domínio, fácil compartilhamento de conhecimento entre autores e o estabelecimento de uma estrutura bem definida para um repositório de conteúdos, estabelecendo uma representação formal de seus objetos de aprendizagem.

3 TRABALHOS RELACIONADOS

Diferentes abordagens têm sido propostas para recuperação de informação sobre dados textuais. Algumas se apresentam viáveis em contextos específicos, outras buscam aplicações mais abrangentes; algumas se destinam a materiais livremente disponibilizados na Web, outras abordam materiais disponíveis em repositórios distribuídos. Da mesma forma, diferentes técnicas e metodologias podem ser aplicadas para solução do problema, utilizando-se de modelos de metadados existentes para então propor extensões a estes modelos; criando-se ontologias para classificação de conteúdos e documentos; recorrendo-se a sistemas baseados em regras, ou aplicando-se algoritmos de aprendizagem de máquina e compreensão de linguagem natural para processamento dos modelos de classificação, anotação e extração de metadados. Há trabalhos que buscam implementar sistemas específicos para tratamento de modelos criados para um dado contexto, outros propõem soluções portáveis baseadas em padrões internacionais, tais como o *Dublin Core*, o LOM e o SCORM. Da mesma forma, os objetivos variam, buscando maneiras eficazes de organização de conteúdos em repositórios, ou formas de tornar mais precisos e eficientes os processos de pesquisa e recuperação desses conteúdos.

Diante disso, para levantamento dos trabalhos relacionados a esta pesquisa, foram pesquisados, em periódicos, anais de eventos e bibliotecas digitais da área da computação[5], aqueles cujos focos de investigação se aproximavam deste trabalho ou estavam a ele relacionados. Foram selecionados aqueles que se dedicavam a empregar processos de geração e anotação automática de metadados, RI e MT, com destaque para aqueles que buscavam solucionar questões relativas a *e-Learning*, Objetos de Aprendizagem e ao SCORM, principalmente, aqueles que buscavam estabelecer relações entre OAs, com base na categoria *relation*. Foram levantados trabalhos relacionados ao longo de todo o período de desenvolvimento desta pesquisa.

Engelhardt e outros (2006) apresentaram uma abordagem, por eles denominada "geração semiautomática de objetos de aprendizagem", que se pauta na utilização de metadados para estabelecer relações existentes entre diversos OAs presentes em um repositório. O conjunto de metadados por eles utilizado foi elaborado com base em um subconjunto retirado do LOM, acrescido de outro conjunto que consiste de metadados por

---

[5] A maior parte das pesquisas foi realizada por meio do Portal de Periódicos da Capes, que permite às universidades conveniadas o acesso às bases de algumas das principais revistas, *journals* e anais de eventos. Também foram realizadas buscas nos principais periódicos nacionais relacionados à informática na educação, tais como o REBIE e SBIE, assim como diretamente na Web, por meio do Google.

eles propostos. A partir desse conjunto, os autores formalizaram as relações semânticas entre os OAs por meio de uma ontologia elaborada com base na *Web Ontology Language* (OWL), aplicável a um conjunto de regras de inferência.

Para Engelhardt e outros (2006), o problema maior para um autor decorre da necessidade de se inserir um OA em um repositório, momento em que se fazem necessárias a descoberta e a definição de relações entre o seu OA e aqueles que lá se encontram. A partir deste entendimento, os autores propuseram um processo automático para derivação destas relações a partir de um conjunto de relações iniciais, que permitem a aplicação de métodos de inferência. Tais relações iniciais podem ser geradas automaticamente ou inseridas manualmente. Ao tratar da ontologia aplicada, os autores afirmam que o LOM possui relações pouco expressivas, obtidas diretamente de outro modelo de metadados, denominado *Dublin Core*, e que, por isso, precisavam ser adaptadas no sentido de aumentar sua expressividade e de modo a contemplar o que chamaram de "hipermídia educacional". Para isso, dividiram seu trabalho em três etapas. Na primeira, selecionaram relações do *Dublin Core* que possibilitaram adequações utilizando-se o menor número possível de modificações e especificações para o contexto da hipermídia educacional. Na segunda etapa, considerando que as semânticas destas propriedades estavam relacionadas a termos puramente técnicos, os autores as redefiniram de forma explícita buscando, segundo eles, torná-las semanticamente fortes o suficiente para que fossem capazes de expressar dependências e relações entre temas. Por fim, propuseram um conjunto de propriedades de relações adicionais, não contempladas pelo LOM.

A partir da estrutura de classificação definida por Engelhardt e outros (2006), foi estabelecida uma rede semântica para todo o repositório, gerando *links* entre os OAs de modo a expressar as relações identificadas entre eles. Tais relações são estabelecidas quando um novo OA é inserido no repositório, momento em que a rede semântica é reavaliada, novas relações são criadas e outras atualizadas para acomodar o novo objeto. Assim, ao acessar um dado OA, o aluno encontra inúmeras possibilidades de navegação por conteúdos relacionados, através da interligação dos diversos OAs que compõem a rede. As soluções apresentadas foram implementadas em um SGA desenvolvido pelos autores, denominado *Hypermidia Learning Objects System* (hylOs). Sua implementação utilizou o *framework* Jena para processamento da ontologia e das regras de inferência.

O trabalho realizado por Engelhardt e outros (2006) realmente permitiu a construção de uma gama de novas relações em sua rede semântica, a cada OA nela inserido. Todavia, ao se pensar em processos que necessitem da definição de uma unidade de aprendizagem,

disciplina, curso ou treinamento, cujo conteúdo seja específico e demande uma ordenação objetiva de conceitos e tópicos pré-estabelecidos, com base em um planejamento didático-pedagógico, tal solução pode não ser adequada. Caso o que se pretenda seja fornecer ao aluno uma fonte de pesquisa, uma biblioteca digital que o auxilie na busca de conhecimento e conteúdos diversos, sua abordagem oferece recursos interessantes. Mas, ao se pensar em um conteúdo estruturado frente a um objetivo educacional específico, a mesma fornece mecanismos que levam a um acesso muito disperso ao seu conjunto de variados OAs, haja vistas que todos os documentos estão previamente inter-relacionados e o acesso ao repositório é feito a partir de uma consulta por parte do aluno. Isso pode fazer com que o número de possibilidades de navegação seja muito grande e diversificado, pois as relações são indicadas para cada OA acessado.

Assim como Lu e outros (2010) e Edvardsen e outros (2009), Engelhardt e outros (2006) utilizam uma extensão ao modelo de metadados definido pelo LOM e um SGA específico para a realização de seu trabalho. Todavia, falta uma abordagem capaz de utilizar metadados para classificar e relacionar OAs sem quaisquer intervenções com relação ao modelo definido pelo SCORM, e sem a necessidade de utilização de um SGA específico para tal, garantindo assim a portabilidade e compatibilidade dos OAs em diferentes SGAs.

Edvardsen e outros (2009) propuseram uma abordagem que utilizou a geração automática de metadados para auxiliar na organização de conteúdos didático-pedagógicos, visando à construção de OAs no modelo SCORM. Neste sentido, desenvolveram um *framework* para geração de metadados em conformidade com o LOM a partir de um dado OA, obtendo como resultado um novo OA, porém no formato SCORM. Para isso, utilizaram processos de recuperação de metadados contextuais com base em informações presentes no SGA e em entidades extraídas a partir dos próprios OAs, combinando diferentes abordagens para geração automática de metadados a partir de seus respectivos conteúdos.

Para recuperação dos metadados contextuais que, segundo Edvardsen e outros (2009), descrevem o contexto em que os OAs foram publicados e não os OAs propriamente ditos, foram utilizadas informações disponíveis no SGA da universidade onde atuam. Também recorreram a informações referentes aos cursos ofertados, presentes em um catálogo mantido pela universidade. A partir destes dados, foram obtidos os metadados relativos ao contexto específico dos cursos e da própria universidade. Os metadados relativos às informações de publicação de cada OA, por sua vez, foram gerados automaticamente com base nos dados de *login* do usuário no SGA, que trazem informações a respeito do próprio usuário, do tipo de curso a que se destina o OA publicado e a data de publicação. Todos estes metadados foram,

por fim, referenciados por elementos previstos no LOM. Todavia, suas fontes denotam grande dependência de elementos externos aos OAs e da forma como estes dados são apresentados. Isso porque, em sua maioria, não foram extraídos diretamente dos conteúdos dos OAs.

A extração dos metadados a partir dos OAs foi realizada pelos autores com base na análise do código dos documentos, pois assim, segundo eles, pode-se ter acesso diretamente ao conteúdo intelectual do mesmo, evitando-se a interpretação de elementos inseridos pelas aplicações que os geraram. Desta forma, recorreram a diferentes algoritmos para obtenção de diferentes conjuntos de metadados, obtidos a partir de informações estruturais, visuais ou semânticas.

Apesar de acreditarem que sua pesquisa apresentou resultados inconclusivos com relação à qualidade destes metadados, Edvardsen e outros (2009) concluíram que os SGAs, por sua vez, podem ser utilizados não apenas para publicação dos OAs, mas também como fonte do que eles chamaram de metadados contextuais, que podem ser utilizados como base para geração de metadados específicos para os próprios OAs.

A utilização de informações advindas de fontes externas aos OAs, como no caso do SGA da universidade e do catálogo de cursos, ofereceu recursos adicionais para recuperação e extração de metadados, como demonstrado por Edvardsen e outros (2009). Porém, a abordagem proposta pode restringir, dessa forma, seu escopo de aplicação ao depender de um *framework* desenvolvido para utilizar-se de informações que podem não estar disponíveis em outros sistemas e instituições, fazendo com que o mesmo conjunto de metadados possa não ser obtido em outro contexto. Como resultado da execução do *framework* proposto pelos autores, foi gerado um pacote de conteúdos SCORM contendo metadados dentre os quais, boa parte, não foi obtida diretamente dos OAs que o compõem. Outra questão importante a ser levantada, além do escopo específico de execução do *framework*, reside na utilização destes metadados, o que não foi indicado no trabalho dos autores. Sendo assim, ao se pensar em repositórios de conteúdos, o grande volume de metadados obtidos por Edvardsen e outros pode auxiliar nos processos de armazenamento, classificação e recuperação de OAs. Porém, não há uma proposta clara para utilização destes metadados em um sentido que vai além do registro ou consulta destas informações.

Roy, Sudeshna e Sujoy (2008) também estendem os metadados previstos pelo LOM, em sua categoria *educational*. Em seu trabalho, os autores definiram uma estratégia que utiliza anotação automática de OAs, disponíveis em repositórios de conteúdo, no intuito de possibilitar aos SGAs a seleção apropriada de material de aprendizagem, além de facilitar o trabalho de desenvolvedores de conteúdo no reaproveitamento deste material. Neste sentido,

desenvolveram uma ontologia cujos atributos pudessem caracterizar os materiais de aprendizagem de um ponto de vista pedagógico. Esta estrutura compôs sua base de conhecimento de domínio, que foi organizada hierarquicamente em três camadas, denominadas respectivamente por *term layer*, *concept ontology* e *topic taxonomy*. Diversos termos, presentes na primeira camada, foram associados a conjuntos de conceitos que os referenciam, presentes na segunda camada. Estes conceitos, por sua vez, permitiram identificar assuntos relacionados com a camada *topic taxonomy*.

Tanto o modelo de ontologia, desenvolvido por Roy, Sudeshna e Sujoy (2008), quanto à estratégia para anotação automática de OAs definida por eles, se apresentam como uma proposta consistente e viável para seleção de OAs e reaproveitamento de materiais didático-pedagógicos. Todavia, assim como percebemos em Lu e outros (2010), Edvardsen e outros (2009) e Engelhardt e outros (2006), os autores utilizaram metadados que estendem aqueles previstos na categoria e*ducational.* Além disso, sua abordagem oferece bons mecanismos que permitem a classificação e recuperação de OAs em repositórios, mas não tem como objetivo o estabelecimento de relações entre eles.

Nauerz e outros (2008) apresentaram uma estratégia semelhante àquela adotada por Roy, Sudeshna e Sujoy (2008), ao utilizarem-se de conjuntos de termos para identificar determinadas entidades representativas em um dado conteúdo. Para os autores, uma das maiores dificuldades enfrentadas por usuários da internet consiste em encontrar conteúdo relevante para suas pesquisas. Segundo eles, os usuários necessitam buscar o que denominam por "*background information",* ou seja, conteúdos que lhes ofereçam informações adicionais ou complementares, para melhor compreensão daquilo que de fato estão pesquisando. Neste sentido, uma das principais causas para a dificuldade apontada consiste no crescimento rápido da quantidade de informações disponibilizadas na Web, por vários e diferentes atores, exigindo um esforço para busca e verificação das informações, que pode se tornar uma tarefa cansativa e tediosa, como afirmam os próprios autores. Diante disso, buscando um sistema de recomendação a partir da interação do usuário em sistemas Web, propuseram um *framework* para anotação de conteúdos relacionados, em arquivos do tipo *eXtensible Hypertext Markup Language* (XHTML), que utiliza os serviços de análises de dados não estruturados, denominados UIMA e Calais. Utilizando estes componentes, o sistema por eles proposto foi capaz de analisar automaticamente um dado conteúdo e identificar determinados termos capazes de descrever certos tipos de "entidades", referentes a pessoas, localizações, empresas, dentre outras. Assim, os relacionamentos entre os conteúdos são descritos por meio de *tags* semânticas que contêm tais entidades que, por sua vez, podem ser ligadas a serviços

correlatos, como no caso de uma entidade do tipo “local” e o serviço do *Google Maps*.

Em uma experiência anterior, Nauerz e outros (2008) perceberam que seu sistema, tal como havia sido desenvolvido até então, produzia um grande número de recomendações irrelevantes. Segundo eles, este problema devia-se ao fato de que os processos de geração das *tags* semânticas eram realizados diretamente sobre o conteúdo, com base apenas em seu texto, identificando certos tipos de informação sem levar em conta os interesses e preferências do usuário. Para contornar este problema, os autores introduziram um modelo de usuário que fornece dados sobre seus interesses e, a partir deste modelo, selecionaram os fragmentos de informação para geração das *tags*. Em síntese, os termos utilizados para se construir as entidades de relacionamentos foram obtidos a partir do processo automático de extração e anotação de metadados, que utilizou como referência um modelo de usuário, permitindo a identificação de conteúdos relacionados cuja informação fosse realmente de seu interesse.

A abordagem proposta por Nauerz e outros (2008) se mostra como excelente solução para o que se pode entender como uma máquina de busca para conteúdos relacionados na Web. Porém, o processo de recomendação decorrente de tal abordagem, à semelhança do que ocorre em Engelhardt e outros (2006), na construção de seu repositório, pode não oferecer o suporte necessário quando o objetivo for a construção de uma unidade de aprendizagem coesa, cujo conteúdo seja composto de OAs agrupados não apenas por estarem relacionados, mas em conformidade com uma estrutura organizacional de cunho didático-pedagógico.

Percebe-se que os problemas levantados por Roy, Sudeshna e Sujoy (2008), Edvardsen e outros (2009) e Lu e outros (2010), dentre outros autores, com relação ao conteúdo para *e-Learning*, se referem principalmente à dificuldade de se identificar OAs cujos conteúdos estejam de alguma forma relacionados. A possibilidade de se criar relações entre conteúdos, de modo que um OA possa referenciar outros a partir de um repositório, pode ser pensada de modo que as referências estejam disponíveis diretamente a partir de cada um dos OAs que componham um mesmo conteúdo em um pacote SCORM. Desta forma, à medida que um dado OA fosse acessado, este apresentaria, em um campo específico, uma listagem de outros OAs que estariam a ele relacionados. Esta foi a premissa a partir da qual Lu e Hsieh (2009) propuseram a utilização da categoria *relation* do SCORM – CAM.

Porém, para Lu e Hsieh (2009), as relações descritas pela categoria *relation* são limitadas. Segundo eles, tais relações conseguem descrever apenas relacionamentos orientados pela estrutura do conteúdo, não sendo capazes de estabelecer relações semânticas entre os OAs. Diante disso, conforme exposto pelos autores, novas relações foram propostas por outros pesquisadores, mas, mesmo sendo capazes de estabelecer relações semânticas,

ainda assim apresentam limitações. Quando não redundantes, segundo Lu e Hsieh, tais relações são por vezes equivocadas, ou inapropriadas, a partir do momento em que não são claras quanto ao que realmente precisam expressar. Além disso, até que ponto podem ou não auxiliar o aluno em seu aprendizado constituía-se em um fator, até então, não mensurado formalmente.

Uma vez identificado tal contexto, Lu e Hsieh (2009) realizaram um trabalho de revisão sobre os modelos de relações anteriormente propostos. Com base na análise destes modelos, excluíram as relações duplicadas e desenvolveram um modelo de extensão aos metadados da categoria *relation* definida no SCORM - CAM. Segundo os autores, é possível estabelecer relacionamentos entre os diversos tópicos de um mesmo OA, entre tópicos de um OA e outro OA ou entre dois ou mais OAs, considerando seus conteúdos por completo. O primeiro tipo de relacionamento, entre tópicos de um mesmo OA, segundo Lu e Hsieh, é devidamente abordado pelo SCORM – SN. Sendo assim, apenas os dois últimos foram abordados em seu trabalho de pesquisa.

Cabe observar que o entendimento apresentado por Lu e Hsieh (2009), a respeito das relações descritas pela categoria *relation*, segundo o qual afirmam que estas são capazes de representar apenas relacionamentos estruturais entre os documentos, encontra fundamento a partir de uma possível interpretação de seus metadados. Neste caso, uma interprestação própria dos autores. Como foi apontado nas seções anteriores deste trabalho, não foi definida pela ADL uma forma específica de como os metadados desta categoria devem ser utilizados, sendo apresentada apenas uma forma geral de utilização e indicado que seja adotado o próprio vocabulário descrito em suas especificações. Sendo assim, é possível que outras interpretações permitam lhes conferir um caráter semântico sem a necessidade de alterá-los ou estendê-los.

Para definição de seu modelo, Lu e Hsieh (2009) pautaram-se em duas principais teorias: a *Instructional Design Theory* (IDT) e a *Rhetorical Structure Theory* (RST). Estas teorias forneceram a *Instructional Ontology* e a *Rhetorical-Didactic Ontology*, ambas utilizadas por eles para elaboração dos metadados necessários à sua pesquisa. Tais metadados, por sua vez, deveriam, segundo os autores, servir de extensão ao SCORM – CAM, expressar relações entre OAs, a partir de seus tópicos, além de serem facilmente implementáveis em qualquer SGA compatível com o padrão SCORM.

Após concluir seu modelo de extensão, Lu e Hsieh (2009) obtiveram quinze novas relações. Em seguida, a utilidade das mesmas, na medida em que auxiliavam os alunos em sua aprendizagem, foi testada e analisada. Os resultados de seus experimentos, segundo eles,

indicaram que as novas relações foram consideradas úteis para a maioria dos cento e quarenta e cinco alunos que contribuíram com sua pesquisa. Com base nestes resultados, os autores consideram interessante que se desenhe um conjunto comum de metadados, no sentido do trabalho por eles realizado, e que sistemas de autoria possam ser criados com suporte ao novo modelo.

Diante das premissas apontadas pelos autores, cabe levantar uma questão fundamental com relação ao padrão SCORM e um de seus objetivos primordiais, que é garantir a portabilidade e reutilização de seus pacotes de conteúdos em qualquer SGA que implemente seus modelos. Conforme exposto na documentação oficial da ADL para o SCORM – CAM, um dos motivos que levaram à construção do padrão foi o fato de que, anteriormente, cada SGA, assim como cada desenvolvedor de conteúdos, implementava e utilizava seus próprios mecanismos e esquemas para agregação, sequenciamento, navegação e suporte em tempo de execução, prejudicando o reaproveitamento e portabilidade dos OAs. Sendo assim, a elaboração de um conjunto de metadados, que necessitem de uma implementação específica no SGA, constitui-se em uma solução que vai contra uma premissa essencial do próprio padrão, pois reduz a portabilidade e a compatibilidade do pacote de conteúdo frente a outros sistemas. Todavia, como se pode observar, muitos trabalhos de pesquisa relacionados ao SCORM adotaram tal estratégia.

Em Lu e outros (2010), foi apresentado o modelo de extensão de metadados, elaborado por Lu e Hsieh (2009), sendo aplicado de forma efetiva no protótipo de um SGA por eles desenvolvido. Trinta autores de conteúdo utilizaram o sistema, atribuindo, eles próprios, as relações entre os OAs, por meio de sua interface de usuário. Em seguida, os mesmos avaliaram se houve ganho de aprendizagem com a utilização desse sistema e como pode ser aplicada mais de uma relação a um mesmo OA. Como resultado, diversas alterações foram realizadas nos arquivos XML que fazem a agregação do conteúdo, levando à inserção de diferentes elementos e atributos, criados e interpretados especificamente para o modelo de metadados e para o sistema de gerenciamento desenvolvidos, respectivamente, por Lu e Hsieh (2009) e Lu e outros (2010). Os trabalhos realizados pelos autores propuseram um novo formato de agregação com vistas às relações entre OAs, mas não deixam claro se o protótipo de SGA por eles desenvolvido implementa suporte ao restante do padrão SCORM, que consiste do SCORM-SN e do SCORM-RTE, além do próprio SCORM-CAM. Apesar de apresentarem um extenso e rico estudo acerca do estabelecimento de relações entre OAs e quais modelos de metadados são efetivamente possíveis para tal, os autores obtiveram um modelo distante do SCORM, tratado por um sistema que não suporta o padrão em si, mas sim

um conjunto específico de definições, que não encontrarão suporte em outros SGAs.

Hernández e outros (2009) adotam a seguinte definição de OA: "[...] material educativo digital, autocontido e reutilizável, que possui informações capazes de descrever seu conteúdo (metadados)." (HERNÁNDEZ et al, 2009, p.2, tradução nossa).[6] Classificando os OAs de acordo com a quantidade e variedade de conteúdos, os autores adotam uma definição segundo a qual um OA pode ser de granularidade fina ou grossa. A granularidade fina é caracterizada por um conteúdo menor, mais objetivo, como um exemplo ou um exercício. A granularidade grossa, por sua vez, é atribuída a OAs de conteúdo mais extenso, contendo outros conteúdos, como no caso de um curso completo. Para eles, OAs de granularidade fina têm maior capacidade de reutilização. Partindo deste principio, Hernández e outros desenvolveram uma ferramenta por eles denominada *Looking4LO*, que utiliza processamento de linguagem natural e ontologias para recuperação de informações em documentos, com o objetivo de extrair OAs de granularidade fina, que abordem uma determinada área do conhecimento, a partir de diferentes fontes.

Segundo Hernández e outros (2009), a ontologia fornece o modelo de domínio para a área de conhecimento, sobre a qual se deseja encontrar e extrair conteúdos a partir de um documento, enquanto a utilização de um modelo pedagógico associado permite definir que tipo de conteúdo se está buscando, ou seja, se são exercícios, exemplos, dentre outros. Dessa forma, o sistema por eles desenvolvido recebe como entrada estes dois modelos e uma fonte de documentos. A saída do sistema é um conjunto de OAs, extraído dos documentos, cujo conteúdo esteja dentro do modelo de domínio e pertença a um dos elementos definidos pelo modelo pedagógico, informação esta que é associada ao OA por meio de metadados. Para desenvolvimento do sistema, foi utilizado o *framework* denominado *General Architecture for Text Engineering* (GATE). Para os testes, foi utilizada uma ontologia de domínio para a área de "Redes de Comunicações", à qual também estavam relacionados os documentos que serviram de fonte para extração dos OAs.

Para Hernandez e outros (2009), a ontologia de domínio se apresentou como um fator crítico para o sucesso da extração dos OAs. Mesmo obtendo resultados bastante satisfatórios, segundo eles, a ontologia utilizada era bastante restrita, contendo apenas um ou dois níveis de classes e cujas instâncias permitiam realizar correspondências sobre os documentos da amostragem empregada nos testes. Segundo os autores, as anotações de metadados baseados na ontologia não utilizaram informações contidas em suas relações entre as entidades, não

[6] [...] material educativo digital, auto-contenido y re-utilizable, poseedor de información que permite describir su contenido (metadata).

empregando parte relevante do potencial oferecido pelas ontologias de domínio. Para Hernandez e outros, a identificação das relações semânticas entre os conceitos presentes na ontologia poderiam melhorar consideravelmente a precisão das buscas realizadas pelo sistema. Além disso, a identificação destas relações poderia ser utilizada para melhorar a delimitação dos OAs.

O trabalho desenvolvido por Hernández e outros (2009) resulta em um sistema capaz de auxiliar desenvolvedores de conteúdos na criação de OAs de granularidade fina. Ele é capaz de extraí-los para diferentes áreas do conhecimento, desde que se forneça como entrada diferentes ontologias de domínio. Além disso, utiliza recursos para automatização dos processos a partir do emprego de técnicas de recuperação de informação e anotação automática de metadados. Todavia, a variedade de OAs extraídos pode ser grande ou pequena, dependendo da variedade, tamanho e quantidade dos documentos presentes na fonte de conteúdos fornecida como entrada para o sistema, o que pode resultar em redundância destes OAs, ocasionando, ao final do processo, um baixo aproveitamento de seu conjunto. Além disso, a verificação e seleção dos OAs gerados na saída fica a cargo do usuário do sistema que, por ventura, deseje utilizá-los na composição de um OA de granularidade grossa. Ainda neste sentido, o *Looking4LO* não oferece recursos que auxiliem o desenvolvedor de conteúdos para *eLearning* na composição de um conteúdo mais extenso e complexo, como no caso de um curso ou disciplina, que demande a identificação e seleção de OAs relacionados entre si, capazes de oferecer conteúdos complementares e, por vezes sequenciáveis, compondo, em seu conjunto, o conteúdo final.

Maratea, Petrozino e Manzo (2012) procuraram realizar a extração automática dos metadados definidos na categoria *general,* contemplada pelo SCORM – CAM, para classificar OAs compostos de artigos científicos. Como alguns destes metadados, segundo eles, estão estreitamente relacionados à estrutura e seções do documento, como no caso do *title e description,* e outros são avaliados a partir de seu próprio conteúdo, tais como *language* e *coverage,* diferentes técnicas foram implementadas para cada tipo de metadado. Para extração dos metadados considerando-se informações estruturais, foi aplicado um passo de preprocessamento, sobre cada arquivo em PDF, obtendo-se, para cada um, um arquivo XML que separa e estrutura cada seção do documento. Este arquivo resultante foi então submetido a uma estratégia de análise baseada em regras para extração dos devidos metadados. No caso dos metadados a serem extraídos a partir do texto do artigo, foi utilizado o *Vector Space Model,* como estratégia para o processamento de linguagem natural. Para isso, foram removidas todas as informações de formatação dos documentos.

A partir dos testes aplicados sobre um conjunto de dezessete artigos científicos, Maratea, Petrozino e Manzo (2012) constataram que as técnicas por eles propostas permitiram a correta extração dos metadados, com um bom nível de precisão. Diante disso, propuseram, como trabalhos futuros, a extração de metadados mais complexos e em documentos menos estruturados do que aqueles utilizados por eles. A extração automática de metadados, com base na categoria *relation do* SCORM – CAM, sobre conteúdos didático-pedagógicos, pode se configurar como um caso de estudo que converge com a indicação de trabalhos futuros feita pelos autores.

Huynh e Hoang (2010), por sua vez, buscaram relacionar artigos científicos com base nos metadados extraídos a partir de documentos em PDF disponíveis na Web. Com base nos metadados obtidos, segundo os autores, é possível reconhecer e saber em quais documentos um dado artigo é referenciado. Para isso, desenvolveram um sistema que utiliza informações de layout dos documentos, regras construídas com base em modelos e uma ontologia, por eles construída, para artigos relacionados à computação. Como referência para os metadados a serem extraídos, os autores adotaram o *Dublin Core Metadata.*

Huynh e Hoang (2010) utilizaram, em seu sistema, componentes e APIs do *framework* GATE. O componente denominado *Apache PDFBox library* possibilita processar o *layout* e o estilo dos documentos para inserir anotações com base na fonte, posição dos textos e palavras-chave. Modelos podem ser extraídos utilizando-se o componente *JAPE Grammar* e um dicionário de nomes de entidades, ambos presentes no GATE, por meio dos quais é possível criar regras para extração de metadados. Para identificação de entidades presentes no conteúdo dos documentos, tais como *Location*, *Person*, *Organization* e *Date-Time*, foi utilizado um *plugin* do *framework*, denominado ANNIE, que além de implementar uma série de algoritmos para extração de informação, disponibiliza também um modelo de dicionário e de regras já definidas, todos podendo ser estendidos.

Uma vez extraídos os metadados, o sistema desenvolvido por Huynh e Hoang (2010) possibilita ao usuário a correção e validação dos mesmos, antes de serem finalmente exportados para um arquivo em XML, permitindo que sejam utilizados para organizar os documentos em bibliotecas digitais ou para enriquecer a ontologia de domínio até então utilizada. Os autores apontam para o fato de que, na abordagem por eles proposta, é necessário muito cuidado ao se criar regras e modelos, e que o levantamento de diversos modelos consiste em uma tarefa trabalhosa, que exige tempo e conhecimento de domínio. Sendo assim, propõem como trabalho futuro combinar sua metodologia atual com a utilização de algoritmos de aprendizagem de máquina, no intuito de aumentar sua acurácia e extrair

novos grupos de metadados com base no *Dublin Core*. Ainda neste sentido, os autores afirmam que a criação de regras e modelos para extração de metadados em referências bibliográficas, juntamente com a identificação da relação entre elas, pode auxiliar o usuário a identificar documentos que se referenciam, assim como verificar se uma dada referência é válida.

Após uma breve apresentação dos principais passos executados por seu algoritmo para extração de metadados, assim como a exibição de alguns exemplos de regras por eles definidas, Huynh e Hoang (2010) não deixam claro em seu artigo o modo como os metadados obtidos podem ser utilizados para organização dos documentos em bibliotecas digitais. O mesmo se pode afirmar com relação à utilização dos metadados para identificação de relações entre diferentes artigos.

Uma estratégia muito semelhante à de Huynh e Hoang (2010) foi utilizada por Guo e Jin (2011b), ao desenvolverem o sistema denominado *SemreX*, a partir do *framework* para extração de metadados discutido em Guo e Jin (2011a). Trata-se de um sistema *peer-to-peer* (P2P) para compartilhamento de documentos de textos entre pesquisadores em ciência da computação. Seu sistema implementa um *framework* baseado em regras para extração de metadados relacionados ao título, autores, resumo, periódicos, volume, ano e página presentes nas citações e referências bibliográficas de artigos científicos. Os arquivos em PDF, dos quais são extraídos os metadados, são convertidos pelo sistema em dois formatos diferentes: um arquivo texto simples e um XML. O arquivo texto contém todo o texto do arquivo fonte, porém sem informações de formatação. O XML, por sua vez, utiliza referências espaciais do documento origem para referenciar os blocos de textos e então, para cada um deles, armazenar os dados de formatação. Para os autores, as informações de formatação auxiliam na identificação do tipo de conteúdo, além de auxiliar na extração dos metadados, tornando o processo mais preciso. A partir daí, os autores aplicam algoritmos baseados em regras, com a utilização de bases de conhecimento, para extração dos metadados e subsequente atualização da base de conhecimento utilizada.

As abordagens propostas por Huynh e Hoang (2010), Guo e Jin (2011a), Guo e Jin (2011b) consistem em maneiras eficientes de extração de metadados em artigos científicos. Todavia, exploram os aspectos estruturais dos documentos como uma referência primordial para a estratégia adotada. Tais abordagens podem não ser tão eficientes quando se tem um conjunto de documentos heterogênios, como no caso de conteúdos didático-pedagógicos que, ao comporem Objetos de Aprendizagem, não possuem, necessariamente, uma estrutura padronizada para apresentação de seu conteúdo, como ocorre com os artigos científicos.

Segundo Tuarob e Pouchard (2013), o *DataOne* consiste em uma rede de dados construída para facilitar o acesso a dados sobre ciências ambientais e ecológicas em todo o mundo. Estes dados são obtidos de diferentes provedores e disponibilizados por meio de uma interface de pesquisa denominada *ONEMercury*. Porém, o conjunto de palavras-chave utilizado pelos usuários nas buscas é predefinido, podendo ser alterado apenas pelos administradores do sistema no intuito de se evitar o aparecimento de palavras-chave inválidas, pois este conjunto é utilizado para anotação manual durante o processo de levantamento de dados que, como visto, são provenientes de fontes diversas. Assim, o problema se encontra no fato de que, dessa forma, é necessário lidar com diferentes níveis de anotação nos dados obtidos das diferentes fontes, sendo que muitos deles podem conter anotações sem significado para o *ONEMercury*, fazendo com que os dados sejam perdidos durante as buscas.

Sendo assim, Tuarob e Pouchard (2013) apresentam algoritmos, por eles desenvolvidos, para anotação automática de metadados. A estratégia utilizada pelos autores consiste, na verdade, em transformar o problema de anotação em um problema de recomendação de *tags* com base em uma biblioteca de palavras-chave, como aquela compreendida pelo *ONEMercury*. Em síntese, os metadados mal registrados nos arquivos analisados, com relação ao conjunto utilizado por esta interface de pesquisa do *DataOne*, são novamente anotados com metadados similares, resultando em uma nova anotação e diminuindo as chances de não serem considerados nas pesquisas do usuário. Como pode ser observado, o problema apresentado se deve principalmente à diversidade das fontes de dados empregadas pelo *DataOne*, quando estes precisam ser pesquisados a partir de uma única interface que utiliza uma biblioteca própria de palavras-chave.

Técnicas de recuperação de informação foram empregadas pelos autores nos algoritmos por eles propostos. Todavia, estes demandam, conforme apontado pelos autores, grande tempo de treinamento sobre a biblioteca de palavras-chave, que pode ser ampliada e modificada a qualquer momento pelos administradores do sistema, tornando necessário novo treinamento. Além disso, o processo de recomendação de *tags* ainda precisa ser melhor avaliado, conforme indicado pelos autores, com relação a sua eficiência e escalabilidade. Talvez o emprego de uma ontologia de domínio, que fosse difundida para as diferentes fontes de dados utilizadas pelo *DataOne*, pudesse facilitar a padronização dos dados utilizados e diminuir a necessidade de aplicação destes algoritmos em larga escala, podendo-se associar uma fase anterior de anotação automática com base nesta ontologia, para depois empregar os algoritmos propostos pelos autores.

Para auxiliar nos processos inerentes à Recuperação da Informação, tais como a

geração e anotação automáticas de metadados e processamento de linguagem natural, há diversas ferramentas e *frameworks* disponíveis. Lipinski e outros (2013) apresentaram uma avaliação de diferentes abordagens e ferramentas para extração de metadados nos cabeçalhos de artigos científicos. Nauerz e outros (2008) propuseram um *framework* utilizando os serviços de análise de dados não estruturados, denominados UIMA e Calais. Engelhardt e outros (2006) utilizaram o *framework* JENA como parte de sua solução para análise de OAs multimídia. Maynard (2008) apresentou um *benchmarking* de ferramentas de anotação automática de texto, concluindo que o GATE obteve a melhor avaliação geral.

Como se pode observar, alguns trabalhos apresentam abordagens e metodologias bastante consistentes e viáveis para anotação automática de OAs, seja no intuito de oferecer melhores mecanismos para pesquisa em repositório de conteúdos, como em Roy, Sudeshna e Sujoy (2008), que utilizam a categoria *educational* para classificação de seus conteúdos, ou para criar uma rede semântica sobre o repositório, permitindo o estabelecimento de relações entre OAs, como é feito em Engelhardt e outros (2006). Todavia, ambos utilizam extensões ao modelo de metadados do SCORM e SGAs específicos, o que reduz a portabilidade e compatibilidade do pacote de conteúdo frente a outros sistemas. No caso de Engelhardt e outros, ao se pensar na construção de uma unidade de aprendizagem, a solução proposta pode não ser adequada, pois leva a um acesso disperso a um conjunto de variados OAs, haja vista que os documentos não são recomendados, mas sim pesquisados.

Em Maratea e outros (2012), por sua vez, foi realizada a extração automática dos metadados definidos na categoria *general*, do SCORM, para classificar OAs compostos de artigos científicos, utilizando informações estruturais dos documentos. Propuseram, como trabalhos futuros, a extração de metadados mais complexos, em documentos menos estruturados. Neste sentido, os metadados da categoria *relation*, extraídos de documentos heterogêneos, tais como OAs, são um bom exemplo. As abordagens propostas por Guo e Jin (2011) e Tuarob e Pouchard (2013) também consistem em maneiras eficientes de extração de metadados em artigos científicos, a partir de aspectos estruturais dos documentos.

Por outro lado, há trabalhos cuja proposta pode restringir seu escopo de utilização, reduzindo muito a portabilidade e a compatibilidade dos OAs com diferentes SGAs. É o caso de Edvardsen e outros (2009), cuja abordagem depende de um *framework* desenvolvido para utilizar-se de informações externas aos OAs, fazendo com que o mesmo conjunto de metadados possa não ser obtido em outro contexto, além de demandar um SGA específico. No caso de Lu e Hisieh (2009) foi desenvolvido um modelo próprio de extensão aos metadados da categoria *relation*. Este modelo foi implementado, em Lu e outros (2010), no

protótipo de um SGA desenvolvido por eles, por meio do qual as relações entre os OAs são atribuídas manualmente. Como resultado, diversas alterações foram realizadas nos arquivos XML que fazem a agregação do conteúdo, acabando por obter um modelo distante do SCORM e que é tratado por um sistema que não suporta o padrão em si, mas sim um conjunto específico de definições, que não encontrarão suporte em outros SGAs.

Dentre os estudos levantados, os que mais de perto se relacionam a esta pesquisa são aqueles cujas abordagens foram pautadas no padrão SCORM e que se dedicaram a extrair metadados em OAs com base em seu Modelo de Agregação de Conteúdos. Dentre estes, alguns buscaram, além de extrair os metadados, estabelecer uma relação entre os OAs com base nos metadados definidos pela categoria *relation*, porém, acabaram por propor extensões aos metadados, além de SGAs específicos para que estes pudessem ser aplicados.

4 METODOLOGIA DE PESQUISA

De acordo com Gil (2002), a partir dos objetivos de uma pesquisa é possível classificá-la em relação a três grandes grupos: pesquisas exploratórias, pesquisas descritivas e pesquisas explicativas.

Para Gil (2002), uma pesquisa exploratória "tem como objetivo principal o aprimoramento de ideias ou descoberta de intuições", possibilitando maior familiaridade com o problema apresentado. Segundo o autor, seu planejamento é flexível e pode incluir levantamento bibliográfico, entrevistas com pessoas que tenham experiência com relação ao problema e análise de exemplos. Porém, na maioria dos casos, uma pesquisa de natureza exploratória assume a forma de uma pesquisa bibliográfica ou de um estudo de caso.

Uma pesquisa descritiva, segundo Gil (2002), tem como objetivo a descrição de um dado fenômeno ou de uma população, assim como o estabelecimento de relações entre variáveis. Segundo o autor, uma das principais características que definem uma pesquisa descritiva consiste na utilização sistemática de técnicas padronizadas para coletas de dados, tais como o questionário e a observação. Segundo ele, há pesquisas descritivas que, além de identificar relações entre variáveis, procuram estabelecer a natureza destas relações, o que aproxima estas pesquisas de uma pesquisa explicativa. Há ainda aquelas pesquisas que se caracterizam como descritivas, mas que acabem possibilitando uma nova visão do problema, o que as aproxima de uma pesquisa exploratória.

As pesquisas explicativas são definidas por Gil (2002) como sendo aquelas que têm como objetivo "identificar os fatores que determinam ou que contribuem para a ocorrência dos fenômenos." Para ele, este é o tipo de pesquisa que permite explicar a razão e os motivos das coisas, o que torna a pesquisa complexa e mais passível de erros.

Essa classificação apresenta por Gil (2002) é, segundo ele, importante para o estabelecimento do marco teórico que possibilita conceituar uma dada pesquisa. Todavia, ainda assim é necessário, segundo o autor, traçar um modelo, um desenho operativo da pesquisa denominado por ele como "delineamento", para que seja possível a aplicação de uma análise empírica que permita confrontar a visão teórica com os dados da realidade. Este conceito de delineamento de uma pesquisa converge com os conceitos de "método", apresentados por Wazlawick (2009) e por Cervo e Bervian (2002). Para Wazlawick (2009), há uma distinção entre "metodologia" e "método", onde a metodologia seria o estudo dos métodos, enquanto o método em si constitui-se nos procedimentos adotados em uma pesquisa. Segundo ele, diante de um dado objetivo, o método apresenta os passos a serem executados

para demonstrar que o objetivo proposto foi atingido. Uma caracterização semelhante do método é apresentada por Cervo e Bervian (2002), para os quais o método consiste em um conjunto de processos ordenados de forma que possam ser empregados na investigação e na demonstração da verdade.

Sendo assim, para Gil (2002) as pesquisas podem ser classificadas, então, com relação a um marco teórico, como discutido anteriormente, e com relação ao seu delineamento ou método, ou seja, os procedimentos técnicos utilizados. Neste sentido, diferentes classificações são apresentadas pelo autor, estando entre elas as "pesquisas bibliográficas" e as "pesquisas experimentais".

Conforme Gil (2002), as pesquisas bibliográficas definem a maioria dos estudos exploratórios e são desenvolvidas com base em material anteriormente elaborado na forma de livros e artigos científicos. Permitem ao pesquisador ampliar a descoberta de conhecimentos relacionados ao problema a ser investigado, conhecimentos estes que não poderiam ser obtidos diretamente de sua própria pesquisa. Segundo o autor, a pesquisa bibliográfica pode ser entendida como um processo que contempla as seguintes etapas, realizadas na respectiva ordem: escolha do tema, levantamento bibliográfico preliminar, formulação do problema, elaboração do plano provisório de assunto, busca das fontes, leitura do material, fichamento, organização lógica do assunto e redação do texto.

Uma pesquisa experimental, por sua vez, segundo Gil (2002), é caracterizada principalmente pela definição de um objeto de estudo, seleção das variáveis capazes de influenciá-lo e a definição das formas de controle e de observação dos efeitos das variáveis sobre o objeto. Para o autor, em uma pesquisa experimental deve haver a manipulação por parte do pesquisador sobre os elementos estudados. Também é necessário que o pesquisador introduza mecanismos de controle nas situações experimentais, principalmente por meio de um grupo de controle. Outro fator importante é a aleatoriedade na designação dos elementos que irão compor os grupos experimental e de controle. Ainda segundo o autor, o planejamento de uma pesquisa experimental inclui o desenvolvimento dos seguintes passos, realizados na respectiva ordem: formulação do problema, construção das hipóteses, operacionalização das variáveis, definição do plano experimental, determinação dos sujeitos, determinação do ambiente, coleta de dados, análise e interpretação dos dados e apresentação das conclusões.

Para realização do presente trabalho de pesquisa, a partir do problema apresentado como foco de investigação e com base na literatura, uma possível solução foi proposta e experimentos foram conduzidos para verificação de sua eficácia, com coleta de dados, análise e interpretação destes dados e apresentação das devidas conclusões. Dessa forma, com base

nas definições acerca dos métodos de pesquisa, apresentadas em Gil (2002), Wazlawick (2009) e Cervo e Bervian (2002), pode-se afirmar que esta pesquisa teve, inicialmente, um caráter exploratório, com foco na pesquisa bibliográfica, seguido de uma pesquisa experimental, onde foi formulado o problema, construída a hipótese e definidos os objetivos, e cujos procedimentos metodológicos foram organizados nas seguintes etapas:

1) definição de uma metodologia para recomendação de OAs relacionados;
2) pesquisa e seleção de APIs e *frameworks* que fornecessem os recursos necessários para implementação da metodologia anteriormente definida;
3) organização e montagem de um repositório de OAs;
4) implementação do protótipo de um Sistema de Recomendação e Agregação de Conteúdos Relacionados;
5) realização de testes;
6) análise quantitativa e qualitativa dos resultados alcançados.

As próximas seções definirão em detalhes cada uma destas etapas.

4.1 Definição de uma metodologia para recomendação de OAs relacionados

Esta etapa consistiu do estudo e proposição de um modelo conceitual para os principais processos necessários à recomendação automática de OAs relacionados, em conformidade com o SCORM.

Também foram contempladas as definições das estratégias adotadas, assim como a elaboração dos algoritmos implementados. Recorreu-se à revisão literatura para verificação dos modelos e estratégias já utilizados em outros trabalhos, procurando-se identificar as que foram mais bem sucedidas, segundo os autores, em relação aos processos conhecidamente inerentes à anotação automática de metadados, recuperação da informação e mineração de textos. Isso permitiu que a maior parte dos esforços, tanto no projeto quanto na experimentação da metodologia aqui proposta, fossem destinados às peculiaridades e problemas específicos deste trabalho de pesquisa.

4.2 Pesquisa e seleção de APIs e *frameworks*

Esta etapa teve como objetivo a seleção de ferramentas (APIs e *frameworks*) que oferecessem os recursos necessários para que se pudesse implementar as estratégias definidas na etapa anterior, contemplando a metodologia proposta para recomendação de OAs relacionados. Foram observadas as ferramentas adotadas por outros pesquisadores, a partir da revisão de literatura, e a forma como foram utilizadas em seus trabalhos. Também foram consideradas as avaliações e comparações de diferentes ferramentas, presentes nos trabalhos disponíveis na literatura.

No intuito de se garantir a disponibilidade dos recursos necessários, assim como o acesso ao projeto do sistema e a todos os insumos utilizados, em qualquer parte e a qualquer momento, os seguintes critérios foram importantes para a definição da ferramenta a ser finalmente utilizada, dentre as que foram selecionadas: permitir o desenvolvimento para *desktop*, possibilitar o acesso a repositórios de conteúdos instalados localmente e ser independente de plataforma, diante do que o Java foi desde o início apontado como linguagem de programação preferencial. Além disso, deveria oferecer recursos facilmente reutilizáveis e que atendessem às necessidades de implementação previstas pela metodologia.

4.3 Organização e montagem de um repositório de OAs

Para realização deste trabalho, foram pesquisados e selecionados OAs cujos conteúdos consistissem em textos didático-pedagógicos, relacionados a uma área de conhecimento específica, escritos em inglês e livremente disponibilizados na Web, podendo ser acessados gratuitamente. A escolha do inglês como idioma se deve ao fato de que a maioria das ferramentas, APIs e *frameworks* disponíveis o têm como idioma padrão, evitando-se assim a necessidade de se realizar configurações específicas para o português no ambiente de desenvolvimento, que não era o foco deste trabalho.

Dessa forma foi necessário, em primeiro lugar, definir a área de conhecimento à qual os conteúdos dos OAs deveriam estar relacionados, delimitando-se sua área de domínio. Foram, então, pesquisados repositórios de OAs para as áreas das ciências exatas, biomédicas, humanas, letras e artes. Porém, no intuito de ampliar as possibilidades de exploração e experimentação da metodologia proposta, foram adotados os seguintes critérios: deveriam ser priorizadas as áreas para as quais não fossem encontradas muitas contribuições no sentido de sua formalização ontológica, categorização de seus conceitos mais relevantes e para as quais

fosse possível encontrar um grande volume de documentos, cujos textos fossem variados, heterogêneos e cujos conteúdos não consistissem em textos demasiadamente técnicos. A questão principal, neste sentido, foi que os documentos tivessem caráter prioritariamente didático-pedagógico, tratando de assuntos específicos e cuja leitura fosse acessível a pessoas que não dominam o conteúdo. Da mesma forma, a ontologia de domínio, caso houvesse, precisava apresentar a classificação do conhecimento contido nestes documentos e não na categorização dos documentos em si.

Foram encontrados diversos repositórios, assim como ontologias e dicionários de termos, mas a área que melhor atendeu aos critérios, dentre o material pesquisado, foi a da Música Erudita. Para esta área do conhecimento, foram encontrados diversos documentos na Web, assim como bibliotecas digitais, mas pouco material que apresentasse uma formalização ontológica de sua área de domínio.

O site NAXOS[7] possui uma área destinada à educação musical, com ênfase na história da música erudita, biografia e obras de compositores, óperas e introdução à teoria musical. O acesso ao seu conteúdo é livre, seus textos são acessíveis e de fácil leitura e compreensão. Dele foi obtida a maioria dos documentos que compõem o repositório de conteúdos utilizado nesta pesquisa.

Os documentos foram armazenados localmente, em uma pasta no computador utilizado para realização da pesquisa, estando disponíveis para serem recuperados e utilizados sem a dependência de uma conexão com a internet.

4.4 Implementação de um Sistema de Recomendação e Agregação de Conteúdos Relacionados

Nesta etapa foi projetado e implementado o protótipo de um sistema que contemplasse a metodologia para recomendação de OAs relacionados, elaborada como parte desta pesquisa, em caráter exploratório, no intuito de se verificar sua eficácia diante do objetivo proposto, analisar os resultados obtidos em cada uma das etapas de processamento previstas pela metodologia e identificar possíveis correções e melhorias a serem realizadas.

4.5 Realização de testes

O objetivo desta etapa foi verificar se a metodologia proposta é viável e se contribui para solução do problema anteriormente identificado. Foram convidados a participar dos

[7] The World's Leading Classical Music Group

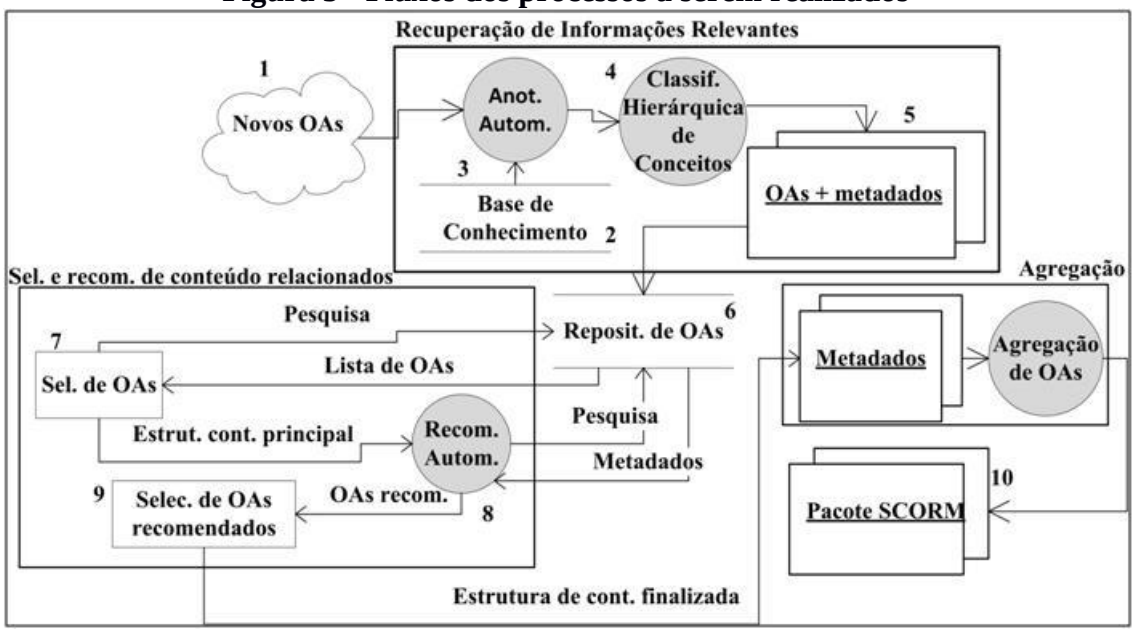
testes, profissionais e especialistas da área de *e-Learning*, no intuito de fornecer elementos de análise capazes de pautar a posterior análise dos resultados obtidos a partir da execução do sistema sobre o repositório de conteúdos.

Para verificação e avaliação dos resultados obtidos, foram desenvolvidos dois casos de testes. O primeiro teve como objetivo principal a verificação da corretude e acurácia alcançadas na geração e anotação automáticas dos metadados e na identificação e classificação de termos chave e conceitos relevantes aos conteúdos dos OAs. O segundo caso de teste, por sua vez, teve como objetivo avaliar a eficácia dos processos de recomendação automática e agregação de conteúdos relacionados. Para realização destes testes, uma amostragem de OAs foi colhida, dentre aqueles que constituem o repositório de conteúdos organizado para realização desta pesquisa.

5 PROPOSIÇÃO DE UMA METODOLOGIA PARA RECOMENDAÇÃO DE OAs RELACIONADOS

O diagrama da Figura 5 apresenta uma visão geral da Metodologia para Recomendação e Agregação de OAs Relacionados, em conformidade com o SCORM, proposta nesta pesquisa. São contempladas três etapas, brevemente descritas a seguir e detalhadas nas próximas seções.

Figura 5 - Fluxos dos processos a serem realizados



Fonte: Elaborada pelo autor

A primeira etapa consiste da recuperação de informações relevantes a cada um dos OAs. Sendo assim, a partir de uma base de conhecimento, representada por (2) na Figura 5, um conjunto de OAs (1) é submetido a um processo de anotação automática de metadados (3), que identifica e classifica seus termos chave e conceitos relevantes. Em seguida, é realizada uma classificação hierárquica destes termos e conceitos quanto ao seu grau de relevância (4). Os OAs, devidamente anotados (5), são, então, armazenados em um repositório (6), a partir do qual podem ser selecionados (7) e utilizados para composição de um dado conteúdo. Este conteúdo é submetido a um processo de recomendação de conteúdos relacionados (8), a partir dos documentos presentes no repositório. Neste processo, outros OAs são pesquisados, com o objetivo de serem agregados como conteúdos relacionados e, ao final, os documentos recomendados podem ser mantidos ou excluídos manualmente (9). Unindo-se documentos pré-selecionados e documentos recomendados, um pacote de conteúdo no formato SCORM é gerado de acordo com as especificações do SCORM-CAM (10).

5.1 Estratégia para recuperação de informações relevantes ao conteúdo de um OA

Para que se possa recomendar e agregar OAs relacionados é necessário que estes sejam submetidos a uma etapa de pré-processamento, que tem como objetivo recuperar informações relevantes aos seus conteúdos, cuja análise permita identificar possíveis relações entre eles.

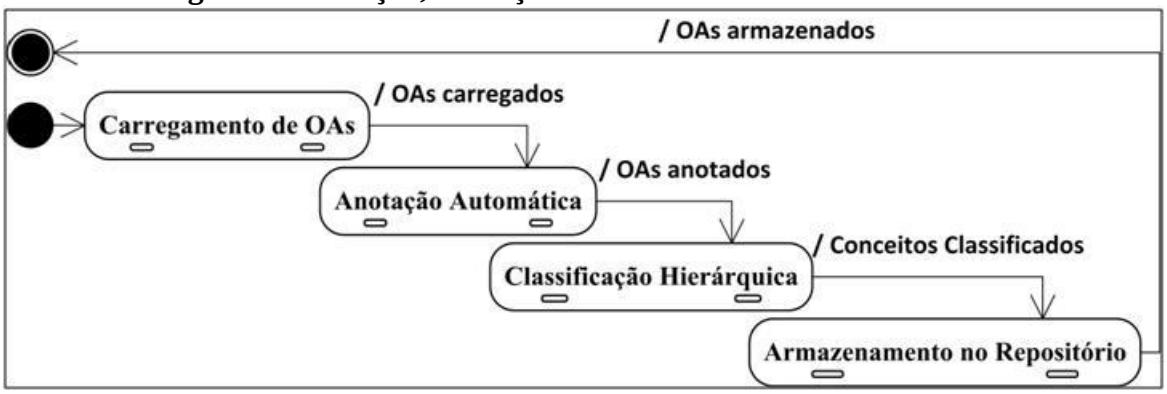
Ao se pensar na construção de um determinado curso ou disciplina, os conteúdos dos OAs utilizados, por se tratarem de materiais didático-pedagógicos, estão associados a uma determinada área do conhecimento. Sendo assim, é necessário recuperar informações, a partir de seu conteúdo, que sejam capazes de representá-lo como um todo, sintetizando os principais assuntos nele abordados, e que sejam relevantes à área de conhecimento à qual esteja relacionado. Dessa maneira, em um texto que trate da biografia de um importante compositor da música erudita, por exemplo, não é interessante que se identifique quaisquer nomes de pessoas ou lugares que nele ocorram, mas sim aqueles que estejam relacionados às áreas de conhecimento compreendidas pelo domínio da música erudita, de modo que possam, então, ser posteriormente analisados quanto ao grau de sua relevância para o conteúdo do documento em si. Estes elementos irão compor o conjunto de termos e conceitos mais relevantes ao documento e que o caracterizam quanto ao conteúdo nele presente.

Neste contexto, é necessário que se tenha, como referência primordial, um modelo de domínio capaz de caracterizar e representar a área de conhecimento à qual pertencem os OAs, sobre os quais se deseje aplicar estratégias para recuperação da informação que, neste caso, torna-se um processo de recuperação de informações relevantes. Sendo assim, conforme discutido na literatura, o emprego de uma ontologia de domínio é fundamental.

Diante disso, foi definida neste trabalho de pesquisa uma estratégia, para recuperação de informação, que utiliza uma base de conhecimento de domínio, composta de uma ontologia de domínio e de um dicionário de termos que a contemplam. Esta estratégia é composta de duas etapas principais. A primeira etapa consiste na geração e anotação automática de metadados para identificação de termos chave e conceitos relevantes, a partir da base de conhecimento de domínio. A segunda consiste na análise dos metadados anteriormente anotados, visando a classificação hierárquica dos elementos por eles identificados, quanto à sua relevância em relação ao conteúdo como um todo. Ao final, obtém-se uma lista de termos e conceitos, ordenados por seu grau de relevância. Ambas as etapas são realizadas durante a inserção de novos OAs no repositório de conteúdos, o que compreende quatro processos

principais, de acordo com a Figura 6. Os dois primeiros consistem do carregamento de novos OAs e da anotação automática de cada um deles, contemplando a primeira etapa. Em seguida, como parte da segunda etapa, que utiliza a saída gerada pela etapa anterior, ocorrem os processos de classificação hierárquica dos elementos anotados e de armazenamento dos OAs, juntamente com seus metadados, no repositório de conteúdos. Cada um destes processos será detalhado a seguir.

Figura 6 - Inserção, anotação e armazenamento de novos OAs

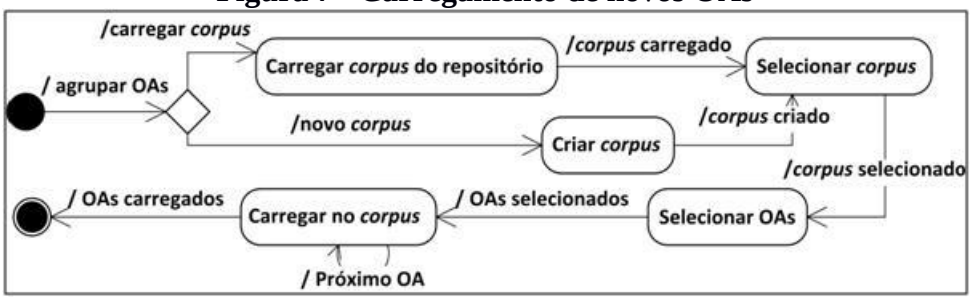


Fonte: Elaborada pelo autor

5.1.1 Carrregamento de novos OAs

Para que sejam carregados, os novos OAs precisam ser inseridos em um *corpus* de documentos, que recebe um nome, permitindo agrupá-los e identificá-los dentro do repositório e possibilitando sua fácil recuperação para composição de um dado conteúdo. Sendo assim, um *corpus* existente pode ser carregado a partir do repositório, ou um novo pode ser criado. Este processo é ilustrado na Figura 7.

Figura 7 - Carregamento de novos OAs

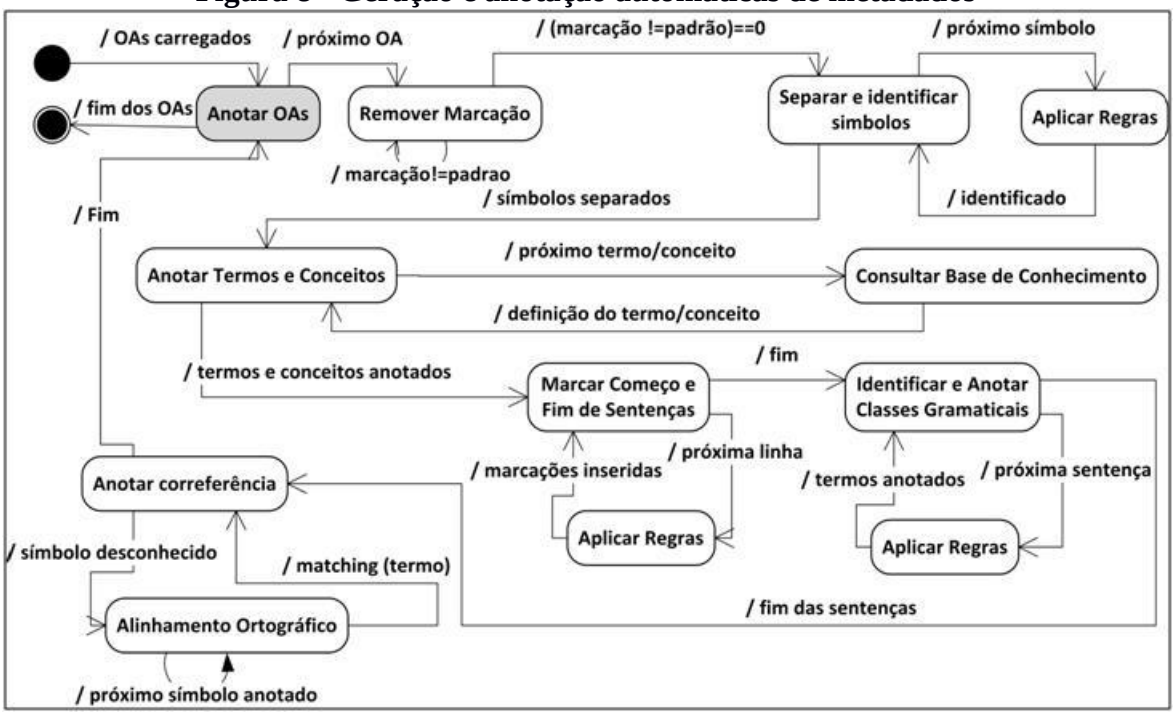


Fonte: Elaborada pelo autor

5.1.2 Geração e anotação automática de metadados

Após o carregamento dos novos OAs em seu devido *corpus*, o processo seguinte consiste na geração e anotação automática dos metadados em cada um dos documentos, conforme ilustrado na Figura 8 e descrito a seguir.

Figura 8 - Geração e anotação automáticas de metadados



Fonte: Elaborada pelo autor

No intuito de otimizar a identificação de símbolos, termos e conceitos, é importante que apenas os dados textuais inerentes ao conteúdo do documento e aqueles relevantes à sua identificação, tais como autores, título e palavras-chave estejam presentes. Dessa forma, o primeiro processo a ser executado consiste na remoção das marcações que não sejam marcações padrão de formatação e estruturação do texto, além de marcações e anotações inseridas no documento por *softwares* de autoria e edição, ou quaisquer outras que não pertençam a um conjunto de entrada padrão. Dessa forma, no caso de OAs em formato HTML, por exemplo, seriam mantidos o texto e os elementos próprios da linguagem de marcação, sendo removidos quaisquer outros dados não pertencentes a este conjunto, evitando que sejam processados nas etapas seguintes e possam interferir nos resultados desejados, além de aumentar desnecessariamente o tempo de processamento de cada documento.

Após a remoção das marcações é necessário, ainda, que os diferentes símbolos presentes no documento sejam separados e identificados como sendo números, símbolos de

pontuação ou palavras. Trata-se de uma fase de pré-processamento importante para os demais processos, pois além dos metadados relacionados à base conhecimento de domínio, outros grupos precisam ser também anotados nas etapas seguintes, com relação à classe gramatical de um termo dentro de uma sentença, ou mesmo para reconhecimento de cada sentença, por exemplo. Estes processos utilizam-se da aplicação de diversos conjuntos de regras, *lexicons* e dados da base de conhecimento para processar cada símbolo presente no documento, o que demonstra o quanto uma correta separação e identificação dos mesmos afeta a eficiência e acurácia do processo de geração e anotação de metadados como um todo.

Uma vez que os diversos símbolos presentes no documento tenham sido identificados, tem início o processo de geração e anotação automática de metadados para identificação dos termos e conceitos relevantes ao conteúdo de cada OA. Todos os termos e conceitos, presentes no documento e que se encontram definidos na base de conhecimento de domínio, devem ser devidamente anotados com seu tipo, posição taxonômica e classificação em relação à ontologia. Além disso, também é necessário que estes recebam anotações relativas à sua classe gramatical, haja vistas que substantivos e nomes próprios, por exemplo, podem apresentar maior potencial de relevância que outros, tais como adjetivos e advérbios. Dessa forma, do início ao fim do documento, cada linha precisa ser analisada, aplicando-se um conjunto de regras no intuito de identificar e separar cada uma de suas sentenças, para que a classe gramatical de um termo ou conceito possa ser identificada por meio de uma análise de sua posição sintática dentro da sentença à qual pertença. Assim, após a subdivisão do texto em sentenças bem definidas, seus termos e conceitos podem receber, então, anotações com metadados que identifiquem sua classe gramatical. Aqueles que não forem devidamente identificados são anotados como “desconhecido”. Todos estes metadados, anotados até então, são de extrema importância para que se possa, futuramente, submeter um ou mais OAs ao processo de recomendação automática de conteúdos, conforme será definido nas próximas seções.

Ao final do processo, é necessário buscar correferências para os símbolos que não tenham sido identificados, mas que podem estar relacionados a termos e conceitos importantes para o conteúdo do OA. Para isso, algoritmos de correspondência, ou *matching*, devem ser aplicados para que cada termo anotado como “desconhecido” possa receber a mesma anotação dada a um termo correspondente, previamente anotado.

5.1.3 Classificação hierárquica de termos chave e conceitos relevantes

Uma vez que um *corpus*, presente no repositório de conteúdos, tenha passado pelo processo de anotação automática de metadados, onde os termos chave e conceitos relevantes foram devidamente identificados e anotados, é necessário determinar quais dentre estes elementos são os mais representativos com relação ao conteúdo didático-pedagógico contido em cada OA. Para isso, cada um deles precisa ser analisado, no intuito de atribuir-lhes um valor que permita ponderar sua relevância para o conteúdo do documento e de acordo com as definições da base de conhecimento de domínio.

As técnicas utilizadas nesta fase, para o cálculo de relevância de cada termo anotado, foram definidas com base na literatura relativa à recuperação da informação e mineração de textos. Conforme descrito na literatura, para a recuperação de informação há algumas técnicas que consideram a relação dos termos e conceitos de um dado texto com todos os textos presentes em uma coleção de documentos. Estas não foram utilizadas, pois se destinam principalmente ao cálculo de similaridade entre documentos, ou entre estes e os dados de entrada de uma consulta, fornecidos por um usuário (MORAIS; AMBRÓSIO, 2007), sendo que o objetivo deste trabalho é estabelecer relações entre diferentes documentos, buscando não a similaridade, mas completude entre seus conteúdos. Diante disso, foram empregadas técnicas que permitem a análise dos termos e conceitos em relação ao próprio documento onde se inserem.

O objetivo é associar, aos metadados anteriormente anotados, métricas acerca da frequência do termo no texto do documento e de sua potencial relevância, com base em usa posição na estrutura do documento, na sentença em que se encontra e em sua classificação gramatical. A cada métrica são associados os devidos pesos sobre os quais é calculada a relevância final do termo em análise, que passa, também, a estar a ele associada como um novo metadado. Os valores de relevância final são utilizados para geração de um vetor de relevância, que contém os termos melhor ponderados, ordenados pelo maior peso, sendo capazes de representar as informações hierarquicamente mais relevantes ao conteúdo do documento onde estão inseridos. Este vetor de relevância será dado como entrada para o processo de recomendação automática de conteúdos relacionados.

5.1.4 Armazenamento dos OAs no repositório de conteúdos

Finalizados os processos de carregamento, anotação automática e classificação hierárquica de termos e conceitos relevantes, os OAs contêm todas as informações necessárias para posterior composição de uma dada unidade de aprendizagem e submissão aos processos de recomendação de conteúdos relacionados. Assim, o *corpus* onde se encontram pode ser fechado e armazenado no repositório de conteúdos.

5.2 Estratégia para recomendação automática e agregação de OAs relacionados

Uma vez que se tenha um repositório de OAs, devidamente indexados e que tenham sido submetidos aos processos de recuperação de informações relevantes, conforme definido na seção anterior, os mesmos estão prontos para serem submetidos aos processos de busca e recomendação de conteúdos relacionados, de acordo com a estratégia definida neste trabalho de pesquisa. Seus principais processos são apresentados na Figura 9.

Figura 9 - Recomendação e Agregação de OAs – principais etapas

Fonte: Elaborada pelo autor

O primeiro processo consiste na seleção de OAs para composição do conteúdo principal de um pacote SCORM. Em seguida, é executado o processo de recomendação automática e agregação de conteúdos relacionados, em conformidade com a categoria *relation*. Por fim, o usuário responsável pelo desenvolvimento do conteúdo didático-pedagógico pode, então, selecionar, dentre os OAs recomendados, aqueles que efetivamente serão agregados ao conteúdo principal e inseridos no pacote SCORM.

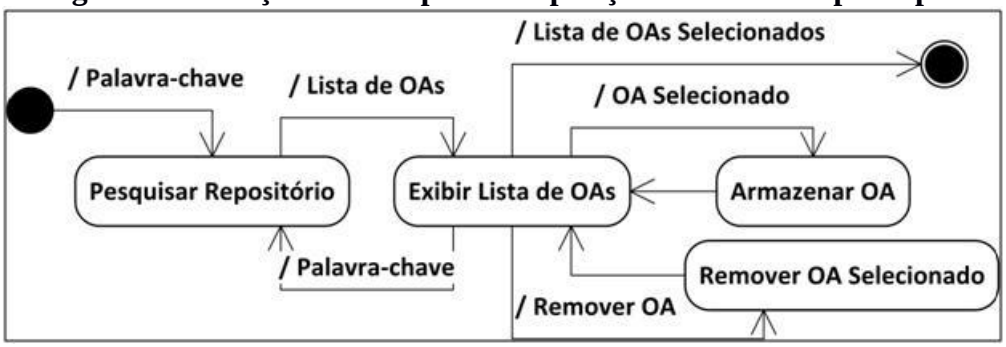
Diante disso, tem-se um processo de recomendação automática, seguido de um processo semiautomático para agregação do conteúdo selecionado. Cada um desses processos é detalhado a seguir.

5.2.1 Montagem do conteúdo principal

O primeiro processo para recomendação de conteúdos relacionados consiste na construção do conteúdo principal do pacote SCORM, quando o repositório é acessado pelo usuário e, a partir dele, são selecionados os OAs que irão compô-lo, conforme ilustrado na

Figura 10.

Figura 10 - Seleção de OAs para composição do conteúdo principal



Fonte: Elaborada pelo autor

Uma vez selecionados os OAs do conteúdo principal, a listagem dos mesmos é dada como entrada para o processo seguinte, que realizará a recomendação automática de conteúdos a eles relacionados, a partir do repositório de conteúdos.

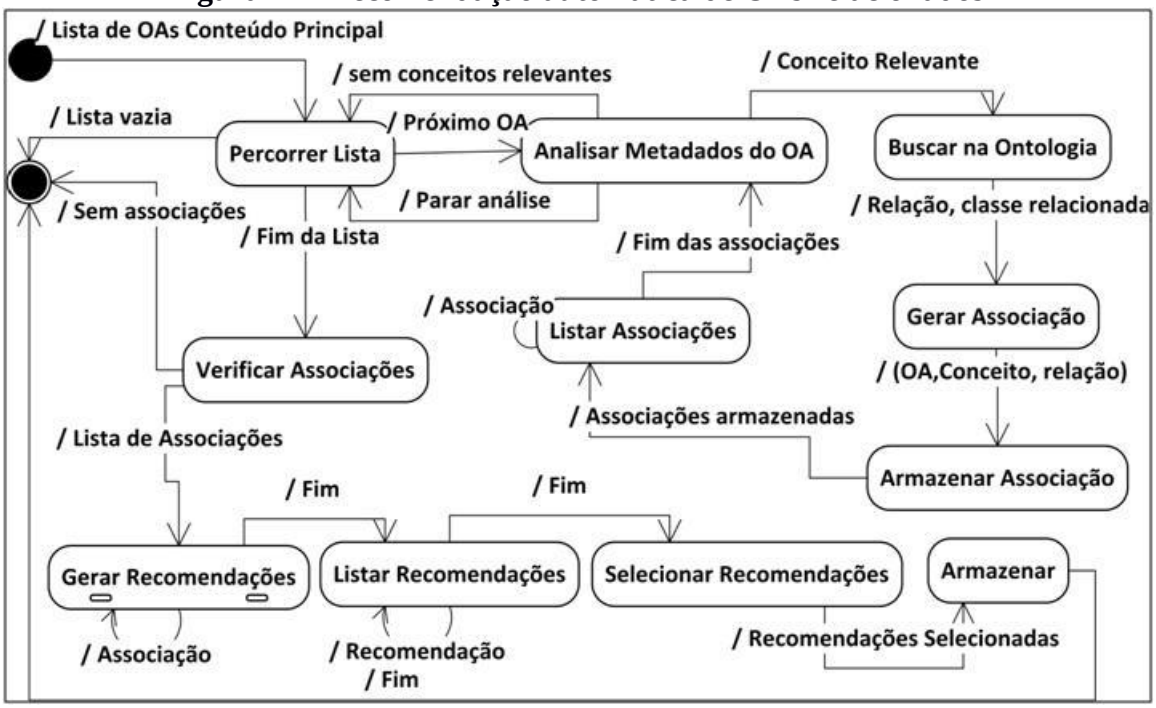
5.2.2 Recomendação automática

As principais etapas do processo de recomendação automática são apresentadas no diagrama da Figura 11. A estrutura do conteúdo principal é percorrida e cada um dos seus OAs tem seus metadados analisados em busca de seus conceitos relevantes, prévia e devidamente anotados e classificados hierarquicamente em seu vetor de relevância. Para cada conceito identificado é estabelecida uma relação com sua classificação na ontologia de domínio que, por sua vez, descreve um grafo no qual as classes de conceitos definem seus vértices e as relações entre elas são definidas por suas arestas. A partir desta relação, entre um conceito anotado e sua classe ontológica, são identificadas outras relações que possam ser estabelecidas entre ele e os demais conceitos presentes em outras classes da ontologia. Sendo assim, para cada relação ontológica identificada, é gerada uma associação do tipo (OA, conceito relevante, relações), onde o número de associações para um OA é definido pela soma das associações geradas para cada um de seus conceitos relevantes, podendo cada um deles ter mais de uma relação dependendo do número de arestas no vértice que define sua classe ontológica. As associações geradas são exibidas ao usuário, enquanto o processo é repetido para o próximo conceito relevante, até que todo o vetor de relevância seja percorrido, ou até que o usuário decida interromper o processo de análise e saltar para o próximo OA.

Uma vez que todos os OAs que compõem o conteúdo principal tenham sido analisados, as diversas associações de conceitos geradas para cada um deles são então

transformadas em recomendações de OAs relacionados, em um processo que será detalhado a seguir. As recomendações realizadas são listadas para que o usuário possa selecionar aquelas que deseja manter. Em seguida, são armazenadas para que possam ser utilizadas no processo de empacotamento, responsável por agregar todo o conteúdo, em conformidade como o SCORM-CAM, e gerar o pacote de conteúdos no formato SCORM.

Figura 11 - Recomendação automática de OAs relacionados

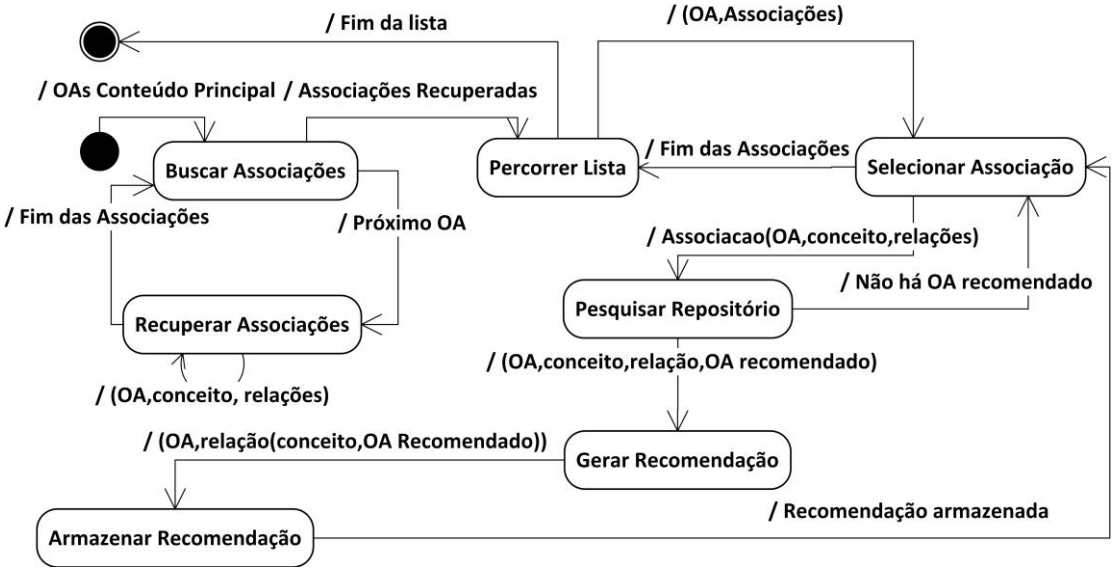


Fonte: Elaborada pelo autor

O processo de Geração de Recomendações, ilustrado na Figura 12, percorre a lista de OAs do conteúdo principal e recupera, para cada um deles, as associações anteriormente geradas e armazenadas. Estas são empregadas para gerar um conjunto de recomendações para cada um dos OAs, a partir dos conceitos e das relações nelas contidos.

Para cada associação de um OA são recuperados seu conceito relevante, as relações nelas descritas e, a partir destas, as classes ontológicas apontadas como associadas ao conceito relevante. A partir destas classes ontológicas é realizada uma busca no repositório de conteúdos por OAs que a elas pertençam e que tenham, entre seus conceitos mais relevantes, o conceito presente na associação que está sendo analisada. Para cada OA encontrado no repositório que corresponda a estes critérios é, então, gerada uma recomendação deste OA como relacionado ao OA do conteúdo principal, para aquele tipo de relação. A Figura 13 destaca o formato de uma recomendação.

Figura 12 - Geração de recomendações a partir das associações de conceitos



Fonte: Elaborada pelo autor

Figura 13 - Formato de uma recomendação

Recomendação((OA, relação (conceito, OA Recomendado))

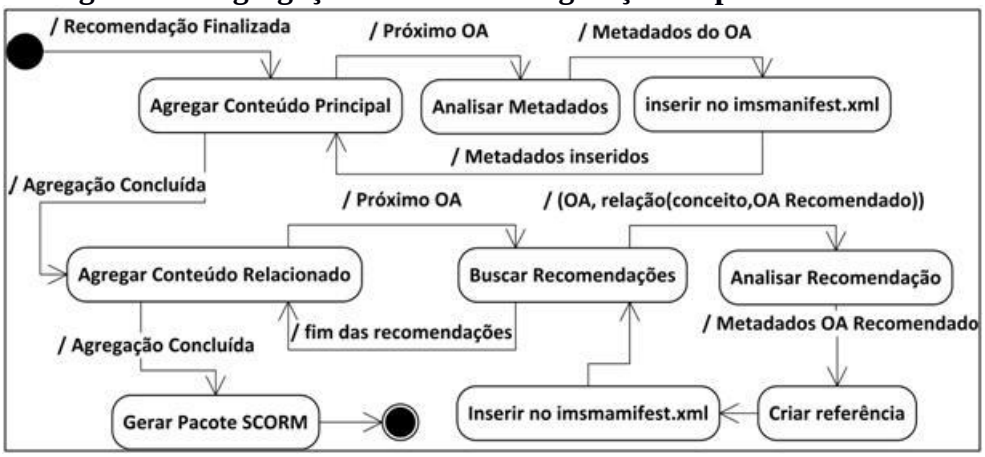
Fonte: Elaborada pelo autor

Todas as recomendações geradas são listadas, permitindo que sejam verificados o título, resumo, palavras-chave e autores de cada OA. Nesse momento, o usuário pode selecionar as recomendações que deseja manter para composição do conteúdo final do pacote SCORM.

5.2.3Empacotamento do conteúdo no formato SCORM

A etapa de empacotamento, ilustrada na Figura 14, contempla os processos responsáveis pela agregação, tanto do conteúdo principal quanto do conteúdo recomendado, gerando, ao final, o pacote SCORM pronto para publicação em um SGA. Nesta etapa, uma lista de OAs é gerada a partir do conteúdo principal e dada como entrada para o processo de agregação do conteúdo, onde o arquivo *imsmanifest.xml* é escrito, armazenando as informações de acesso ao conteúdo por meio do SGA, conforme especificações do SCORM – CAM.

Figura 14 - Agregação de conteúdo e geração do pacote SCORM



Fonte: Elaborada pelo autor

Para cada OA, são recuperados do repositório os metadados relativos ao seu título e à URI contendo seu endereço, para que possa ser criada a devida referência no *imsmanifest.xml.* Após a agregação de todo o conteúdo principal, a lista de OAs é, então, processada para agregação do conteúdo relacionado. Para cada OA do conteúdo principal são, então, recuperadas as recomendações anteriormente geradas para cada um de seus conceitos relevantes. As recomendações são analisadas e os OAs recomendados também têm seu título e URI recuperados do repositório, passando a estar associados ao OA principal, de modo que o OA relativo à recomendação é associado por meio do devido atributo da categoria *relation* e sua referência é, então, inserida no *imsmanifest.xml.* Ao final, todos os OAs são copiados do repositório de conteúdos e encapsulados juntamente com o arquivo *imsmanifest.xml*, compondo, finalmente, o pacote de conteúdos SCORM resultante de todo o processo.

6 RECURSOS DO *FRAMEWORK* GATE

Com base nos trabalhos relacionados, foi selecionado para realização desta pesquisa o *framework* GATE. Também substanciaram tal escolha, a leitura prévia de sua documentação e a análise da estrutura e recursos fornecidos pelo *framework* em relação aos processos demandados para este trabalho. Nesta seção será apresentado o GATE, com ênfase nos recursos e ferramentas nele presentes e que foram efetivamente utilizados.

Devido à sua arquitetura, que agrupa funcionalidades em módulos independentes, disponibilizados na forma de *plugins*, o GATE pode ser utilizado por meio de sua própria IDE, denominada *Gate Developer*, para construção de aplicações padrão e que exigem pouca customização, ou pode ser embutido em aplicações Java para maior flexibilidade na implementação de diferentes processos, estratégias e metodologias, como é o caso do trabalho desenvolvido nesta pesquisa. É capaz de processar arquivos em uma grande variedade de formatos, tais como HTML, PDF, RTF, XML e *Microsoft Word*, dentre outros, a partir de repositórios de conteúdos locais, em rede ou na Web. Sendo documentos individuais, agrupados em um *corpus* ou armazenados em repositórios, estes arquivos podem ser submetidos a diferentes processos de geração e anotação de metadados, recuperação da informação e mineração de textos, empregando-se diferentes técnicas e algoritmos, tais como análise de modelos, aplicação de regras de inferência, aprendizagem de máquina e compreensão de linguagem natural. Além disso, no *Gate Developer* estão disponíveis ferramentas, com interface gráfica, para criação e edição de ontologias e dicionários, inserção e edição manual de metadados e construção de aplicações.

Sendo assim, devido à sua arquitetura e ao conjunto de ferramentas que oferece para processamento de texto e recuperação de informação, o GATE se constitui como um *framework* ao mesmo tempo robusto e flexível, acompanhado de uma documentação detalhada e abrangente.

6.1 Recursos de processamento e recursos de linguagem

A estrutura básica de organização do GATE divide os recursos a serem empregados em Recursos de Processamento e Recursos de Linguagem. O primeiro grupo consiste dos *plugins* e recursos de processamento para uma dada aplicação. O segundo compreende arquivos a serem analisados, um ou mais *corpus* de documentos, armazenados ou não em repositórios, ontologias, *lexicons*, modelos e esquemas de anotação, caso sejam empregados.

Cada *plugin* executa uma atividade específica. Muitas vezes um determinado passo de processamento, realizado por um *plugin*, necessita da saída gerada por outro *plugin*, em um passo anterior. Assim, é necessário que estes estejam organizados na forma de um *pipeline*, permitindo que a ordem de execução dos processos permita a correta geração de entradas e saídas para cada plugin.

Uma vez organizado o *pipeline* de processamento, com os *plugins* necessários ao trabalho que se deseja realizar, este será executado sobre um dado conjunto de Recursos de Linguagem. Pode ser submetido ao *pipeline* um único arquivo ou um *corpus* com vários arquivos. Como saída, tem-se todos os arquivos submetidos ao processamento, mais os dados resultantes, tais como os metadados gerados e as anotações inseridas nos documentos.

6.2 Seleção de recursos de processamento do GATE

Esta seção apresenta, de forma sucinta, os principais Recursos de Processamento utilizados para realização desta pesquisa. Foram selecionados por suas funcionalidades, com base nos processos demandados pela metodologia proposta neste trabalho e apresentada na Seção 5.

6.2.1 JAPE – Java Annotation Patterns Engine

O JAPE é um transdutor que emprega máquinas de estados finitos para reconhecimento de anotações em textos a partir de expressões regulares. Sua gramática define um conjunto de fases, cada uma delas consistindo de modelos e regras, que são executadas sequencialmente compondo uma cascata de transdutores sobre as anotações. A documentação do GATE detalha a sintaxe empregada na descrição das regras e modelos, que são armazenados em arquivos texto e disponíveis para edição.

Sendo utilizado por diversos *plugins* e recursos de processamento, o JAPE constitui-se como um dos componentes mais importantes do GATE.

6.2.2 O plugin ANNIE

O *plugin* denominado *Nearly-New Information Extraction System* (ANNIE) baseia-se em algoritmos de Máquinas de Estados Finitos e utiliza o JAPE para processamento de Expressões Regulares. Para as etapas de pré-processamento e pós-processamento, o ANNIE utiliza-se ainda de outros *plugins* do *framework*, que devem ser organizados dentro de um

pipeline conforme as etapas de processamento a serem realizadas e os resultados a serem obtidos. A seguir, são apresentados, de forma sucinta, os recursos de processamento e *plugins* utilizados pelo ANNIE.

6.2.2.1 O recurso de processamento *Document Reset*

Este recurso de processamento, em sua configuração padrão, permite remover de um documento todos os conjuntos de anotações nele presentes, inclusive marcações de formatação. Através de seus parâmetros é possível indicar, ainda, uma lista dos tipos de anotações a serem removidas de um conjunto, fazendo com que as demais sejam mantidas. Da mesma forma, é possível indicar conjuntos inteiros de anotações a serem removidas. A remoção de marcações de formatação é opcional.

6.2.2.2 O recuso de processamento *Tokeniser*

Este recurso de processamento aplica regras de inferência para dividir o texto de um documento em símbolos simples, tais como números, pontuação, espaço em branco e palavras de diferentes tipos. Seu objetivo é minimizar o trabalho na etapa de processamento e análise de símbolos, intensificando a carga de trabalho sobre as regras gramaticais, por serem mais flexíveis e poderem ser adaptadas, alteradas e estendidas. A Figura 15 apresenta um exemplo de regra simples, que se aplica a palavras que começam com letra maiúscula.

Figura 15 - Exemplo de regra

```
‘UPPERCASE_LETTER’ ‘LOWERCASE_LETTER’* >
Token;orth=upperInitial; kind=word;
```

Fonte: Cunningham, 2012

6.2.2.3 O recuso de processamento *English Tokeniser*

Este Recurso de processamento utiliza o *JAPE Transducer* para processar a saída genérica do *Tokeniser*, aplicando uma série de regras de modo a adaptá-la ao formato requerido para o processamento das classes gramaticais.

6.2.2.4 O recurso de processamento *Gazetter*

O *Gazetter* identifica nomes de entidades em textos a partir de listas de termos, presentes em arquivos-texto e agrupadas por um arquivo principal. A organização das listas e sua entrada, uma para cada linha, no arquivo principal, permite estabelecer relações taxonômicas entre os termos de diferentes listas, agrupando-os como principais ou secundários, por meio de parâmetros que compõem a entrada de cada lista. Esta estrutura de entrada é ilustrada na Figura 16.

Figura 16 - Estrutura de entrada de uma lista de termos no arquivo principal

```
…
Nome_da_lista : tipo_principal : tipo_secundario
Lista_de_paises : localização : país
Lista_de_cidades : localização : cidade
Numeros : número
Numeros_ordinais : número : ordinal
...
```

Fonte: Elaborada pelo autor

Cada entidade identificada no texto, pelo *Gazetter*, a partir de uma das listas do arquivo principal, recebe uma anotação *Lookup*, que indica o tipo principal e o tipo secundário, associados àquela lista e identificados pelos parâmetros de sua entrada. É possível determinar um segundo tipo de anotação para uma lista específica, acrescentando-a como último parâmetro na linha de entrada da lista, no arquivo principal.

O dicionário de termos elaborado para compor a base de conhecimento de domínio, empregada neste trabalho, foi formatado em conformidade com o *Gazetter*.

6.2.2.5 O recurso de processamento *OntoGazetter*

O *OntoGazetter* consiste em um dos principais Recusos de Processamento utilizados neste trabalho. Permite associar os termos contidos em um dicionário, elaborado de acordo com o *Gazetter*, a determinadas classes de uma ontologia, por meio de um editor de ontologias e de um editor de mapas que associam os arquivos de listas do dicionário com as respectivas classes. Dessa forma, este importante plugin associa aos termos anotados um metadado indicando sua classe ontológica, de acordo com o mapeamento e a respectiva

ontologia.

Este *plugin*, na realidade, não é contemplando pelo *ANNIE* em sua composição padrão, oferecida pelo GATE. Presente no *framework* como um *plugin* adicional, foi introduzido no *pipeline* especificamente para desenvolvimento deste trabalho, devido à necessidade de utilização de uma ontologia de domínio, conforme discutido nas seções anteriores.

6.2.2.6 O recuso de processamento *Sentence Splitter*

Trata-se de um transdutor de estados finitos que divide o texto em sentenças, identificando-as por meio de uma anotação do tipo *Sentence.* No caso dos finais de frases, uma anotação *Split* é inserida, podendo ser de dois tipos: *internal* ou *external.* O primeiro tipo refere-se à presença de símbolos de interrogação, exclamação ou dois pontos, enquanto o segundo refere-se a uma quebra de linha no texto.

6.2.2.7 O recurso de processamento *Part of Speech Tagger*

Este recurso de processamento identifica a classe gramatical das palavras a partir de um dicionário léxico e de um conjunto de regras, obtidos como resultado do treinamento de algoritmos de aprendizagem de máquinas sobre um extenso *corpus*, obtido do *Wall Street Jounal*. Os termos são anotados com *tags* específicas, associadas a cada classe gramatical, identificando-os como substantivos, pronomes, adjetivos, dentre outras classes.

6.2.2.8 O recurso de processamento *Semantic Tagger*

Aplicando uma série de regras baseadas na linguagem JAPE, este recurso de processamento atua sobre as anotações inseridas em fases anteriores, para identificar entidades semânticas nas frases em que se inserem.

6.2.2.9 O recurso de processamento *OrthoMatcher*

Este recurso identifica relações de identidade entre as entidades anotadas pelo *Semantic Tagger*, de modo a estabelecer correferências. Seu objetivo não é encontrar e nomear novas entidades, mas atribuir um tipo a uma entidade não classificada, a partir de uma entidade correspondente já classificada, empregando um processo de *matching*.

6.2.3 SerialDataStore

Além dos diversos *plugins*, o GATE também disponibiliza uma API completa para desenvolvimento de aplicações em Java. Um dos componentes desta API, importante para realização deste trabalho, consiste na classe *SerialDataStore*, disponível no pacote "gate.persist", que permite armazenar e manipular um *corpora* com grande número de documentos. Este componente oferece recursos para persistência, organização e recuperação de documentos serializados em disco, no momento exato em que precisem ser processados.

6.2.4 Ontology

A classe *Ontology*, disponível no pacote "gate.creole.ontology", oferece uma série de recursos para processamento de ontologias, possibilitando a criação, o carregamento e o *parser* de ontologias nos formatos OWL e RDF, além da edição e manipulação de classes, atributos e relações.

6.3 Simulação com o *GATE Developer*

A utilização de uma aplicação básica, construída no *Gate Developer*, não seria suficiente para implementar a metodologia desenvolvida neste trabalho, devido aos vários processos específicos por ela definidos e respectivas estratégias de implementação. Todavia, quando da seleção do *framework* a ser utilizado, alguns processos foram simulados, conforme descrito nesta seção, no intuito de verificar a viabilidade de sua utilização para o desenvolvimento deste trabalho. Dessa forma, o *Gate Developer* foi instalado e configurado. Apenas o *plugin* ANNIE, em sua composição padrão, foi carregado e configurado em um uma aplicação. A base de conhecimento de domínio ainda não continha a ontologia, mas apenas parte do dicionário de termos, pois a mesma ainda estava sendo desenvolvida para aplicação neste trabalho e os termos para composição do dicionário ainda estavam sendo coletados. Assim, os recursos para processamento com base na ontologia não foram utilizados.

Foram selecionados nove documentos em PDF e vinte e três em HTML, disponibilizados livremente na Web, cujos conteúdos se referem à historia da música, aprendizado de música e teoria musical. No GATE, foram configurados dois repositórios, para serialização e indexação destes documentos, sendo um deles para os aquivos em PDF e o outro para os aquivos em HTML. Da mesma forma, foram criados dois *corpus* separados por repositório, denominados "*OA Music PDF*" e "*OA Music HTML*", para agrupamento dos

respectivos documentos, que foram, assim, carregados e armazenados nos repositórios.

Após a execução da aplicação, os termos presentes no dicionário foram devidamente anotados nos documentos. Como pode ser visto na Figura 17, o compositor “Mozart” foi devidamente identificado e anotado com os metadados do tipo *Lookup* (Seção 6.2.2.4), indicando que se trata de um termo pertencente ao domínio da música, no grupo de compositores.

Figura 17 - Documento anotado e exportado para XML


```
id="11119"/> <Node id="11120"/>came<Node id="11124"/> <Node id="11125"/>to<Node id=
/>involve<Node id="11135"/> <Node id="11136"/>baroque<Node id="11143"/> <Node id="11144"
<Node id="11150"/>
de id="11156"/> <Node id="11157"/>works<Node id="11162"/> <Node id="11163"/>by<Node id=
/>Haydn<Node id="11171"/>,<Node id="11172"/> <Node id="11173"/>Mozart<Node id="11179"/>
<Annotation Id="4068" Type="Token" StartNode="11173" EndNode="11179">
<Feature>
  <Name className="java.lang.String">length</Name>
  <Value className="java.lang.String">6</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">orth</Name>
  <Value className="java.lang.String">upperInitial</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">kind</Name>
  <Value className="java.lang.String">word</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">string</Name>
  <Value className="java.lang.String">Mozart</Value>
</Feature>
</Annotation>
```


Fonte: Elaborada pelo autor

Para prosseguimento do trabalho fez-se necessário que a base de conhecimento de domínio estivesse completa, visando maior precisão no processo de anotação automática, reduzindo o número de termos irrelevantes ao domínio e ampliando o reconhecimento de informações relevantes, conforme previsto na metodologia.

O *pipeline* definido nesta etapa de exploração ainda não refletia a estratégia para recuperação de informações relevantes definida neste trabalho, pois apenas anotava os termos identificados como pertencentes ao dicionário. Todavia, ficou claro que os processos de geração e anotação automáticas de metadados contemplariam os resultados esperados, cumprindo uma etapa inicial, mas fundamental à realização do restante desta pesquisa.

Para implementação do sistema de recomendação de conteúdos relacionados,

desenvolvido como parte deste trabalho e que aplica a metodologia proposta, não foi utilizado o *GATE Developer*. Os *plugins* do *framework* foram importados separadamente em uma ferramenta para desenvolvimento de *softwares* utilizando a linguagem de programação Java.

7 SISTEMA DE RECOMENDAÇÃO E AGREGAÇÃO DE CONTEÚDOS RELACIONADOS

Nesta seção apresenta-se o trabalho de implementação realizado como parte desta pesquisa. São apresentadas e detalhadas as estratégias e soluções desenvolvidas para atender a cada um dos processos definidos na metodologia.

Para desenvolvimento do protótipo de um sistema de recomendação e agregação de conteúdos, foi utilizada a IDE NetBeans. Foi criado um projeto de aplicação Java, para o qual foram importados os *plugins* do GATE e implementados os diversos módulos necessários a cada etapa de processamento: *AssignerRelevance*, para anotação automática de metadados e classificação hierárquica dos conceitos relevantes; *AssociationsBuilder*, para geração de associações entre os conceitos relevantes e a ontologia de domínio; *RecommendationsBuilder*, para geração de recomendações de conteúdos relacionados, a partir das associações preestabelecidas; e *DocScoreRecommendationsBuilder*, que atribui um *score* a cada documento recomendado, identificando o que melhor corresponde à relação a ser estabelecida.

A Seção 7.1 apresenta a base de conhecimento de domínio, contendo o dicionário de termos e a ontologia de domínio. As demais seções discutem cada uma das etapas de processamento e o papel de cada um dos módulos implementados.

7.1 Base de conhecimento de domínio

Os conteúdos dos OAs utilizados neste trabalho pertencem ao domínio da área de conhecimento da Música Erudita. Dessa forma, seus termos e conceitos relevantes são automaticamente anotados a partir de uma base de conhecimento, composta de um dicionário de termos e de uma ontologia de domínio, elaborados e desenvolvidos como parte desta pesquisa. Os elementos presentes no dicionário contemplam a ontologia, enquanto esta permite agrupá-los em classes de conceitos e identifica as ligações e relações taxonômicas entre eles. O dicionário contém 37.183 termos e conceitos, distribuídos em 47 arquivos. Estes arquivos recebem nomes que indicam o grupo de termos neles contidos e uma extensão de arquivo “lst”. Um arquivo principal, que lista todos os demais arquivos e define a classificação taxonômica, recebe o nome “lists.def”. A Figura 18 apresenta parte deste arquivo.

Figura 18 - Parte do arquivo "lists.def"

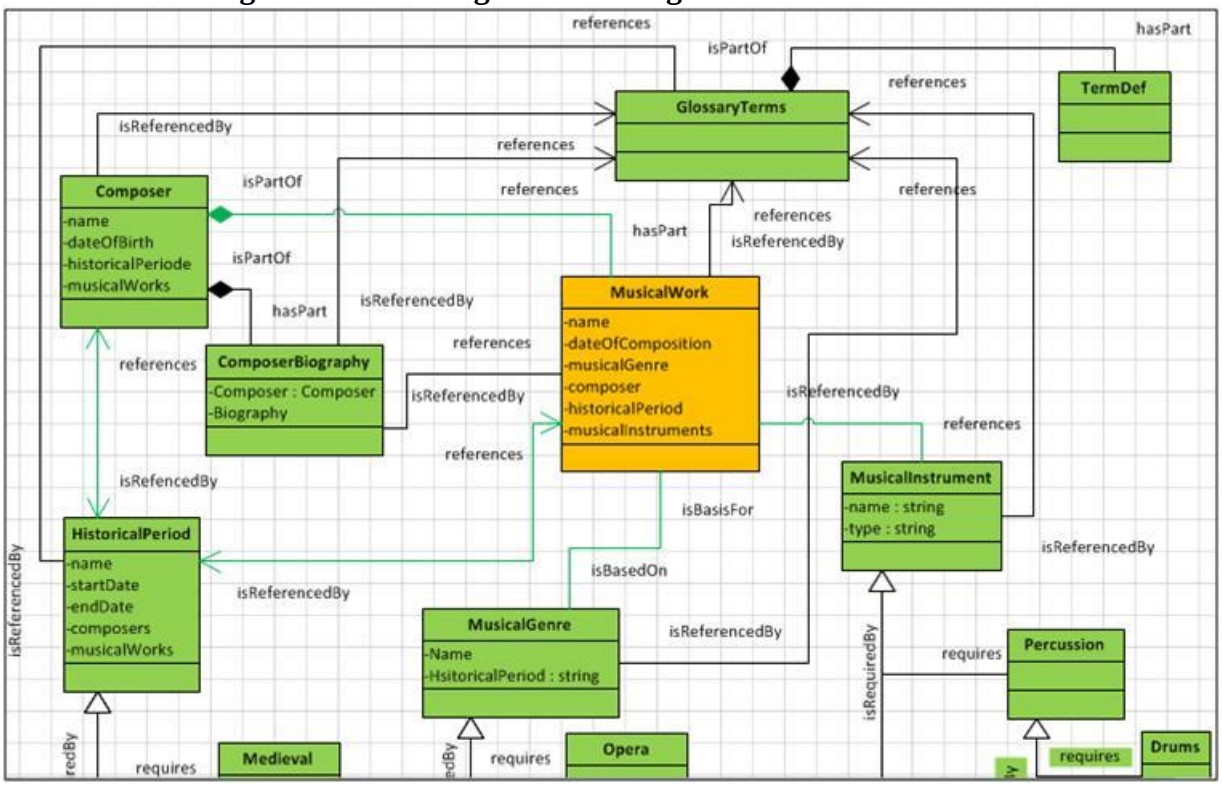
```
historical_period_late_romantic_dates.lst:late_romantic:date
historical_period_medieval.lst:historical_period:medieval
historical_period_medieval_dates.lst:medieval:dates
historical_period_renaissance.lst:historical_period:renaissance
historical_period_renaissance_dates.lst:renaissance:dates
historical_period_twenty_centure.lst:historical_period:twenty_century
historical_period_twenty_centure_dates.lst:twenty_centure:dates
musical_genre.lst:musical_genre
musical_genre_chanber.lst:musical_genre:genre_chanber
musical_genre_opera.lst:musical_genre:genre_opera
musical_genre_orchestral.lst:musical_genre:genre_orchestral
musical_genre_solo_instrumental.lst:musical_genre:genre_solo_instrumental
musical_genre_vocal.lst:musical_genre:genre_vocal
musical_genre_orchestral_ballet.lst:musical_genre_orchestral:ballet
musical_genre_orchestral_incidental_music.lst:musical_genre_orchestral:incidental_music
musical_genre_orchestral_overture.lst:musical_genre_orchestral:overture
```

Fonte: Elaborada pelo autor

A ontologia a ser utilizada precisava apresentar uma classificação que contemplasse a área de domínio da Música Erudita. A única ontologia encontrada foi a OntoMusica. Todavia, esta é bastante restrita, contendo poucas classes de conceitos e não permitindo o estabelecimento das relações necessárias para a realização desta pesquisa. Sendo assim, foi proposta uma nova ontologia, contendo 39 classes, distribuídas entre superclasses e subclasses, com 31 relações distribuídas entre elas. Suas classes e respectivas relações foram modeladas com a *Unified Modeling Language* (UML)[8] e construídas utilizando o editor de ontologias do GATE, conforme ilustrado nas Figuras 19 e 20.

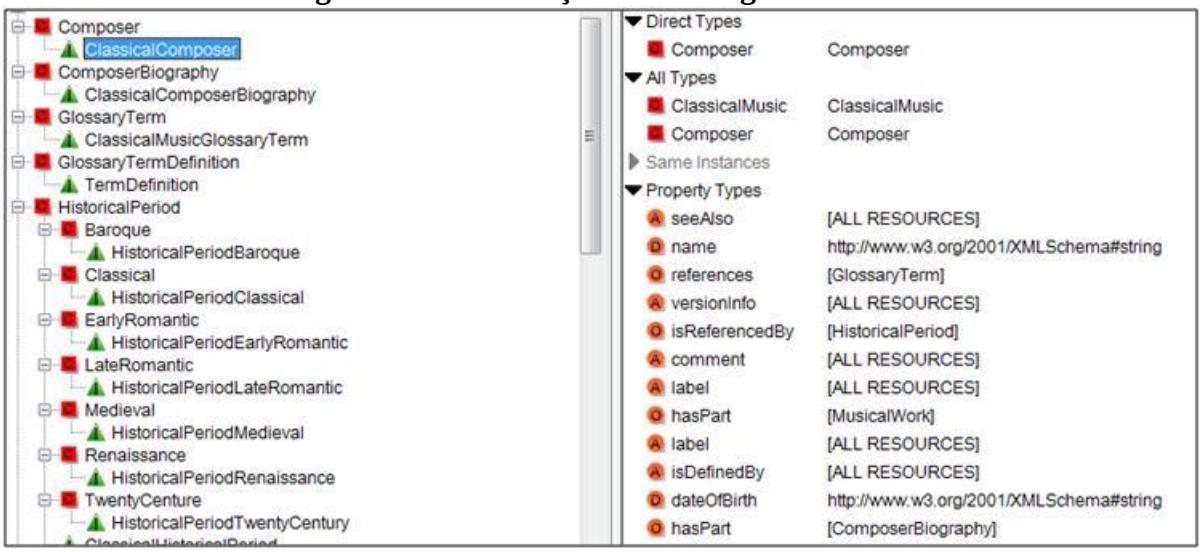
[8] A UML consiste em um conjunto de especificações para modelagem de dados e sistemas. Permite a criação de classes de objetos, que representam objetos e entidades do mundo real, estabelecendo diferentes relações entre eles.

Figura 19 - Modelagem da ontologia de domínio – recorte



Fonte: Elaborada pelo autor

Figura 20 - Construção da ontologia – recorte



Fonte: Elaborada pelo autor

A proposta de modelagem da ontologia via UML partiu do princípio de que as relações a serem estabelecidas não se apoiariam na estruturação dos documentos, considerando tópicos e subtópicos como foi feito por Lu e Hsieh (2009). Elas seriam estabelecidas entre classes de documentos, definidas com base na organização dos temas apresentada pelo *site* NAXOS e na caracterização dos textos a partir de seu tema central. Esta

caracterização se deve ao fato de que as relações que se busca estabelecer neste trabalho não visam documentos quaisquer, mas documentos que consistem em OAs, entendidos como uma unidade de aprendizagem cujo conteúdo apresenta um determinado assunto e se encerra em si, podendo ou não ser estendido, mas, de todo modo, sendo capaz de ser compreendido por si só (HERNANDÉZ, 2009).

Dessa forma, utilizando um diagrama de classes, foi possível mapear as relações definidas na categoria *relation* em relações estabelecidas na UML, por meio de associações, agregações, heranças e especializações. Esta abordagem permitiu conferir um caráter semântico às relações da categoria *relation*, quando se tem, por exemplo, duas classes denominadas *Composer* e *ComposerBiography*, cuja associação se dá por meio de uma agregação, com a qual se estabelece que *Composer* possui uma relação do tipo *hasPart* com *ComposerBiography* e esta, no sentido contrário, estabelece uma relação do tipo *isPartOf* com *Composer*. Haja vistas que se o diagrama de classes estivesse modelando um sistema para catalogação de compositores, por exemplo, a relação de agregação com sua biografia estaria coerente. Da mesma forma, observa-se que as relações *hasPart* e *isPartOf* podem conter o mesmo significado denotado à agregação da UML. Em outro exemplo, uma obra musical é parte de um compositor, no sentido de agregação em que se um compositor deixa de existir, suas obras também deixam. Logo, trata-se de uma agregação do tipo *hasPart* e *isPartOf*, entre as classes *Composer* e *MusicalWork.*

O mesmo foi percebido para as demais relações e respectivas associações na UML. Entre compositor e período histórico, pode-se considerar que um compositor está associado a um período, mas se este período deixa de existir no estudo da história, o mesmo não ocorre com os elementos concretos que estavam a ele associados. Assim, se um compositor estava associado a um dado período histórico, e este deixou de existir, o mesmo pode ser novamente associado a outro período. Dessa forma, adotou-se o entendimento de que um compositor tem como referência um período histórico e um período histórico referencia um compositor, recebendo as associações *references* e *isReferencedBy.* Por sua vez, a compreensão da história da música, requer a compreensão de cada um de seus períodos, na medida em que cada período pode ser visto como uma especialização da classe *HistoricalPeriod*. O mesmo ocorre com gêneros musicais, onde o conhecimento dos mesmos requer o conhecimento de cada um em particular, sendo estes uma especialização da classe dos gêneros musicais. Nestes casos, tem-se uma relação do tipo *requires* e *isRequiredBy*, no sentido que a classe geral requer classes especializadas e estas são requeridas por ela. No caso de obra musical e gênero musical, tem-se que, se uma obra musical se caracteriza como uma ópera, então teve sua

composição baseada neste gênero musical, que por sua vez lhe serviu de base, em uma associação. Assim, tem-se uma relação do tipo *isBasisFor* e *isBasedOn*.

As relações recebem os seguintes pesos para associação dos metadados no SCORM: *requires/isRequiredBy* e *isBasisFor/isBasedOn* são relações fortes no sentido didático-pedagógico, onde um conteúdo necessita do outro para ser compreendido, logo, os OAs estarão relacionados como pré-requisitos. Por sua vez, *references/isReferencedBy* e *isPartOf /hasPart* pressupõem relações de complementação não obrigatória, onde os documentos se complementam mas não dependem uns dos outros para serem compreendidos, logo, os OAs não serão relecionados como pré-requisitos, mas apenas como material complementar.

Por fim, foi criado um mapa, por meio do editor do GATE, que associa cada lista de termos do dicionário a uma classe da ontologia. Este mapa é armazenado em um arquivo texto que, juntamente com o dicionário de termos e o arquivo RDF que contém a ontologia, compõe a base de conhecimento de domínio definida e empregada nesta pesquisa.

7.2 Recuperação de informações relevantes

Conforme discutido na Seção 5.1, para que se possa recomendar e agregar OAs relacionados é necessário que tenham sido submetidos a uma etapa de pré-processamento, que tem como objetivo recuperar informações relevantes aos seus conteúdos, possibilitando a análise e identificação de possíveis relações entre eles. Para implementação desta etapa foi utilizado um *plugin* do *framewok* GATE denominado ANNIE, discutido em Cunningham e outros (2012), e o módulo *AssignerRelevance*, implementado como parte deste trabalho. O ANNIE subdivide o texto em símbolos e sentenças, anotando os termos quanto à sua classe gramatical e indicando sua classe ontológica, por meio dos recursos de processamento apresentados na Seção 7 deste trabalho. O *AssignerRelevance*, por sua vez, implementa os algoritmos para geração dos demais metadados, definidos para esta etapa, aos quais foram associadas as métricas necessárias à recuperação da informação, assim como realiza a classificação hierárquica destes elementos, a partir da relevância atribuída a cada um deles.

A partir da base de conhecimento de domínio, os processos de geração e anotação automática dos metadados, assim como a classificação hierárquica dos termos e conceitos relevantes, são realizados quando da inserção de novos OAs no repositório de conteúdos, como definido pela metodologia proposta. Para manipulação do repositório foi implementado um *Serial Data Store*, utilizando a API do GATE, que serializa e armazena os OAs no respectivo *corpus*.

Uma vez que novos OAs a serem inseridos no repositório de conteúdos tenham sido carregados pelo sistema a partir de seu repositório original, a etapa seguinte consiste nos processos de geração e anotação automática dos metadados em cada um dos documentos. Para isso, o sistema utiliza o plugin ANNIE. Este recebe como entrada o *corpus* contendo os OAs, o dicionário de termos, seu mapeamento para a ontologia e a ontologia de domínio. A saída do ANNIE consiste nos OAs contendo termos e conceitos anotados quanto à sua classe gramatical e sua classificação ontológica. Como uma etapa de preparação dos OAs para a recomendação de conteúdos a eles relacionados, após a anotação dos metadados, é necessário que os termos e conceitos anotados sejam classificados quanto ao seu nível de relevância, em relação ao conteúdo como um todo. Por fim, os OAs são persistidos, com as devidas anotações, no repositório de conteúdos.

Para a classificação hierárquica de termos chave e conceitos relevantes, o sistema utiliza o módulo *AssignerRelevance*, que foi implementado de modo a receber como entrada uma lista de OAs e fornecer como saída a mesma lista, porém com novos metadados inseridos em cada OA, para cada termo ou conceito anteriormente anotado, e uma classificação hierárquica de seu conjunto com base no nível de relevância de cada um deles em relação ao texto como um todo. Os novos metadados inseridos pelo *AssignerRelevance* contêm métricas de relevância para cada um dos termos e conceitos. A partir da análise dos novos metadados gerados, procura-se associar ao termo ou conceito um determinado peso, que pode ser inferido com base em um conjunto de diferentes indicadores de relevância. O Quadro 1 relaciona alguns destes indicadores com as abordagens comumente utilizadas para mensurá-los (MORAIS; AMBRÓSIO; 2007). Utilizando uma combinação das abordagens presentes no Quadro 1, o cálculo da relevância para cada termo anotado é realizado a partir da formulação proposta neste trabalho, ilustrada na Figura 21 e apresentada a seguir.

Quadro 1 – Indicadores de relevância e possíveis abordagens

Termos mais utilizados ao longo do texto, à exceção das *stop words*.	Cálculo das frequências, absoluta e relativa, do termo no respectivo documento.
Termos presentes em títulos, palavras-chave e resumos.	Localização do termo nas diferentes seções do texto a partir de uma análise estrutural do documento.
Termos que são substantivos e complementos.	Análise semântica e identificação da posição sintática do termo, na sentença em que ocorre.
Termos que podem estar sendo definidos por outro termo da sentença.	Análise semântica e sintática para verificar as relações entre dois termos na sentença. Ex.: Mozart is a composer.

Fonte: Elaborado pelo autor

A partir das formulações apresentadas em Morais e Ambrósio (2007), Roy, Sudeshna e Sujoy (2008), Tuarob, Pouchard e Giles (2013), foi proposta para este trabalho a seguinte formulação: seja **VT** um vetor de termos relevantes, $\mathbf{R_{ti}}$ a relevância de um termo t_i, $\mathbf{F_{abs(ti)}}$ a frequência absoluta de um termo t_i, $\mathbf{F_{rel(ti)}}$ a frequência relativa de um termo t_i, $\mathbf{T_{tit}}$ um termo presente no título de um documento d_i, $\mathbf{T_{KW}}$ um termo presente nas palavras-chave de um documento d_i, $\mathbf{T_S}$ um termo que é substantivo, $\mathbf{T_{Frel}}$ o termo de maior frequência relativa e $\mathbf{Sent_{ti}}$ a sentença onde o termo ocorre. Definem-se as seguintes funções: (1) retorna a frequência absoluta do termo t_i para o documento d_i; (2) retorna a frequência relativa do termo t_i para o documento d_i, onde **N** é o número total de termos no documento; (3) recebe um termo t_i e a sentença onde ele ocorre e retorna 1.5, caso seja seguido de um substantivo precedido por um verbo de ligação, aumentado sua relevância em, ou 1 caso contrário; (4) retorna um fator de relevância de um termo t_i para o documento d_i, onde ocorre, a partir de uma avaliação com base na combinação dos parâmetros em $\mathbf{F_{rel(ti)}}$, $\mathbf{T_{tit}}$, $\mathbf{T_{KW}}$ e $\mathbf{T_S}$; (5) retorna a relevância final de um termo t_i para o documento d_i onde ocorre, utilizando-se dos valores retornados por (3) e (4) para confirmar o peso de (1).

Figura 21 - Funções definidas para o cálculo de relevância

$$F_{abs(t_i)} = FreqAbs(t_i, d_i) \quad (1)$$

$$F_{rel(t_i)} = FreqRel(t_i, d_i) = \frac{FreqAbs(t_i, d_i)}{N} \quad (2)$$

$$FuncDef\left(t_i, Sent_{t_i}\right) \quad (3)$$

$$R(t_i, d_i) = \begin{cases} 2.0, (t_i = T_{Frel}) \wedge \left((t_i = T_{tit}) \vee (t_i = T_{KW})\right) \wedge (t_i = T_s) \\ 1.5, \left((t_i = T_{tit}) \vee (t_i = T_{KW})\right) \wedge (t_i = T_s) \\ 1.0, \left((t_i = T_{tit}) \vee (t_i = T_{KW})\right) \\ 0.75, (t_i = T_{Frel}) \wedge (t_i = T_s) \\ 0.25, (t_i = T_s) \end{cases} \quad (4)$$

$$FuncRel(t_i, d_i) = \begin{cases} F_{abs(t_i)} \; x \; R(t_i, d_i) \; x \; FuncDef\left(t_i, Sent_{t_i}\right), R(t_i, d_i) > 0 \\ 0 \end{cases} \quad (5)$$

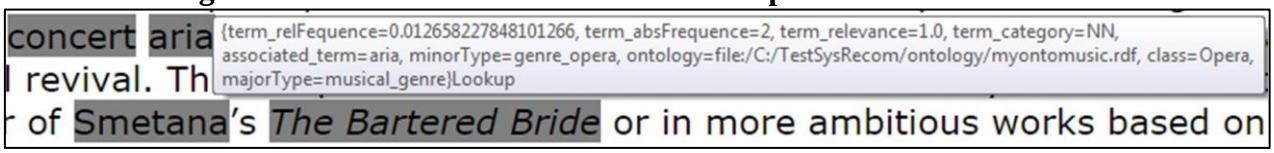
Fonte: Elaborada pelo autor

Empregando-se as funções definidas na Figura 21, para cada termo t_i em d_i, calcula-se $\mathbf{F_{abs(ti)}}$ = **FreqAbs** (t_i, d_i) e $\mathbf{F_{rel(ti)}}$ = **FreqRel** (ti, di). Para todo termo t_i em d_i define-se assim sua relevância $\mathbf{R_{ti}}$: $\mathbf{R_{ti}}$ = [t_i, **FuncRel** (t_i, d_i)]. Se $\mathbf{R_{ti}}$ >=0.25, **VT** ← [ti, R_{ti}]. A variação entre 0.25 e 2.0 para o retorno de (4) divide, mantém ou dobra o peso inicial atribuído ao termo por

(1), gerando assim o seu valor de relevância final.

A Figura 22 ilustra parte de um OA contendo os termos anotados e respectivos metadados gerados até esta etapa. A Figura 23 apresenta parte do arquivo gerado para registro dos dados gerados para cada AO.

Figura 22 - Termo anotado em um OA e respectivos metadados.



Fonte: Elaborada pelo autor

Figura 23 - Parte do arquivo de registros das anotações e metadados gerados.

```
-->128
AnnotationImpl: id=2646; type=Lookup; features={term_relFequence=0.012658227848101266,
term_absFrequence=2, term_relevance=1.0, term_category=NN, associated_term=aria,
minorType=genre_opera, ontology=file:/C:/TestSysRecom/ontology/myontomusic.rdf,
class=Opera, majorType=musical_genre}; start=NodeImpl: id=1682; offset=4111;
end=NodeImpl: id=1683; offset=4115
```

Fonte: Elaborada pelo autor

7.3 Construção de associações

Uma vez cumpridas todas as etapas anteriores, para cada OA presente no repositório, tem-se como resultado um VT, associado a cada documento. A próxima etapa consiste em se identificar as possíveis relações para cada OA, com base nos elementos de VT. Para isso, o sistema utiliza o módulo *AssociationsBuilder*, que recebe como entrada a lista de OAs presentes no *corpus* e fornece como saída as associações possíveis para estes documentos, com base nas relações descritas pela ontologia de domínio, a partir das classes às quais os termos e conceitos relevantes em VT estão associados. Estas associações são inseridas em forma de metadados em cada um dos OAs. Para isso, também foi desenvolvido um *parser* para a ontologia de domínio, utilizando-se recursos da API do GATE, que é utilizado pelo *AssociationsBuilder* e foi chamado de *OntologyParser*.

Assim, para cada elemento de VT, seus metadados são analisados e a classe à qual se associam na ontologia, anteriormente anotada, é resgatada. A partir dela, o *parser* retorna a superclasse e as subclasses a ela associadas, assim como as possíveis relações que estabelece com as demais classes, que foram definidas durante a modelagem da ontologia e obedecem ao vocabulário da categoria *relation* do SCORM: *requires* e *isrequiredby*, *ispartof* e *haspart*, *references* e *isreferencedby*, *isbasedon* e *isbasisfor*. Cada associação contém o termo relevante, sua classe, subclasses e as relações, que têm sua classe como domínio e a classe

associada como alcance, formando um grafo sobre a ontologia onde as classes são os nós e as relações são as arestas que as conectam. Concluídas as associações, os OAs se encontram com todas as informações necessárias ao processo de recomendação automática de conteúdos relacionados. A Figura 24 ilustra parte do arquivo gerado para verificação das associações anotadas nos respectivos OAs.

Figura 24 - Parte do arquivo de registros das associações geradas

```
Building terms ontological association to document:
ComposerBiography_Antonin_Dvorak.html_00021
Processing 68 annotated relevant terms.
-----------------------------------------------------
-->
Relevant term: Dvorak
Ontological class: Composer
Ontological super class: ClassicalMusic
Ontological sub classes: don't have.
Searching relations in the ontology:

Relation name: references
Domain class: Composer
Range class: GlossaryTerm

Relation name: isReferencedBy
Domain class: Composer
Range class: HistoricalPeriod
```

Fonte: Elaborada pelo autor

Após a geração das devidas associações para cada OA, conclui-se a etapa de recuperação de informações relevantes. Assim, os OAs podem, então, ser finalmente persistidos no repositório de conteúdos, juntamente com seus metadados, permanecendo à disposição de autores de conteúdos que queiram utilizá-los para composição de uma unidade de aprendizagem.

7.4 Recomendação e agregação de OAs relacionados

Uma vez que se tenha selecionado um conjunto de OAs a partir do repositório, para composição de um dado conteúdo didático-pedagógico, estes podem ser submetidos ao processo de recomendação automática de conteúdos relacionados. Para isso, é utilizado o módulo do sistema denominado *RecommendationsBuilder,* que foi implementado de modo a receber como entrada uma lista de OAs e fornecer como saída outra lista, contendo um conjunto de OAs recomendados como conteúdos relacionados aos OAs da lista de entrada. Este processo consiste, então, na geração de um conjunto de recomendações, para cada OA da

lista de entrada, de modo que cada recomendação aponte outro OA presente no repositório e indique o tipo de relação que estabelece com o OA ao qual está sendo recomendado.

Iterando sobre a lista de entrada, cada um dos OAs listados é recuperado a partir do repositório. Dentre os metadados anotados em cada OA, nas etapas anteriores, se encontram as diversas associações geradas a partir de seus termos e conceitos mais relevantes, com base na estrutura da ontologia de domínio. Assim, para cada associação encontrada, as relações que a compõem são analisadas e as classes de termos para as quais apontam como seu alcance são identificadas. Dessa forma, têm-se, através destas relações, arcos que conectam o documento a diversas outras classes de conceitos, a partir de cada um de seus termos mais relevantes. Assim, para cada relação presente, em cada uma das associações geradas, para cada um dos termos mais relevantes, em cada um dos OAs da lista de entrada, é realizada uma busca no repositório de conteúdos por outros OAs cujos termos mais relevantes pertençam à classe de alcance do termo através da relação em análise. Em cada OA encontrado seu VT é analisado. Caso contenha o termo fonte da associação em análise, este OA é então recomendado como conteúdo relacionado ao OA principal e o tipo de relação é qualificada como sendo do tipo descrito na associação do termo fonte. A Figura 25 ilustra parte do arquivo que registra as recomendações geradas para os respectivos OAs.

Figura 25 - Parte do arquivo de registros das recomendações geradas

```
Building recommendations to document:
HistoricalPeriod_SUMMARY_OF_WESTERN_CLASSICAL_MUSIC_HISTORY.htm_00055
in corpus html0aCorpus
Processing 363 ontological associations
--------------------------------------------------------------------------------
-->
Relevant term:Ludwig
Relevante term class: Composer

Searching for hasPart relation in range class = ComposerBiography
Relation found:
HistoricalPeriod_SUMMARY_OF_WESTERN_CLASSICAL_MUSIC_HISTORY.htm_00055 term Ludwig hasPart
ComposerBiography_Ludwig_Minkus.html_0003C

Searching for hasPart relation in range class = MusicalWork
```

Fonte: Elaborada pelo autor

No intuito de se refinar as recomendações geradas e evitar um número alto de recomendações irrelevantes, mesmo que estas tenham sido construídas por todos os processos anteriormente descritos, é realizada uma última etapa de processamento, para então retornar ao usuário a lista de OAs recomendados. Trata-se de um processo de *ranking* dos documentos recomendados para cada uma das associações.

As recomendações previamente geradas por vezes trazem mais de um documento recomendado, para o mesmo tipo de relação, a partir do termo relevante contido em uma mesma associação. Sendo assim, é importante que se determine quais destes documentos são os mais recomendados. Este processo é realizado pelo módulo do sistema denominado *DocScoreRecommendationsBuilder*. Este módulo recebe como entrada as recomendações para cada OA em análise, gerada pelo *RecommendationsBuilder*. A partir destas recomendações, os documentos apontados como relacionados a cada termo de uma recomendação recebem, então, um *score*, que indica dentre estes documentos, para quais deles o termo fonte da relação estabelecida é mais relevante.

É importante que se perceba que, nesta fase, os termos presentes nas recomendações já foram ponderados quanto à sua relevância para cada documento, por meio do módulo *AssignerRelevance* e, por isso, foram considerados para a construção de associações e recomendações. Sendo assim, o processo realizado pelo *DocScoreRecommendationsBuilder* não consiste em atribuir nova relevância ao conceito, mas sim em agrupar os diferentes documentos apontados como relacionados ao OA em análise a partir daquele termo, que consta como relevante para todos eles, e dizer em qual deles o conceito é mais relevante. Para isso, este módulo do sistema utiliza um *plugin* do GATE denominado *SearchPR*, que recebe um termo e uma coleção de documentos e retorna, para cada documento, um *score* que indica o quanto aquele documento é importante, considerando o termo dado como entrada.

Dessa forma, a lista de recomendações retornada ao usuário é capaz de apontar, para cada relação estabelecida, o OA recomendado que está mais fortemente relacionado ao OA para o qual foram geradas a recomendações.

8 TESTES E RESULTADOS ALCANÇADOS

Como parte deste trabalho foi organizado um repositório de OAs, composto de 8.967 documentos, cujos conteúdos estão compreendidos dentro do domínio da música erudita. Para a realização dos testes e avaliação dos resultados, foi necessária a execução de uma etapa de anotação manual de metadados e indicação de termos relevantes, como será descrito a seguir. Sendo assim, devido à dificuldade inerente à realização manual destes processos, uma amostragem foi gerada a partir do total de OAs presentes no repositório, mantendo-se a mesma proporção no percentual de documentos para cada categoria. Os documentos foram selecionados de forma automática e aleatoriamente, conforme a Tabela 1. Dentre os 111 documentos obtidos, 10 se limitavam a apresentar o nome de um compositor e a discografia com suas obras. Estes não foram analisados, resultando ao final em 101 documentos.

Tabela 1 - Composição da amostragem inicial

Classificação	Nº Documentos	% Amostragem
Biografias de Compositores	37	33,3
Períodos Históricos	2	1,8
Termos de Glossário	14	12,6
Obras Musicais	58	52,3
Total de documentos:	111	100,0

Fonte: Elaborada pelo autor

Quatro profissionais da Educação aceitaram o convite para contribuir com esta pesquisa, de forma voluntária e de acordo com os procedimentos descritos no documento a elas encaminhado, denominado "Termo de Consentimento Livre e Esclarecimento", devidamente submetido ao Comitê de Ética em Pesquisa da universidade. A cada colaboradora foi solicitado, após devidamente acordado, que realizasse a anotação manual de todos os 101 documentos da amostragem final, indicando também os termos que julgassem mais relevantes, dentre os termos anotados, a cada documento, a partir dos quais deveriam ser recomendados outros documentos que estivessem a eles relacionados, no intuito de estender ou complementar seu conteúdo. Apenas três colaboradoras concluíram a atividade dentro do prazo previsto para trinta dias, resultando em 303 documentos anotados. Foram utilizados para análise os dois resultados que apresentaram maior número de anotações, cujas colaboradoras serão identificadas como A e B. Cada documento foi analisado e teve as anotações manualmente computadas. Os termos indicados como mais relevantes, por cada colaboradora, foram por elas listados em uma planilha, para cada documento.

Os resultados obtidos pelas colaboradoras A e B são apresentados e comparados na Tabela 2. Nela estão presentes a contagem total de termos manualmente anotados sobre a amostragem, por ambas colaboradoras, quantos destes são relevantes ao domínio e qual foi a acurácia total alcançada no processo de anotação manual. Também são apresentados os resultados gerados por cada uma das colaboradoras individualmente, obtendo-se o percentual de acurácia para cada uma delas. As médias de anotações por classe de documento também compõem esta tabela, permitindo uma análise mais refinada do comportamento das colaboradoras quanto ao processo de anotação, como será discutido a seguir. Outra informação importante, que consta na Tabela 2, é a contagem final de termos indicados como mais relevantes sobre a amostragem, tanto por cada uma das colaboradoras quanto pela soma dos resultados. A importância desta informação reside no fato de que a base para as relações entre os documentos, segundo a metodologia proposta, se encontra no vetor de termos relevantes gerado para cada documento.

Foram considerados não relevantes os termos e conceitos anotados que não se referiam ao domínio da música erudita, tendo como referência principal a ontologia de domínio, e aqueles que se referiam a informações genéricas, tais como “ganhador de quatro prêmios”, “nascido em Paris”, dentre outros. Essas anotações foram recorrentes no caso da Colaboradora A. Estes termos também não foram computados na contagem dos termos indicados como mais relevantes para cada documento.

Tabela 2 - Resultados da anotação manual

Anotações Sobre a Amostragem			**Col. A**		**Col. B**		**Acurácia**	
Total	Relevantes	Acurácia	Total	Rel.	Total	Rel.	**Col. A**	**Col. B**
1892	**1231**	**65,06**	**749**	**492**	**1143**	**739**	**65,69 %**	**64,65 %**
Média de Anot. / documento/ classe	Biog. de Compositores		8,43	4,50	20,07	15,50	53,39 %	77,22 %
	Períodos Históricos		156,0	117,0	123,0	117,0	75,00 %	95,12 %
	Glossário		7,14	5,93	5,50	5,43	83,00%	98,70 %
	Obras Musicais		4,36	2,81	6,46	1,90	64,59 %	29,40 %
Indicação de termos mais relevantes			**287 termos**		**191 termos**		**478 total**	

Fonte: Elaborada pelo autor

A Tabela 3 indica o número de resultados coincidentes, que consiste no conjunto interseção dos termos anotados por ambas as colaboradoras, para cada classe de documentos, assim como dos termos que foram indicados, também por ambas, como mais relevantes sobre a amostragem. Estes dados permitem uma análise das possíveis variações no padrão de anotação adotado por diferentes atores. No caso deste trabalho, entre as colaboradoras A e B.

Tabela 3 - Contagem de resultados coincidentes

Classes de documentos	Anot. coincidentes	Termos mais relevantes
Biografia de Compositores	53	53 coincidências para os termos indicados como mais relevantes.
Períodos Históricos	41	
Glossário	41	
Obras Musicais	13	
Total	**148**	

Fonte: Elaborada pelo autor

Durante a análise do material, percebe-se a dificuldade em se manter a coerência no processo de anotação manual. Ao anotar documentos pertencentes à mesma classe, com o mesmo formato e padrão para disponibilização das informações, ora um determinado conjunto de termos foi marcado como relevante, ora não, pela mesma colaboradora. Por exemplo, nos textos relativos às obras musicais, a Colaboradora A manteve o padrão de anotação, marcando sempre o autor, o tipo de obra, a data e local de apresentação, com pouquíssima variação, não anotando elementos presentes no texto da sinopse, o que talvez possa ser consequência da atividade cansativa e repetitiva que consiste na anotação manual de metadados. A Colaboradora B, por sua vez, ora inseria anotações na sinopse da obra, ora na ficha técnica, ora nos nomes dos personagens das óperas, sendo difícil identificar o critério adotado. Além disso, a quantidade de anotações da Colaboradora B diminuiu consideravelmente entre os primeiros e últimos documentos anotados. Nos documentos maiores, ela indicou como os termos mais relevantes apenas aqueles presentes na primeira página.

Concluído o trabalho por parte das colaboradoras, assim como a análise do material por elas produzido, a mesma amostragem foi submetida ao Sistema de Recomendação e Agregação de Conteúdos Relacionados, executando-se todos os processos, da anotação à recomendação. Os resultados da anotação automática são apresentados na Tabela 4.

Tabela 4 - Resultados da anotação automática

Anotações Sobre a Amostragem		Total	Verdadeiros Positivos	**Acurácia**
	Termos Anotados:	**6228**	**4988**	**80,09 %**
Média de Anotações /documento/ classe	Biografia de Compositores	84,93	71,14	83,77 %
	Períodos Históricos	1213,00	1083,00	89,28 %
	Glossário	23,71	23,07	97,29 %
	Obras Musicais	39,07	26,95	69,98 %

Fonte: Elaborada pelo autor

Percebe-se uma diferença nos parâmetros de análise, entre os resultados automático e manual, sendo que no segundo não há ocorrência de termos não relevantes, devido à consistência da base de conhecimento e ao fato de que apenas termos nela presentes são anotados. Todavia, a atenção, neste caso, se volta para a geração de falsos positivos. Estas ocorrências se devem a problemas difíceis da recuperação da informação e que fogem ao escopo deste trabalho, tais como o tratamento de homônimos e de duplicação, cuja maior ocorrência se deve aos nomes de compositores e termos de ocorrência ampla, facilmente presentes em contextos fora da área de domínio, como nos casos de “time” e “scale”, que constam no glossário como “tempo musical” e “escala musical”, mas foram responsáveis por parte dos falsos positivos. Em um dos documentos, a frase “large-scale composition” teve “scale” anotado como termo relevante ao domínio, pertencente ao glossário musical.

Outro conjunto de falsos positivos foi gerado para compositores cujos primeiros ou segundos nomes são monossilábicos e podem ser interpretados como *stop words* ou pertencentes a outras classes gramaticais que não sejam a dos nomes próprios. Os nomes He, She, An são exemplos de ocorrências como estas. O tratamento desta questão não é trivial. A anotação correta fica, então, a cargo do nome completo. Os OAs da classe *MusicalWork* foram os que mais apresentaram falsos positivos devido à grande presença de nomes de personagens de óperas, que muitas vezes foram anotados como primeiros nomes de compositores. Ao contrário, os OAs pertencentes ao glossário foram os que apresentaram o menor número de falsos positivos, pois consistem de textos com descrições técnicas, cujos termos são específicos e dificilmente incorrem em homônimos ou apresentam duplicação.

No caso específico dos compositores, a divisão do dicionário em uma lista para nomes completos, uma para primeiros nomes e outra para segundos nomes fez com que, em alguns momentos, nomes tais como Michael Tilson Thomas tivessem mais de uma anotação, sendo “Michael” anotado como segundo nome de um compositor e também como primeiro nome, “Tilson Thomas” como segundo nome e “Michael Tilson Thomas” como nome completo de

compositor, o que gerou um falso positivo. Outro exemplo de falso positivo com homônimos ocorreu com Bernard Haitink, que é um maestro e teve Bernard anotado como o primeiro nome de um compositor. O mesmo observado para "York", que é o primeiro nome de um compositor, mas foi assim anotado quando ocorreu em *New York*. *Silent Woods* teve *Woods* marcado como segundo nome de um compositor. Esta é uma questão não trivial.

Outra ocorrência de falsos positivos ligados a nomes de compositores se deve àqueles que possuem segundos nomes iguais aos nomes de países. Na biografia de Kosku Yamada, Berlin aparece tanto como a cidade alemã quanto como o compositor Berlin Musikhochschule e todas às vezes foi marcado como primeiro nome de compositor. Por isso, gerou falsos positivos nos casos em que se tratava da cidade de Berlim. Autores com nomes de meses do ano, como *April*, também geram falsos positivos, pois na língua inglesa os nomes dos meses são grafados com a primeira maiúscula. Todavia, o nome do autor ainda será anotado corretamente quando ocorrer por completo, pelo primeiro nome ou pelo segundo nome, quando diferentes deste caso. Outro caso interessante e não trivial ocorre quando um compositor possui um dos nomes cujo termo pertence a outra classe de termos. Por exemplo, *Ballet*. *Ballet* é um estilo de composição ou trabalho musical, mas há um compositor que possui *Ballet* como segundo nome. Neste caso, o termo recebe as duas anotações, uma delas sendo falso positivo. Casas de espetáculo com nomes de compositores ou personalidades homônimas de compositores também geraram falsos positivos.

Todavia, é interessante atentar para o fato de que estes falsos positivos, em grande parte, foram eliminados ou receberam ponderação muito baixa no processo de classificação hierárquica, conforme os parâmetros da função do cálculo de relevância, descritos na Seção 3.1, não sendo considerados para as etapas de criação de associações e recomendação de conteúdos relacionados. Este fator importante é claramente percebido para os nomes de compositores onde primeiros nomes geraram falsos positivos. Como os compositores são citados na maior parte do texto por seus nomes completos ou por seus segundos nomes, estes obtiveram um fator de relevância muito maior que seus primeiros nomes, fazendo com que estes ficassem, na maioria das vezes, fora dos processos de associação ontológica e recomendação automática que é o objetivo principal deste trabalho e não o processo de anotação.

Ainda assim, para os falsos positivos que cheguem a compor associações e entrem no processo de recomendação, na maioria das vezes os documentos recomendados a partir deles recebem, no passo final, um *score* menor do que aqueles que foram recomendados para verdadeiros positivos. Isso demonstra como a metodologia proposta atua como esperado a

partir da execução do conjunto de seus processos, fornecendo mecanismos capazes de minimizar o numero de falsos positivos na recomendação de conteúdos relacionados. Ainda assim, aqueles falsos positivos que acabem por serem considerados podem impactar ao final de todo o processo, gerando recomendações não relevantes e, por isso, o último filtro para as recomendações geradas, antes da agregação do conteúdo, é realizado manualmente pelo responsável por sua composição.

A utilização de uma única lista com os nomes completos poderia se apresentar como possível solução para o problema das diversas marcações sobre nomes de compositores, tais como Michal Tilson Thomas, mas criaria um problema com falsos negativos, quando os autores fossem referenciados apenas pelo segundo nome, o que é muito comum na literatura. A utilização de duas listas no dicionário, uma com nomes completos e outra com segundos nomes, além da resolução de homônimos e duplicação, talvez seja uma solução plausível.

Para o problema dos compositores com nomes de países, utilizar-se de uma solução trivial, tal como inserir uma listagem de países, pareceria algo plausível à primeira vista, mas poderia gerar um grande número de marcações não relevantes, número maior do que os falsos positivos gerados pelos nomes dos compositores, entendendo-se que “Países” não consiste em uma classe de termos relevantes ao domínio da música. Seriam geradas duas anotações para o nome tendo-se, assim, um verdadeiro positivo e um falso positivo, além de suas anotações para o caso de um país, tendo-se novamente um verdadeiro positivo e um falso positivo. Uma listagem de nomes de meses, para os autores que os têm em seus primeiros nomes, consiste na mesma questão levantada para o nome dos países.

No caso de homônimos entre compositores, uma possível abordagem seria subdividir a classificação de compositores, agrupando-os por períodos históricos ou gêneros musicais e criando-se heurísticas com base nessa relação. De todo modo, a solução para homônimo é um caso não trivial. Para recomendação de conteúdos relacionados, por exemplo, identificar exatamente a qual dos compositores de mesmo nome um dado texto se refere talvez exija a análise de uma combinação de outros elementos do texto, para se tentar associar, por exemplo, o nome do compositor ao período histórico ou às obras musicais, aos quais o texto se refira. Todavia, a atribuição de um *score* aos documentos recomendados, conforme feito neste trabalho e descrito na Seção 8.3, se apresenta como um recurso eficiente para reduzir a possibilidade de que o autor de conteúdos selecione, ao final, um texto equivocado quanto à existência de uma relação com o conteúdo principal, ao qual está sendo recomendado.

Durante a análise do processo de anotação manual, identificou-se que locais importantes, como casas de ópera, teatros e renomadas escolas de música foram anotados de

forma recorrente pelas colaboradoras, mas não estão representados na ontologia. Da mesma forma, percebe-se que nos textos, muitos termos e conceitos relevantes ao domínio não são anotados por estarem em outro idioma, tais como francês, alemão e italiano, como é comum no domínio da música. Nestes casos, têm-se sempre falsos negativos para o processo de anotação automática, o que incorrerá, no caso da metodologia proposta, na falta de recomendação para tais termos.

Os termos que são iguais, mas que são identificados por anotações diferentes, não têm suas frequências absolutas somadas, na função de relevância proposta, sendo interpretados individualmente, como estão anotados. Por exemplo, “ERNESTO LECUONA”, “Ernesto Lecuona”, “Ernesto” e “Lecuona”, foram anotados corretamente no documento sobre este autor, mas como termos independentes e assim foram ponderados. Se sua ponderação fosse somada e eles aparecessem como um único termo no vetor de relevância para o documento, provavelmente traria associado ao nome do autor um valor muito maior de relevância. Todavia, percebe-se que mesmo anotados separadamente, os termos aparecem nos primeiros lugares na classificação hierárquica por relevância, sendo “Lecuona” com relevância igual a 6.0, “Ernesto” com 4.5 e “Ernesto Lecuona” com 3.0. Isso garante que o autor seja considerado em primeiro lugar na fase de recomendação de conteúdos.

Como descrito na Seção 3.1, a fase de associação não processa todos os termos anotados, mas apenas os de maior relevância em cada documento. Sobre a amostragem utilizada foram geradas 3508 associações ontológicas sobre os 101 OAs. Para testar especificamente o processo de recomendação de conteúdos relacionados, foi dado como entrada o OA da amostragem que contém o maior conteúdo, pertence à classe Períodos Históricos e totaliza 363 associações. As relações estabelecidas por estas associações foram analisadas pelo sistema e geraram um conjunto de 12 recomendações finais para este documento sobre os 101 OAs contidos na amostragem. Estas recomendações estão divididas da seguinte forma: um OA como *isrequiredby*; sete como *isreferencedby* e quatro como *haspart*. As 12 recomendações geradas automaticamente estavam corretas e são apresentadas no Quadro 2. A Figura 26 ilustra uma das recomendações.

Para realização dos testes, foi utilizada uma máquina com processador *Intel Core* I3 com 2.27 GHz por núcleo, 4 GB de RAM e utilizando o *Microsoft Windows* 7. Os processos de carregamento, armazenamento, anotação automática e geração de associações sobre a ontologia, para toda a amostragem, foram executados em 1 minuto, 43 segundos e 2 décimos de segundo. O processo de geração de recomendações foi executado em 5 décimos de segundo.

Quadro 2 – Resultado da recomendação automática

HistoricalPeriod_SUMMARY_OF_WESTERN_CLASSICAL_MUSIC_HISTORY.htm	
isrequiredby	Music_Theor_Online_Music_of_the_20th_Century.htm
isreferencedby	ComposerBiography_Victor_Herbert.html, ComposerBiography_Colin_Matthews.html, ComposerBiography_Antonin_Dvorak.html, ComposerBiography_Henry_Purcell.html, ComposerBiography_Erroll_Garner.html, ComposerBiography_William_Byrd.htm, ComposerBiography_Ludwig_Minkus.html.
haspart	GlossaryTermDefinition_Recorder.htm, Charles_Wakefield_Cadman.htm, Gustav_Mahler.html, Charles_Wakefield_Cadman.htm

Fonte: Elaborado pelo autor

Figura 26 - Parte do arquivo gerado para as recomendações finais

Relevant term: classical music
Relevant term class: Classical
Relevant term superclass = HistoricalPeriod
HistoricalPeriod_SUMMARY_OF_WESTERN_CLASSICAL_MUSIC_HISTORY.htm_00055 term classical music isRequiredBy
HistoricalPeriod_Music_Theor_Online_Music_of_the_20th_Century.htm_00054

Fonte: Elaborada pelo autor

9 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve como objetivo a proposição de uma metodologia para recomendação automática de OAs relacionados, em conformidade com o SCORM. Assim, possibilitou o estabelecimento de relações entre OAs, utilizando a categoria de metadados *relation*, tal como definida pelo padrão e sem a necessidade de utilização ou desenvolvimento de SGAs específicos para interpretar tais metadados. Além disso, percebe-se que as diversas abordagens presentes na literatura se pautam em um processo de pesquisa por conteúdos relacionados que é realizado pelo usuário, quando a metodologia proposta neste trabalho adota outra perspectiva, segundo a qual os OAs são recomendados como conteúdos relacionados a um dado conteúdo de referência, previamente selecionado, visando à construção de uma unidade de aprendizagem.

Partindo da hipótese apresentada na Seção 1.3, foi elaborada e construída, como parte deste trabalho, uma base de conhecimento de domínio, contendo um extenso dicionário de termos e uma ontologia capaz de apresentar uma conceituação da área de domínio utilizada e o estabelecimento de relações entre as diversas classes de conceitos. Esta base de conhecimento foi efetivamente empregada pelo protótipo de um Sistema de Recomendação de Conteúdos Relacionados, também desenvolvido nesta pesquisa e que implementa a metodologia proposta. Os seguintes módulos do sistema foram implementados e atenderam a cada uma das etapas da metodologia: *AssignerRelevance*, para anotação automática de metadados e classificação hierárquica dos conceitos relevantes; *AssociationsBuilder*, para geração de associações entre os conceitos relevantes e a ontologia de domínio; *RecommendationsBuilder*, para geração de recomendações de conteúdos relacionados, a partir das associações preestabelecidas; e *DocScoreRecommendationsBuilder*, que atribui um *score* final a cada documento recomendado. O *plugin* ANNIE foi utilizado apenas para o processo de anotação automática de metadados, empregando a base de conhecimento.

Com base na análise dos testes realizados, observa-se que a metodologia proposta nesta pesquisa é viável e produz os resultados esperados, com boa precisão e eficiência, além de superiores àqueles alcançados unicamente por seres humanos. Pode ser aplicada a diferentes áreas do conhecimento, para a composição de conteúdos didático-pedagógicos, sendo necessário apenas o emprego de uma base de conhecimento de domínio relacionada à área desejada, comprovando a veracidade da hipótese anteriormente apresentada. Diante disso, verifica-se que a recomendação automática de OAs relacionados pode auxiliar os desenvolvedores de conteúdo para *e-Learning* na composição de OAs em conformidade com

o SCORM, reduzindo o tempo e o esforço necessários ao desenvolvimento e agregação de conteúdos relacionados e facilitando o seu reaproveitamento.

Há diversas possibilidades de ampliação desta pesquisa e melhorias nos resultados obtidos. Devido ao fato de que a metodologia proposta é fortemente dependente da base de conhecimento, falsos negativos podem ocorrer para os termos e conceitos que não estão nela presentes. Reduzir o número de falsos positivos no processo de geração e anotação de metadados é de extrema importância, pois estes podem impactar no final do processo, gerando falsos positivos também nos documentos recomendados, pois falsos positivos que passem à fase de geração de associações podem influenciar a recomendação. Isso demonstra que a acurácia do processo de recomendação está relacionada à acurácia da fase de anotação e classificação hierárquica. Possíveis abordagens para redução destas ocorrências consistem na resolução dos problemas de recuperação da informação, anteriormente levantados, e na avaliação de diferentes métricas na função de cálculo de relevância, sendo que a primeira abordagem procuraria reduzir diretamente o número de anotações sobre falsos positivos e a segunda procuraria reduzir o número de falsos positivos anotados que passariam à fase de geração de associações. A identificação de novas classes na ontologia pode ampliar a cobertura para a recomendação de conteúdos relacionados sob a área de domínio.

Desse modo, este trabalho de pesquisa cumpriu com os objetivos propostos, apresentando uma metodologia eficiente para recomendação de OAs relacionados, em conformidade com o SCORM. O sistema implementado traduz a metodologia e apresenta bons resultados quanto aos processos de anotação automática de metadados, associação ontológica para identificação de relações e recomendação automática de conteúdos relacionados. Contudo, há muito que se pesquisar e avançar na melhoria destes processos, seja por meio da experimentação e comparação de diferentes métricas para o cálculo de relevância; seja revisando, ampliando e modificando a ontologia de domínio ou acrescentando-se o tratamento para múltiplos idiomas ou aplicando-se estratégias para redução de falsos positivos e resolução de homônimos; seja pela proposição de melhorias na própria metodologia.

A constatação mais importante é que a busca por meios de auxiliar os desenvolvedores de conteúdos, melhorar os processos inerentes à construção de materiais didático-pedagógicos e contribuir para com o desenvolvimento e acesso à educação são fatores para os quais a computação tem muito a contribuir.

REFERÊNCIAS

ADVANCED DISTRIBUTED LEARNING. **The Advanced Distributed Learning (ADL) Initiative**: history. 1999. Disponível em: < http://www.adlnet.gov/overview/>. Acesso em: 11 jul. 2013.

ADVANCED DISTRIBUTED LEARNING. **SCORM 2004:** content aggregation model [CAM]. 4. ed. ADL, 2009a. Disponível em: <http://www.adlnet.gov/scorm/_scorm-2004-4th/>. Acesso em: 11 jul. 2013.

ADVANCED DISTRIBUTED LEARNING**. SCORM 2004:** run-time environment [RTE]. 4. ed. ADL, 2009b. Disponível em: <http://www.adlnet.gov/scorm/scorm-2004-4th/>. Acesso em: 11 jul. 2013.

ADVANCED DISTRIBUTED LEARNING. **SCORM 2004:** sequencing and navigation [SN]. 4. ed. ADL, 2009c. Disponível em: <http://www.adlnet.gov/scorm/_scorm-2004-4th/>. Acesso em: 11 jul. 2013.

ARAÚJO, Moysés; FERREIRA, Maria Alice G. V. **Educação a distância e web semântica**: modelagem ontológica de materiais e objetos de aprendizagem para a plataforma CoL. São Paulo: Universidade de São Paulo, 2003. Disponível em: <http://www.pcs.usp.br/~interlab/ artigoWebSemantica4.pdf >. Acesso em: 30 out. 2013.

BAGHELA, Vishwadeepak Singh; TRIPATHI, S. P. Text mining approaches to extract interesting association rules from text documents. **International Journal of Computer Science Issues**, v. 9, n. 3, p. 545-552, may 2012. Disponível em: <http://www.ijcsi.org/ papers/IJCSI-9-3-3-545-552.pdf >. Acesso em: 11 mar. 2014.

BORGES, Vanessa Araujo; BARBOSA, Ellen Francine. Using ontologies for modeling educational content. In: 7th INTERNATIONAL WORKSHOP ON ONTOLOGIES AND SEMANTIC WEB FOR E-LEARNING, 9, 2009**,** Brighton. **Proceedings…** Disponível em: < http://www.inct-sec.org/actrep/sites/default/files/highlights/SWEL-Ellen.pdf >. Acesso em: 30 de out. 2014.

BHOWMICK, Plaban Kumar et al. A framework for manual ontology engineering for management of learning material repository. **International Journal of Computer Science and Applications**, v. 7, n. 2, p. 30-51, 2010. Disponível em: <http://www. tmrfindia.org/ijcsa/v7i23.pdf >. Acesso em: 21 de jan. 2014.

BONTCHEVA, Kalina et al. 2004. Evolving GATE to meet new challenges in language engineering. **Natural Language Engineering**, v. 10, n. 3-4, p. 349-373, Sept. 2004. Disponível em: <http://journals.cambridge.org/action/displayAbstract?fromPage=online &aid=252241&fileId=S1351324904003468>. Acesso em: 24 nov. 2013.

CHEN,Yen-Liang; LIU, Yi-Hung; HO, Wu-Liang. Text mining approaches to extracts interesting association rules from text documents. **Journal of the American Society for Information Science and Technology**, v. 64, n. 2, p. 280-290, jan. 2013. Disponível em <http://www.researchgate.net/publication/251231645_A_Text_Mining_Approach_to_Assist_ the_General_Public_in_the_Retrieval_of_Legal_Documents >. Acesso em: 12 fev. 2014.

CORDEIRO, A. D. **Gerador Inteligente de Sistemas com Auto-aprendizagem para**

Gestão de Informações e Conhecimento. 2005. Tese (doutorado) - Universidade Federal de Santa Catarina, Departamento de Engenharia da Produção, Santa Catarina.

CUNNINGHAM, Hamish et al. **Developing language processing components with GATE version 7 (a user guide).** Sheffield: University of Sheffield Department of Computer Science, 2012. Disponível em: <https://gate.ac.uk/sale/tao/tao.pdf>. Acesso em: 11 jul. 2013.

EBECKEN, N. F. F.; LOPES, M. C. S.; COSTA, M. C. A. Mineração de Textos. In: REZENDE, S. O. (Org.). **Sistemas Inteligentes**: fundamentos e aplicações. Barueri: Manole, 2003. Cap. 13, p. 337 – 370.

EDVARDSEN, Lars F.H et al. Using automatic metadata generation to reduce the knowledge and time requirements for making SCORM learning objects. In: IEEE INTERNATIONAL CONFERENCE ON DIGITAL ECOSYSTEMS AND TECHNOLOGIES, 3, 2009, Istanbul. **Proceedings…** New York: IEEE, 2009. p.253-258.

ENGELHARDT, Michael et al. Reasoning about elearning multimedia objects**.** In: INTERNATIONAL WORKSHOP ON SEMANTIC WEB ANNOTATIONS FOR MULTIMEDIA, 2006, Edinburgh. **Proceedings…** Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.120.8484&rank=1>. Acesso em: 9 nov. 2013.

FOREST, Dominic; SYLVA, Lyne Da. Text mining and information retrieval. **Canadian journal of information and library science**, v. 35, n. 3, p. 217-227, set. 2011. Disponível em: <http://muse-jhu-edu.ez93.periodicos.capes.gov.br/journals/canadian_journal_of_ information_and_library_science/v035/35.3.article.html>. Acesso em: 20 ago. 2014.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002. 175 p.

GUO, Zhixin; JIN, Hai. A rule-based framework of metadata extraction from scientific papers. In: INTERNATIONAL SYMPOSIUM ON DISTRIBUTED COMPUTING AND APPLICATIONS TO BUSINESS, ENGINEERING AND SCIENCE, 10, 2011a, Wuxi. **Proceedings…** New York: IEEE, 2011a. p. 400-404.

GUO, Zhixin; JIN, Hai. Reference metadata extraction from scientific papers: cluster and grid computing lab services computing technology and system lab Huazhong University of Science and Technology. In: INTERNATIONAL CONFERENCE ON PARALLEL AND DISTRIBUTED COMPUTING, APPLICATIONS AND TECHNOLOGIES, 12, 2011b, Gwangju. **Proceedings…** New York: IEEE, 2011b. p. 45-49.

HERNÁNEZ, Alvaro et al. Convirtiendo el contenido de archivos en objetos de aprendizaje. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 20, 2009, Florianópolis. **Anais...** Florianópolis: UFRGS, 2009. Disponível em: <http://www.niee.ufrgs. br/eventos/SBIE/2009/conteudo/artigos/completos/62173_1.pdf>. Acesso em: 8 set. 2014.

HUYNH Tin; HOANG, Kiem. GATE framework based metadata extraction from scientific papers. In: INTERNATIONAL CONFERENCE ON EDUCATION AND MANAGEMENT TECHNOLOGY, 2010, Cairo. **Proceedings…** New York: IEEE, 2010. p. 188-191.

LIPINSKI, Mario et al. Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL

LIBRARIES, 13, 2013, New York. **Proceedings…** New York: CM/IEEE, 2013. p. 385-386.

LU, Eric Jui-Lin et al. Extended relation metadata for SCORM-based learning content management systems. **Educational Technology & Society**, v.13, n. 1, p. 220-235, Jan. 2010. Disponível em: <http://www.ifets.info/journals/13_1/21.pdf >. Acesso em: 21 abr. 2013.

LU, Eric Jui-Lin; HSIEH, Chin-Ju. A relation metadata extension for SCORM content aggregation model. **Computer Standards & Interfaces**, v.31, n. 5, p.1028-1035, Sept. 2009.

MARATEA, Antonio; PETROSINO, Alfredo; MANZO, Mario. Automatic generation of SCORM compliant metadata for portable document format files. In: INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND TECHNOLOGIES, 13, 2012, Ruse. **Proceedings...** New York: ACM, 2012. p. 360-367.

MARGARITOPOLOUS, Merkourios; MANITSARIS, Athanasios; MAVRIDIS, Ioannis. **On the identification of inference rules for automatic metadata generation**. 2007. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.1723&rep=rep1&type=pdf>. Acesso em: 24 out. 2013.

MAYNARD, Diana. Benchmarking ontology-based annotation tools for the semantic web. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 6, 2008, Marrakech. **Proceedings...** Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.161.1242>. Acesso em: 9 nov. 2013.

MORAIS, Edison Andrade Marins; AMBRÓSIO, Ana Paula L. **Mineração de textos**. Goiânia: Instituto de Informática, Universidade Federal de Goiás, 2007. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>. Acesso em: 22 jan. 2014.

NAUERZ, Andreas et al. Personalized recommendation of related content based on automatic metadata extraction. In: CONFERENCE OF THE CENTER FOR ADVANCED STUDIES ON COLLABORATIVE RESEARCH: MEETING OF MINDS, 2008, Ontario. **Proceedings…** New York: ACM, 2008.

REDONDO, Rebeca P. Díaz; VILAS, Ana Fernández; ARIAS, Jose J. Pazos. Educateca: A Web 2.0 approach to e-Learning with SCORM. **Software Services for e-World –IFIP. Intelligent Systems Reference Library**, v. 32, p. 195-207, 2012. Disponível em: <http://link.springer.com/chapter/10.1007%2F978-3-642-25694-3_10>. Acesso em: 28 abr. 2013.

REY-LÓPEZ, Marta et al. An extension to the ADL SCORM standard to support adaptivity: The t-learning case-study. **Computer Standards & Interfaces**, v. 31, n. 2, p. 309-318, Feb. 2009.

ROY, Devshri; SUDESHNA Sarkar; SUJOY Ghose. Automatic extraction of pedagogic metadata from learning content. **International Journal of Artificial Intelligence in Education**, v. 18, n. 2, p. 97-118, Apr. 2008.

SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. **Os grandes desafios da computação no Brasil: 2006 – 2016.** São Paulo: SBC, 2006. Disponível em: <http://www.sbc.org.br/index.php?option=com_jdownloads&Itemid=195&task=finish&cid=11&catid=50>. Acesso em: 11 jul. 2013.

SU, Jun-Ming. et al. Constructing SCORM compliant course based on high-level petri nets. **Computer Standards & Interfaces,** v. 28, n. 3, p. 336-355, Jan. 2006.

TUAROB, Suppawong; POUCHARD ,Line C.; GILES, C. Lee. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In**:** ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 13, 2013, New York. **Proceedings…** New York: ACM, 2013. p. 239-248.

WAZLAWICK, Raul Sidnei. **Metodologia de pesquisa para ciência da computação.** 1. ed. Rio de Janeiro: Elsevier, 2008. 159 p.

CERVO, Amado Luiz; BERVIAN, Pedro Alcino. **Metodologia científica.** 5. ed. São Paulo: Prentice Hall, 2002. 242 p.

YUAN, Xiaojun; BELKIN, Nicholas J. Investigating information retrieval support techniques for different information-seeking strategies. **Journal of the American Society for Information Science and Technology**, v. 61, n. 8, p. 1543-1563, Apr. 2010. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/asi.21314/pdf>. Acesso em: 20 ago. 2014.

APÊNDICE A – TERMO DE CONCENTIMENTO LIVRE E ESCLARECIMENTO.

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Título do projeto: Recomendação e Agregação de Conteúdos Relacionados em Conformidade com o Padrão SCORM.

Prezado Sr (a) ,

Você está sendo convidado (a) a participar de um projeto de pesquisa na área de informática na educação, que busca propor e desenvolver um sistema computacional para recomendação automática de conteúdos didático-pedagógicos relacionados entre si. As relações que buscamos estabelecer, através do sistema proposto, se dão no âmbito de conteúdos complementares que podem ser agregados a um dado conteúdo principal. Por exemplo, dado um documento cujo conteúdo fale sobre a história da música clássica, nosso sistema de recomendação precisa identificar, entre centenas de outros documentos, presentes em um repositório de conteúdos, aqueles que o requerem, para serem melhor compreendidos, e quais ele próprio requer, para ser também melhor compreendido. Assim, poderemos compor uma relação onde os documentos se complementem da seguinte forma: documento 1 é requerido pelo documento 2 e documento 1 requer o documento 3.

Você foi selecionado(a) por sua formação e ampla experiência como profissional da Educação. A sua participação nesse projeto consiste em ajudar na produção de um material que servirá como parâmetro para que possamos, posteriormente, mensurar a eficiência obtida pelo sistema que estamos desenvolvendo. Sendo assim, iremos selecionar uma amostragem de documentos, entre uma série de documentos didático-pedagógicos, e você deverá identificar aqueles cujos conteúdos estão relacionados entre si. Esta amostragem será depois submetida ao sistema computacional, para verificarmos se ele consegue produzir um resultado parecido àquele produzido por você, que é o nosso gabarito. Para que você possa relacionar os documentos, são necessários três passos principais:

1) Primeiro você deverá identificar, em cada documento, uma lista de termos e conceitos que são os mais relevantes ao conteúdo presente no documento. De imediato, se encontram entre estes alguns termos do título, do resumo e das palavras chave, quando houver. Além destes, há também aqueles presentes no corpo do texto. Todos os termos

e conceitos deverão ser registrados em uma planilha simples do Excel, que lhe será fornecida no formato que melhor lhe convier, impresso ou eletrônico.

2) Depois de identificar e listar os termos e conceitos mais relevantes, em cada documento, você precisará enumerá-los em ordem decrescente de relevância, ou seja, os mesmos deverão ser listados do mais relevante para o menos relevante.

3) Após a realização dos passos 1 e 2, para todos os documentos que você analisar, o último passo consiste em indicar, para cada um destes documentos, aqueles que estão relacionados a ele, apontando outros documentos que dele necessitem e outros que ele próprio necessite para ser compreendido. Os documentos apontados como relacionados deverão ser listados em campo próprio, na mesma planilha utilizada para listagem dos termos e conceitos relevantes para o documento.

Contaremos com um prazo de 20 dias para realização desta etapa. Não será necessária a sua presença em um local específico nem será estabelecido qualquer horário para realização de suas atividades. A partir de um mínimo de 10 documentos, você irá determinar a quantidade máxima que irá compor sua amostragem, de acordo com sua disponibilidade de tempo para contribuição com esta pesquisa.

Teremos um único encontro presencial para repasse do material necessário e esclarecimento de dúvidas que por ventura você tenha, com relação às tarefas a serem realizadas. O local e horário deste encontro serão definidos de acordo com a sua disponibilidade. Feito isso, toda a comunicação passará a ser realizada por telefone ou por e-mail.

Sua participação é muito importante e voluntária. Você não terá nenhum gasto e também não receberá nenhum pagamento por participar desse estudo. Os materiais necessários serão compostos de papel, caneta, lápis e borracha, caso opte por utilizar o formato impresso da planilha. Neste caso, todo o material lhe será fornecido. Caso opte pelo formato eletrônico e ainda necessite, você pode solicitar o material de papelaria.

As informações obtidas nesse estudo serão confidenciais, sendo assegurado o sigilo sobre sua participação, quando da apresentação dos resultados em publicação científica ou educativa, uma vez que os resultados serão sempre apresentados como retrato de um grupo e não de uma pessoa. Caso haja consentimento de sua parte e sim assim desejar, seu nome poderá ser citado em seções de agradecimentos, onde couber, de forma genérica, sem qualquer referência às atividades realizadas especificamente por você. Você poderá se recusar

a participar ou a realizar qualquer uma das atividades previstas, a qualquer momento, não havendo nenhum prejuízo pessoal se esta for a sua decisão.

Você receberá uma cópia deste termo onde consta o telefone e o endereço do pesquisador responsável, podendo tirar suas dúvidas sobre o projeto e sua participação, agora ou a qualquer momento.