UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Juarez Angelo Piazza Sacenti

SEMÂNTICA EM ANOTAÇÕES DE DADOS GEOGRÁFICOS COLETADAS EM SISTEMAS COLABORATIVOS DA WEB

Florianópolis

2013

Juarez Angelo Piazza Sacenti

SEMÂNTICA EM ANOTAÇÕES DE DADOS GEOGRÁFICOS COLETADAS EM SISTEMAS COLABORATIVOS DA WEB

Trabalho de Conclusão de Curso submetido ao Curso de Ciências da Computação para a obtenção do Grau de Bacharel em Ciências da Computação.
Orientador: Prof. Dr. Renato Fileto
Coorientador: Me. André Salvaro Furtado

Florianópolis

2013

Juarez Angelo Piazza Sacenti

SEMÂNTICA EM ANOTAÇÕES DE DADOS GEOGRÁFICOS COLETADAS EM SISTEMAS COLABORATIVOS DA WEB

Este Trabalho de Conclusão de Curso foi julgado aprovado para a obtenção do Título de "Bacharel em Ciências da Computação", e aprovado em sua forma final pelo Curso de Ciências da Computação.

Florianópolis, 06 de Novembro 2013.

Prof. Dr. Renato Cislaghi
Coordenador

Banca Examinadora:

Prof. Dr. Renato Fileto
Orientador

Me. André Salvaro Furtado
Coorientador

Prof. Dr. Ronaldo dos Santos Mello

Profa. Dra. Vânia Bogorny

A Erna Bettini Sacenti.

AGRADECIMENTOS

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e ao Programa Institucional de Bolsas de Iniciação Científica da UFSC (PIBIC/CNPq - BIP/UFSC), pelo apoio ao projeto de iniciação científica entitulado "Semântica em Anotações de Dados Geográficos coletada em sistemas colaborativos da web".

Ao instituto *Marie Curie International Research Staff Exchange Scheme Fellowship* e ao programa *7th European Community Framework Programme*, por suportarem o projeto *Semantic Enrichment of trajectory Knowledge discovery* (SEEK), inspirador deste trabalho e parceiro em trabalhos futuros.

Ao projeto Wikimapia, por disponibilizar os dados geográficos anotados utilizados por este projeto.

Ao orientador, Dr. Renato Fileto, por acreditar em meu potêncial do início ao fim deste trabalho e incentivar ao máximo meu desenvolvimento e amadurecimento como pesquisador.

Ao coorientador, André Salvaro, pelo suporte teórico e paciência.

Ao professor Dr. João Candido Lima Dovicchi e a Edilson Prudêncio, por apresentar a área de dados geográficos.

A Doroti Sacenti e Gislaine Pigmentel pelo carinhoso auxílio na revisão do texto.

A Odete Maria de Oliveira, pelo incentivo e guia nos caminhos acadêmicos.

Aos meus pais, Mirian Célia Piazza Sacenti e Juarez Angelo Sacenti, por tornarem esta conquista possível.

A Willian Ventura Koerich, pelo auxílio no desenvolvimento e parceria em trabalhos futuros.

Ao *Laboratory for Systems Integration and Advanced Applications* (LISA) e a seus frequentadores que tornaram este projeto mais descontraído e fácil.

Ao Grupo de Banco de Dados da UFSC (GBD-UFSC) que proporcionou maturidade e perspectiva de pesquisa mais ampla.

A equipe que auxiliou e continua auxiliando na preparação para a missão futura.

Enfim, a todos que direta ou indiretamente se envolveram na execução deste projeto, por acreditarem na formação deste pesquisador.

RESUMO

A coleta de dados geográficos utilizando satélites e dispositivos com sistema de posicionamento global (GPS) é insuficiente para levantar informações de como o homem interpreta o espaço. Diversas abordagens têm sido propostas para suprir esta lacuna, incluindo padrões de anotação e coleta colaborativa (*crowdsourcing*) de informação geográfica. Contudo, os dados coletados desta forma não apresentam semântica suficientemente formal para ser utilizada computacionalmente e suficientemente detalhada para suprir as necessidades das aplicações. As soluções da *web* semântica, como as anotações semânticas, ontologias e dados ligados (*linked data*), podem contribuir para resolver os problemas de ambiguidade e interoperabilidade das anotações colaborativas. Este trabalho pretende publicar dados do projeto Wikimapia, um sistema colaborativo ainda não conectado a *web* de dados, segundo o conceito dos dados ligados. Análises comparativas dos sistemas colaborativos Wikimapia e *OpenStreetMap* (OSM), e dos dados desses sistemas, são realizadas visando justificar a contribuição dos dados coletados no Wikimapia para os dados ligados do projeto OSM, o *Linked Geo Data* (LGD). As análises dos dados são restritas a um estudo de caso da cidade italiana de Milão, as palavras das categorias dos elementos espaciais do Wikimapia são confrontadas lexicamente com os nomes das entidades e suas possíveis características, definidas em ontologias e conjuntos de dados ligados a elas associados, tais como GeoNames, DBpedia e LGD.

Palavras-chave: anotações colaborativas, anotações semânticas, dados geográficos, *linked data*, ontologias.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 INTRODUÇÃO

A informação de caráter espacial é relevante para diversas aplicações. Cerca de 80% dos dados estão relacionados com alguma informação espacial (MALINOWSKI; ZIMNYI, 2008). O dado geográfico é utilizado para representar computacionalmente informações espaciais sobre a crosta da Terra, *i.e.*, objetos e fenômenos do mundo real associados a características espaciais e cuja localização geográfica é relevante para interpretação. Sistemas de apoio à tomada de decisão em domínios de aplicação como análise de tráfego, expansão de doenças e turismo precisam considerar dados geográficos em suas análises. Entretanto, a coleta de dados geográficos é uma tarefa trabalhosa e cara no desenvolvimento de sistemas de informação geográfica (SIG). Satélites e dispositivos com sistema de posicionamento global (GPS) são utilizados para coletar informações geográficas, porém são insuficientes para levantar informações de como o homem interpreta o espaço.

Diversas abordagens têm sido propostas para suprir esta lacuna, incluindo padrões de anotação (GIL; KOZIEVITCH; TORRES, 2011) e a coleta colaborativa de informação geográfica, ou *volunteered geographic information* (VGI) (GOODCHILD, 2007). Contudo, ao menos parte dos dados coletados atualmente não apresentam semântica suficientemente formal para ser utilizada computacionalmente ou suficientemente detalhada para suprir as necessidades das aplicações.

Os Bancos de Dados Geográficos (BDG) atuais focam principalmente na forma dos elementos espaciais, sem explorar suficientemente a semântica dos dados. A informalidade e a superficialidade dificultam o processo de integração de bases de dados geográficas e a obtenção de informação espacial detalhada com anotações semânticas bem definidas. Tais problemas se agravam quando a coleta e a anotação de dados é realizada de forma colaborativa, com natureza e modelagem de dados não formalmente definidos.

As soluções da *web* semântica, como as anotações semânticas, ontologias e dados ligados (*linked data*), contribuem para resolver os problemas de ambiguidade e interoperabilidade das anotações colaborativas. Existem algumas iniciativas seguindo a abordagem de desenvolver sistemas de coleta colaborativa de dados geográficos utilizando estas soluções, como o OurMap (GONZALEZ et al., 2013) e o YUMA (SIMON; JUNG; HASLHOFER, 2011). Outras iniciativas buscam converter dados colaborativamente coletados em dados com semântica bem definida, segundo o conceito dos dados ligados, como o *Linked Geo Data* (LGD)[1] (STADLER et al., 2012) do projeto *OpenStreetMap* (OSM).

[1] Disponível em: `http://linkedgeodata.org/About`. Acesso em nov. 2013.

Este trabalho segue a segunda iniciativa: define um processo baseado no método utilizado por (STADLER et al., 2012) para publicar dados ligados em um sistema de coleta colaborativa ainda não ligado na *web* de dados, o projeto Wikimapia. O objetivo principal é disponibilizar o conteúdo do Wikimapia segundo o conceito dos dados ligados, correlacionando-o com os dados ligados já existentes. Este trabalho e o método nele desenvolvido devem também servir de referência para futuras publicações. Análises comparativas dos sistemas colaborativos Wikimapia e OSM, e de seus dados, realizadas neste trabalho objetivam justificar a contribuição dos dados coletados no Wikimapia para a área de dados geográficos ligados, complementando coleções como LGD e GeoNames[2].

A análise de dados do Wikimapia procura identificar e publicar de maneira padronizada a semântica presente nas anotações visando reconhecer não somente classes de elementos geográficos (*e.g.*, hotel, restaurante, hospital), como também outras características relevantes para aplicações (*e.g.*, tipo de acomodação, tipo de comida, especialidades médicas). As análises dos dados são restritas a um estudo de caso da cidade italiana de Milão, onde as palavras das categorias dos elementos espaciais, encontradas nos sistemas colaborativos, são confrontadas lexicamente com os nomes das entidades e suas possíveis características, definidas em ontologias e conjuntos de dados ligados a elas associados, tais como GeoNames e DBpedia.

1.1 JUSTIFICATIVA

O alto valor agregado ao desenvolvimento de bases geográficas e a escassez de fontes de dados adequadas são desafios determinantes para as pesquisas e aplicações de dados espaciais. Sistemas colaborativos da *web* são fontes de dados de rápida expansão e atualização constante. Entretanto, estes dados apresentam problemas como vandalismo, inconsistência e esparsialidade (MOONEY et al., 2011), requerendo assim tratamento adequado (*e.g.*, análises, limpeza, integração) antes do seu uso.

A interoperabilidade do dado ligado possibilita a combinação de diferentes fontes geográficas. Os dados geográficos ligados possuem expressividade suficiente para realizar experimentos como a busca por similaridade semântica no espaço geográfico ou o enriquecimento semântico de trajetórias de objetos móveis. Consequentemente, a utilização dos dados geográficos ligados proporciona melhores resultados em recuperação de informação e novas funcionalidades nas aplicações.

[2]Disponível em: `http://www.geonames.org/`. Acesso em nov. 2013.

1.2 OBJETIVO GERAL

O objetivo geral desse trabalho é uma pesquisa prospectiva sobre a semântica encontrada em anotações de dados geográficos produzidas por ferramentas colaborativas e verificar sua qualidade e a possibilidade de utilização em pesquisas futuras.

1.2.1 Objetivos Específicos

Os objetivos específicos desse trabalho são:

1. Revisar a literatura técnico-científica a respeito de bancos de dados geográficos, *volunteered geographic information* e anotações, particularmente anotações semânticas.

2. Avaliar a possibilidade da extração de dados geográficos com anotações de ferramentas colaborativas da *web* e avaliar a qualidade semântica das fontes dos dados.

3. Verificar a possibilidade de correlacionar anotações coletadas colaborativamente com conceitos de ontologias já existentes e de converter a coleção de anotações de um sistema colaborativo em ontologia.

1.3 METODOLOGIA

A execução desse trabalho foi dividida em etapas. Cada etapa resultou na escrita de um relatório contendo resumos e descrições das tarefas realizadas. Esses relatórios foram utilizados posteriormente para a redação dos documentos requeridos nas entregas 1 e 2 do trabalho de conclusão de curso (TCC).

Na etapa inicial, o estudo da fundamentação técnico-científica, foi realizada uma revisão bibliográfica sobre o estado da arte. Ela se dividiu em três pesquisas: nas áreas de dados geográficos, *volunteered geographic information* e anotações semânticas. A segunda pesquisa aprofundou-se em anotações semânticas de dados geográficos e ontologias. Não foram descartadas referências de linhas de pesquisa a estas relacionadas, como a *web* semântica e a *web* 2.0, desde que contribuissem para o projeto.

Na segunda etapa, conversão de dados geográficos em dados ligados, são apresentadas a definição do problema, os trabalhos correlatos e a proposta do processo de conversão de dados geográficos anotados em dados ligados.

Na terceira etapa, a implementação da proposta, foi planejada a implementação das etapas do processo restrito a um caso de estudo: objetos geográficos do Wikimapia, restritos aos limites da cidade de Milão, Itália.

Foram previstas ainda duas etapas: a redação do documento parcial do TCC, onde foram reunidos e revisados os relatórios de cada etapa para o desenvolvimento dos documentos de entrega solicitados pelas disciplinas INE5433 e INE5434, e a preparação da defesa, quando a apresentação final foi planejada.

1.4 ESTRUTURA DO TRABALHO

O capítulo 2 aborda os fundamentos teóricos a respeito de dados geográficos (seção 2.1), coleta colaborativa de dados geográficos (seção 2.2) e anotações geográficas (seção 2.3). O capítulo 3, conversão de dados geográficos anotados em linked data, apresenta a definição do problema (seção 3.1) e a proposta do processo de conversão de dados geográficos anotados em dados ligados (seção 3.2). O capítulo 4, caso de estudo: Wikimapia, apresenta alguns resultados da implementação da proposta no sistema colaborativo Wikimapia, ilustrando as etapas de extração de dados (seção 4.1), limpeza de dados (seção 4.2), enriquecimento semântico (seção 4.3), alinhamento e/ou mapeamento ontológico (seção 4.4), triplificação (seção 4.5) e resultados obtidos (seção 4.6). O capítulo 5 apresenta trabalhos correlatos e o capítulo 6, a conclusão e trabalhos futuros (seção 6.1).

2 FUNDAMENTOS

Este capítulo examina as dificuldades referentes a semântica na utilização dos dados geográficos, como os sistemas colaborativos contribuem para a coleta de informações geográficas e de que forma as diferentes tecnologias de anotação digital são empregadas em dados geográficos.

2.1 DADOS GEOGRÁFICOS

O termo dado geográfico se refere a todos os dados associados a características espaciais e que tenham referência a uma localização na superfície da Terra. O dado geográfico, ou geoespacial, pode se referir a elementos (*i. e.*, construções, rios e estradas) ou a fenômenos (*i. e.*, massas de ar, disseminação de doenças e temperatura) do espaço. Os dados geográficos são responsáveis pela representação computacional do espaço geográfico conceitualizado a partir do mundo real (CASANOVA et al., 2005). A representação computacional do espaço geográfico e de suas informações geoespaciais, seja sobre elementos ou fenômenos geográficos, é o objeto de estudo da ciência de geoprocessamento.

Dados geográficos geralmente são representados computacionalmente por mapas construídos por imagens, aeroespaciais ou capturadas por satélites, ou geometrias constituídas de coordenadas coletadas, por exemplo, por dispositivos com GPS. O dado geoespacial se difere dos demais pela necessidade de expressar a localização geográfica e as relações topológicas com outros dados geoespaciais.

O dado geográfico necessita de dois componentes para representar a geoinformação (RIGAUX; SCHOLL; VOISARD, 2000):

- O componente referenciado como atributo espacial ou geométrico. Este componente descreve a localização, forma, orientação e tamanho do objeto no espaço, de duas ou três dimensões.

- O componente referenciado como atributo temático, descritivo ou não-espacial. Os atributos temáticos descrevem o dado geográfico da mesma forma que atributos de dados relacionais e por isso também são referenciados como atributos alfanuméricos.

Diferente do atributo temático, o atributo espacial do dado geográfico não possui um tipo de dado padrão, por isso ele também é chamado de dado ou objeto espacial. Existem dois modelos de representação para dados espaciais: o vetorial e o matricial. A figura 1 ilustra os dois diferentes modelos.

Figura 1 – **Representações de dados geográficos (página do Wikimapia[1])**

(a) Representação Vetorial (b) Representação Matricial

A representação vetorial, ilustrada na figura 2(a), utiliza geometrias para representar elementos espaciais geralmente individualizados e com limites bem definidos. As estruturas geométricas dos dados espaciais são associadas a atributos alfanuméricos. As geométricas de elementos espaciais são representadas usando três formas básicas: ponto, linha e polígono; e composições dessas geometrias básicas. Cada elemento possui uma ou mais coordenadas, que referenciam o espaço de acordo com um determinado sistema de projeções cartográficas.

A representação matricial, ilustrada na figura 2(b), baseia-se em imagens *raster* ou matrizes e é excelente para representar fenômenos e características do espaço que são contínuos e não bem delimitados. A variação de cores entre *pixels* da imagem é utilizada para representar por exemplo a elevação do terreno, a temperatura, a radiação do terreno ou a cor da região em uma foto aérea. Outras informações alfanuméricas podem estar ligadas a cada *pixel*, como a altitude, o nível da copa das árvores o tipo ou uso do solo.

Ambas representações podem ser interligadas coerentemente pela utilização do mesmo sistema de coordenadas. Os sistemas de coordenadas são necessários para expressar a posição de pontos sobre uma superfície, seja ela representada como elipsóide, esfera ou plano, segundo a apostila do Instituto Brasileiro de Geografia e Estatística (IBGE) (MAGALHÃES et al., 1998). Existem diversos sistemas de coordenadas, cada qual respeitando uma determinada projeção cartográfica do planeta. Este trabalho utiliza apenas o padrão *World Geodetic System* (WGS) 84, um dos sistemas de coordenadas estabelecidos pela *Association of Oil and Gas Producers Geomatics Committee* (OGP *Geomatics Committee*). Sua descrição está publicada em *European*

[1]Disponível em: <`http://wikimapia.org/#lang=pt&lat=-27.601345&lon=-48.518221&z=17&m=w`>. Acesso em jun. 2013.

Petroleum Survey Group Geodetic Parameter Dataset (*EPSG Geodetic Parameter Dataset*).

A representação computacional do objeto espacial - o atributo espacial - torna o dado geográfico um dado especializado, diferente do dado relacional, e o caracteriza como um dado complexo. O modelo relacional de dados é limitado, insuficente para representar e manipular dados geométricos. A fim de persistir o dado geográfico, ou a aplicação trata o dado espacial isoladamente, ou o sistema de gerenciamento de bancos de dados (SGBD) precisa de técnicas especializadas para definir, construir, manipular, consultar e atualizar o BDG. Um SGBD com estas técnicas é considerado geograficamente extendido.

Os BDGs são componentes fundamentais de muitas aplicações, sendo a principal tecnologia os sistemas de informação geográfica. Todas elas necessitam do dado geográfico para solucionar problemas do mundo real, seja pela organização, estudo ou a simples disponibilização da informação geoespacial.

2.1.1 Sistema de Informações Geográficas

Um SIG, ou *Geographic Information System* (GIS), realiza o tratamento computacional dos dados geográficos e recupera informação não apenas com base nas características alfanuméricas, mas também através da localização espacial (CâMARA; MONTEIRO; MEDEIROS, 2004). O SIG é uma coleção integrada de *softwares* e dados para visualização e análise de mapas. Todo SIG completo deve suportar a inserção e integração dos atributos descritivos e estruturas geométricas dos dados geográficos, oferecer mecanismos para análises, consultas, recuperação e visualização do conteúdo da base de dados geográficos. Entretanto, o SIG não precisa apresentar todas estas funcionalidades.

O SIG também pode realizar transformações e operações espaciais sobre os dados geográficos, combinando-os para criar novas representações do espaço geográfico e suportando simulações para facilitar o trabalho de especialistas de diversas áreas, *e. g.*, administração pública, transporte, aplicações militares e sistemas ambientais.

O BDG é o componente central de um SIG, responsável pelo armazenamento e recuperação do dado geográfico. Entretanto, o SIG possui outras ferramentas para processar a informação geoespacial que formam o gerenciador de entrada e integração dos dados geográficos, o analisador espacial, o mecanismo de visualização e plotagem e a interface gráfica.

Uma representação mais robusta do dado geográfico é necessária para

facilitar sua manipulação dentro do SIG. A feição (*feature*) geográfica, também chamada de objeto geográfico ou entidade geográfica, é uma abstração de um elemento ou fenômeno do mundo real (*e. g.*, uma construção, uma floresta, uma área de baixa pressão atmosférica) implementada em uma linguagem de programação de alto nível qualquer (diferente do dado geográfico não necessáriamente implementado em uma linguagem de alto nível). Os atributos da feição (*Feature attributes*) são as características destes fenômenos - como o nome, o tipo ou um valor de domínio do dado - e podem ser atributos tanto alfanuméricos quanto geométricos (OGC, 2011).

Do ponto de vista da aplicação, a utilização apropriada do SIG implica na escolha das representações computacionais mais adequadas para capturar a semântica de seu domínio de aplicação. Enquanto que, do ponto de vista tecnológico, o desenvolvimento apropriado de um SIG significa amplitude, tanto em estrutura quanto em mecanismos de manipulação do dado geográfico, para representar a grande diversidade de concepções do espaço (CASANOVA et al., 2005). A representação computacional do espaço é fundamental no desenvolvimento de um SIG e sua escolha não afeta apenas o modelo lógico do BDG como também todos os seus outros componentes.

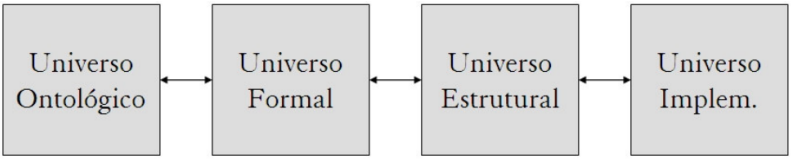
2.1.2 Representação Computacional do Espaço

Como já foi visto, os dados geográficos são responsáveis pela representação computacional do espaço geográfico conceitualizado no mundo real (CASANOVA et al., 2005). Entretanto, a tradução da informação geográfica para o computador é o problema fundamental da geoinformação.

Do mesmo modo que é necessário criar um modelo conceitual e lógico baseando-se nas necessidades do domínio, antes de criar um banco de dados relacional, deve-se realizar a modelagem do espaço sob o ponto de vista do domínio da aplicação geoespacial. Para isto, é utilizado o paradigma dos quatro universos (CâMARA, 1995) que define o processo de especificação e conversão da informação, através dos universos ontológico, formal, estrutural e de implementação, os quais são ilustrados na figura 2 e descritos nos itens a seguir.

- No universo ontológico (real) existem os objetos e fenômenos de interesse para nossa aplicação, alguns relacionados com o espaço, outros não. A primeira etapa do paradigma dos quatro universos exige a identificação destes objetos e fenômenos, sendo importante não apenas a avaliação da equipe de desenvolvimento como também a opinião dos especialistas do domínio da aplicação.

Figura 2 – **Paradigma dos quatro universos (CASANOVA et al., 2005)**



- No universo formal (conceitual ou matemático), os objetos e fenômenos de interesse e as relações entre eles são descritos de maneira formal (com fundamentação matemática), idealizados e transformados em entidades e relações, respectivamente. Neste universo já é possível classificar a entidade geográfica em dado contínuo, como a umidade, previsão de preciptação, temperatura e massas de ar, ou objetos individualizáveis, como uma estrada, uma construção e uma divisa política. O processo de transição da informação do mundo real para o conceitual é chamado de modelagem conceitual. Realizando a modelagem conceitual podemos obter o modelo de dados conceitual.

- No universo estrutural (de representação), as entidades geográficas são associadas a formas de representação geométrica, matricial ou vetorial, que podem ainda ser especializadas. A escala, a projeção cartográfica escolhida e a época de aquisição do dado influenciam nesta decisão. Nesta etapa adequamos o modelo conceitual ao projeto, considerando detalhes de implementação e obtendo o modelo de dados lógico.

- Por último, no universo de implementação ocorre a realização prática dos modelos de dados conceitual e lógico através de linguagens de implementação e ferramentas como o BDG.

A modelagem conceitual de uma aplicação geográfica apresenta entidades geográficas (rios, cidades, países), que também podem ser chamadas de temas. Em um SIG, a informação geoespacial correspondente a um tópico particular é reunida em um tema (RIGAUX; SCHOLL; VOISARD, 2000). O tema, similar a uma relação no modelo relacional de dados, possui um esquema e instâncias.

As instâncias são os objetos ou fenômenos geográficos, entidades do mundo real referentes ao tema. Por exemplo, a Ilha de Santa Catarina é um objeto geográfico cujo tema pode ser ilha se estivermos construindo um SIG para analisar as ilhas catarinenses. Outros exemplos de temas são as possíveis instâncias do tema ciclone tropical (*e.g.*, Catarina, Anita, 1970 Bhola), divisão política, corpos de água e malha viária.

Os temas constumam ser organizados em níveis de informação. O esquema de um tema apresenta a sua coleção de atributos, descritivos ou espaciais, formalizando qual informação pode-se obter de seus objetos geográficos. Seguindo o último exemplo, se o tema ilha estiver definido pelo esquema: **Ilha(nome, geo)**, então o objeto geográfico Ilha de Santa Catarina deve possuir um atributo alfanumérico *nome*, por exemplo "Ilha de Santa Catarina", e um atributo espacial *geo* contendo o polígono dos limites da ilha.

A coleção de esquemas dos temas de interesse de uma aplicação, junto com a descrição das possíveis relações entre elementos espacias referentes a cada tema, formam um modelo conceitual, também chamado de modelo geográfico. O modelo geográfico é essencial tanto para a geração do modelo lógico dos dados espaciais quanto para a implementação das *features* e dos componentes que interagem com elas.

A transposição dos elementos do mundo real para o computador é complexa e extensa. O problema fundamental da geoinformação, assim como do desenvolvimento de um SIG sob o ponto de vista da aplicação, é obter a representação do espaço geográfico mais apropriada para o domínio da aplicação, em termos de qualidade da informação e performance da implementação. Este problema, aliado com algumas más práticas de desenvolvimento, resultam na dificuldade de utilização da informação geográfica, tanto para os usuários de uma aplicação específica quanto para a integração e interoperabilidade com outras aplicações.

Todavia, a representação do espaço não é o único problema. A obtenção e manutenção dessas informações descritivas, não simplesmente obtidas por GPS e imagens de satélite, são geralmente as tarefas mais caras e trabalhosas de um SIG. Pequenos grupos de reconhecimento e coleta de informação não são a solução ideal.

2.2 COLETA COLABORATIVA DE INFORMAÇÃO GEOGRÁFICA

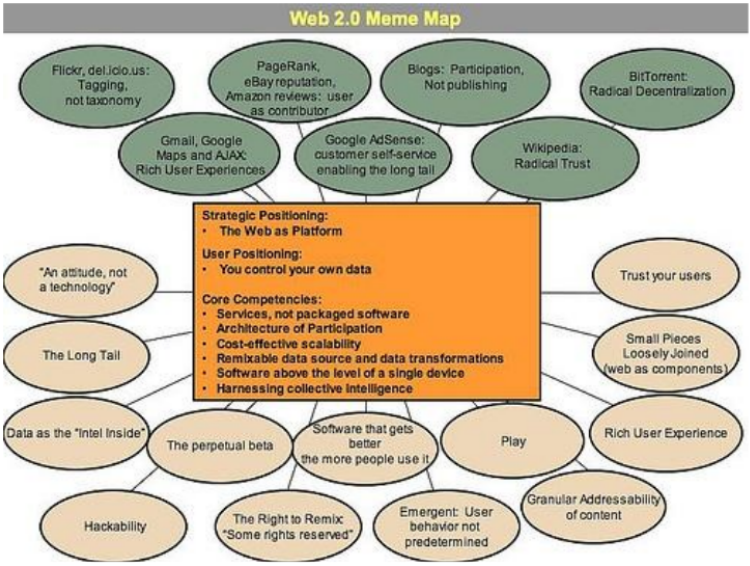
Sistemas de coleta colaborativa de informação geográfica são providos de apoio geralmente voluntário de usuários da *web* para obter informações geográficas complementares as obtidas por GPS e imagens de satélite. Esta forma de obtenção de informação geográfica também é chamada de *citizen science*, sistemas geográficos voluntários, colaboração em massa (*crowdsourcing*) geográfica e VGI (GOODCHILD, 2007). Ainda não há consenso quanto ao termo mais apropriado.

O sistema de coleta colaborativa combina os componentes de um SIG com elementos da: *web* 2.0 (SCHARL; TOCHTERMANN, 2007), inteligência coletiva (SMITH, 1994) e neogeografia (TURNER, 2006). A infraes-

trutura resultante possibilita a coleta de dados geográficos, principalmente de atributos temáticos, incluindo, por exemplo, nomes e tipos de lugares (*e.g.*, estabelecimentos prestadores de serviços, locais de interesse turístico, pontos de ônibus), informações ambientais (*e.g.*, qualidade do ar, desastres ecológicos), culturais (*e.g.*, construções e utilização do espaço) e populacionais (*e.g.*, categorias de pessoas que frequentam um local). A coleta colaborativa desses atributos facilita a construção de SIGs e aplicações.

A *web* 2.0 propõe uma rede interativa, bi-direcional, em que o usuário faz parte da construção do conteúdo do sistema *web* (O'REILLY, 2005). O conceito de *web* 2.0 é construído em torno de alguns elementos principais tais como pessoas, grupos, temas de interesse e suas relações. A figura 3 apresenta um esboço elaborado durante uma sessão de *brainstorm* da companhia O'Reilly Media contendo princípios, práticas e exemplos do que poderia ser uma *web* mais interativa.

Figura 3 – **Esboço da *web* 2.0 (O'REILLY, 2005)**



O projeto Wikipedia e aplicações privadas como Facebook[2], Twitter[3],

[2]Disponível em: <`https://www.facebook.com/`>. Acesso em nov. 2013.
[3]Disponível em: <`https://twitter.com/`>. Acesso em nov. 2013.

Instangram[4] e Zuodao[5] apresentam os conceitos da web 2.0. O processo de tradução linguistíca do conteúdo de algumas aplicações, como o Facebook, também se baseiam em tecnologias que permitem a colaboração dos usuários da *web*.

As informações descritivas dos elementos geográficos são frutos da percepção humana do espaço e de suas propriedades. Cada indivíduo coleta informações geográficas desde a infância, construindo conceitos elaborados da região onde vive. O apoio de grandes contingentes de indivíduos via *web*, com diferentes origens, culturas, interesses e localizações geográficas, de diferentes níveis de aperfeiçoamento técnico em descrição geográfica, contribui para uma coleta rápida de informação sob diferentes perspectivas. No movimento VGI, os cidadãos coletam dados espaciais e informações sobre sua própria ou outra localidade.

Entretanto, medidas devem ser tomadas para garantir a confiança das informações colaborativamente coletadas. Não há garantia de que a informação foi coletada por um especialista ou um amador. Além disso, há que se desconsiderar a desatualização, má-especificação e vandalismo dos dados (MOONEY et al., 2011). Por outro lado, a diversidade de pontos de vista garante uma fonte de informação rica e dinâmica, com crescimento rápido e muitas possibilidades de aplicação. O voluntário, amador em alguns domínios, pode apresentar informação profissional para outros domínios, principalmente quando se trata de sua localidade.

Um sistema de coleta colaborativa de informação geográfica precisa de soluções inteligentes para que a informação possa ser comunicada, montada, integrada e interpretada. Políticas de acesso e manutenção dos dados são necessárias para minimizar a depreciação e garantir a confiança da informação. Assim como a coleta de informação é colaborativa, o controle de qualidade e confiança dos dados pode ser realizado coletivamente pelos próprios usuários (FRITZ et al., 2009).

A coleta colaborativa de dados geográficos é utilizada por diversos projetos. A companhia *Environmental Systems Research Institute*[6] (ESRI) utiliza a tecnologia de coleta colaborativa para o monitoramento de desastres e emergências (JOHNSON, 2000), tais como o vazamento de óleo no Golfo do México, o terremoto do Chile, o terremoto no Haiti; e denúncia de problemas de saneamento e fiscalização de obras públicas (eGov2.0). Outras empresas, como a Copacabana Tech[7], utilizam os serviços da ESRI para

[4]Disponível em: <`http://instagram.com/#`>. Acesso em nov. 2013.
[5]Disponível em: <`http://www.zuodao.com/`>. Acesso em nov. 2013.
[6]Disponível em: <`http://www.esri.com/`>. Acesso em jul. 2013.
[7]Disponível em: <`http://www.copacabanatech.com/`>. Acesso em jul. 2013.

disponibilizar outros serviços, como a ferramenta Map4Divers[8] que ajuda a compartilhar informações de locais de mergulho. O projeto Geo-Wiki[9] visa melhorar a qualidade dos mapas da utilização do espaço (FRITZ et al., 2009).

Dois projetos destacam-se por objetivar descrever o mundo inteiro sob a perspectiva de seus usuários. O Wikimapia[10] tem o objetivo de obter descrições úteis de qualquer elemento da superficie terrestre. O projeto OpenStreetMap[11] (OSM) pretende construir um mapa livre e editável do mundo inteiro.

2.2.1 Open Street Map

O OSM é um projeto colaborativo com o objetivo de criar um mapa livre e editável de todo o mundo. O projeto foi fundado em Julho de 2004 por Steve Coast. A principal motivação eram as licenças restritivas da maioria dos mapas gratuitos, que limitavam a utilização criativa e a produção de dados geográficos. Seus dados estão disponíveis para utilização sob a linceça *Open Data Commons Open Database License* (ODbL). O OSM é a força motora por trás do termo VGI e apresenta ferramentas para criar, manipular e divulgar dados geográficos fornecidos voluntariamente.

Os dados são coletados a partir de dispositivos GPS portáteis, fotografias aéreas, outras bases geográficas como o sistema *Topologically Integrated Geographic Encoding and Referencing* (TIGER), dentre outras fontes. As contribuições obtidas são desde o conhecimento dos usuários, esforços de grupos organizados e dados doados em cooperações, como os dados de Netherlands doados pela compania *Automotive Navigation Data* (AND)[12] . Os usuários podem inserir dados coletados por GPS ou utilizar ferramentas de edição vetorial desenvolvidas pela comunidade. Inspirado em sistemas colaborativos não geográficos como o Wikipédia, o OSM possibilita a edição de mapas e guarda um histórico completo das modificações.

Para representar os dados, o OSM possui as entidades geográficas:

- *Node*: um par de coordenadas geográficas, a longitude e latitude relativas ao sistema de referência geográfica WGS 84.

- *Way*: uma sequência ordenada de *nodes*. *Ways* com *nodes* inicial e final idênticos são chamados de fechados. Quando fechados, são usados na

[8]Disponível em: <`http://map4divers.appspot.com/`>. Acesso em jul. 2013.
[9]Disponível em: <`http://www.geo-wiki.org/`>. Acesso em jul. 2013.
[10]Disponível em: <`http://www.wikimapia.org/`>. Acesso em jul. 2013.
[11]Disponível em: <`http://www.openstreetmap.org/`>. Acesso em jul. 2013.
[12]Disponível em: <`http://www.and.com/`>. Acesso em jul. 2013.

representação de construções e faixas de terra, e quando não fechados, rios e estradas.

- *Relation*: um conjunto de *Nodes* e/ou *Ways* e/ou *Relations*, usado para representar relacionamentos entre os componentes espaciais como os multipolígonos. Cada entidade participante de uma *Relation* possui um *role* (papel).

Cada uma dessas entidades geográficas possui um identificador único e pode ter qualquer número de anotações, no formato de chave-valor. Essas anotações, chamadas de *tags* na terminologia do OSM, são usadas para descrever informações como o tipo de um objeto (chave), como restaurante, rua, lagoa (valores), por exemplo, e seus detalhes mais relevantes como endereço, se o acesso é restrito, se é iluminado, entre outros. Tanto a chave quanto o valor de uma *tag* não possuem restrição de escrita, porém existem padrões e convenções[13] a serem seguidas, dentro do processo ágil da comunidade, nos casos mais comuns de utilização das *tags*. Serviços como o TagWatch[14] criam estatísticas do uso das *tags* no OSM, sendo ferramentas úteis tanto para analisar novas tendências quanto para realizar correções nas anotações.

Um exemplo de dado OSM é a *Relation* da fronteira administrativa da Alemanha, cujo identificador é 51477 [15]. Esta relação tem em torno de 1000 *Ways* para definir a fronteira e mais de 30 anotações que descrevem, por exemplo, o nome do país representado, a Alemanha, em diferentes línguas. Também existem algumas anotações de metadados como quando e quem foi o último usuário a editar a relação.

A figura 4 apresenta a arquitetura do OSM. Os dados do OSM estão armazenados em um banco relacional Postgres com a extensão geográfica PostGIS. O OSM oferece uma *Application Programming Interface* (API) com características do protocolo de transferência de estado representativo (*Representational State Transfer* - REST). Uma versão integral da base também é publicada semanalmente. Atualmente o conteúdo total ocupa cerca de 6 GB de dados comprimidos em formato Bzip2. O projeto também publica as alterações recentes na base de dados, na frequência de minutos, horas e dias, possibilitando a sincronização de bases derivadas do OSM com o projeto. Ambas as publicações, completa e atualizações, podem ser manipuladas a partir do *Osmosis*, uma ferramenta JAVA para processar dados OSM. A API OSM fornece os dados às ferramentas de edição. Essas interfaces como,

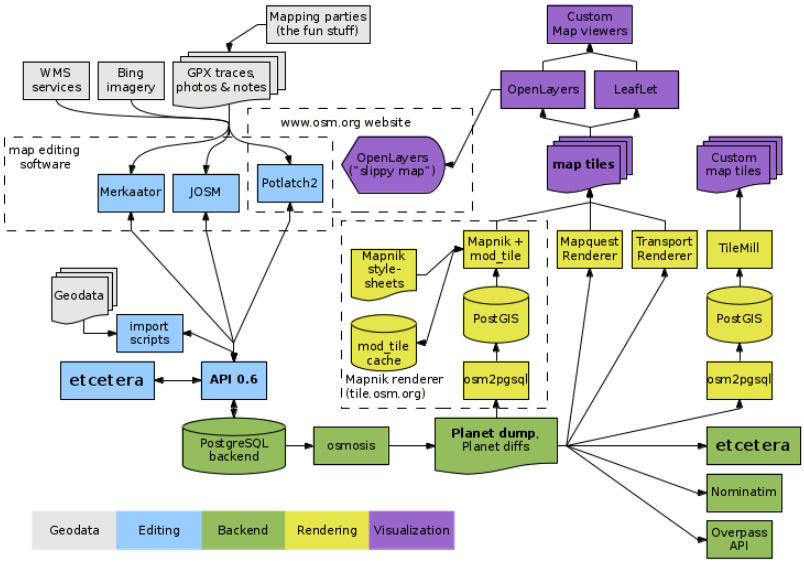
[13]Disponível em: <`http://wiki.openstreetmap.org/wiki/Map_Features`>. Acesso em jul. 2013.

[14]Disponível em: <`http://tagwatch.stoecker.eu/`>. Acesso em jul. 2013.

[15]Disponível por <`http://www.openstreetmap.org/browse/relation/51477`>. Acesso em jul. 2013

por exemplo, *online Potlatch* e as de *desktop JOSM* e *Merkaartor* são usadas pelos colaboradores para descrever, corrigir e enriquecer a base OSM.

Figura 4 – **Arquitetura OSM (página do OSM[16])**



A mudança do Google Maps, um serviço de pesquisa e visualização de dados geográficos, em não mais oferecer informações e serviços gratuitamente incentivou a adesão de interessados do mundo todo ao projeto OSM. Entre os novos usuários institucionais, podemos citar o Foursquare, o TripAdvisor, a Wikipédia e até mesmo a Apple no seu aplicativo iPhoto, que substituiram o Google Maps pelo OSM. O projeto ganhou certa projeção quando foi usado pelas equipes de ajuda humanitária no Haiti após o furacão Isaac para mapear toda a área afetada em apenas 2 dias. Grupos de mapeadores, as *mapping parties* se reunem com frequência para melhorar os dados de algumas regiões. O crescimento do projeto é enorme: Desde a fundação, são mais de um milhão de usuários que já contribuiram com mais de 3 bilhões de pontos coletados por GPS, mais de 1 bilhão de *Nodes*, 174 milhões de *Ways* e quase 2 milhões de *Relations*.

[16]Disponível em: <`http://wiki.openstreetmap.org/w/images/1/15/OSM_Components.png`>. Acesso em nov. 2013.

2.2.2 Wikimapia

O Wikimapia é um projeto com o objetivo de criar e manter um mapa atualizado, completo, gratuito e multilíngue de todo o mundo. O Wikimapia foi criado por Alexandre Koriakine e Evgeniy Saveliev em maio de 2006 sob o *slogan* "Vamos descrever o mundo todo", pretendendo dispor de informação detalhada sobre cada lugar no planeta Terra. Seu conteúdo é inteiramente criado por voluntários da *Internet*, não havendo restrições para colaborar. Todo conteúdo é disponível para reuso sob a lincença *Creative Commons license Attribution-ShareAlike 3.0*.

Os colaboradores do Wikimapia usufruem de um SIG para realizar consultas, criar, editar e visualizar os dados. Sua interface é baseada no Google Maps e foi muito bem aceita pelos usuários. As imagens *raster* são fornecidas pelo Google, as mesmas usadas pelo Google Maps, e pelo OSM. Entretanto, a qualidade dessas imagens não são garantidas pelo projeto. A criação de novos lugares é basicamente uma conversão de imagens rasterizadas em estruturas vetoriais, utilizando ferramentas de desenho para definir um polígono ou uma linha.

Após a seleção ou inserção da geometria de um elemento geográfico, a ferramenta disponibiliza formulários *web* para coletar atributos descritivos. A descrição pode incluir *links* de vídeos, imagens e páginas na *web*, em especial para a Wikipédia. O usuário é incentivado a classificar uma ou mais vezes o local segundo uma hierarquia de classes não fixa (aberta). Se não encontrar a opção que deseja é possível expandir as opções. Uma categoria do Wikimapia pode ser principal, secundária ou uma categoria sinônimo. Esta classificação é aproveitada para filtrar locais durante a visualização do mapa. O usuário é orientado a realizar qualquer descrição de um ponto de vista neutro.

A medida que um colaborador contribui com o projeto, adquire experiência e sobe na hierarquia de usuários. Esta hierarquia possui 3 níveis e quanto maior, mais liberdade e responsabilidade são concedidas ao usuário.

- Usuário nível 0: permissão para adicionar e editar locais, e usar o sistema de mensagem pessoal.
- Usuário nível 1: todos os usuários atingem este nível depois de alguns dias. Nele, o usuário pode também alterar polígonos, adicionar ruas, estradas, rios, ferrovias, deletar locais e contribuir com o fórum.
- Usuário nível 2: O usuário avançado é responsável por tarefas vitais no projeto como banir vândalos e coibir atentados aos dados do mapa.

Além da ferramenta para visualização e edição dos dados geográficos e anotações, o Wikimapia disponibiliza também o Wikimapia *Cityguide*, um

guia com informações de algumas cidades como Moscou[17] e Nova Iorque[18]. Há uma *wiki* com documentação e orientações para os usuários, um *blog* com as últimas notícias do projeto, um fórum para discussão sobre novas *features*, correções de problemas e *bugs* e uma API[19]. Entretanto, a documentação é focada na manipulação da interface. Informações quanto a representação dos dados e arquitetura do projeto são escassas.

O Wikimapia já possui mais de vinte milhões de objetos em sua base de dados e ultrapassa um milhão e quinhentos mil usuários. Com a recente integração com o Facebook, é esperado um crescimento ainda maior. Entretanto, por ser uma base de dados aberta, há sempre o risco de dados desatualizados, vandalismo e ausência da descrição de novos locais. Um bom exemplo é a criação de categorias. Antigamente a anotação da categoria era livre e muitos usuários repetiam a informação do título do local, ou colocavam informações impertinentes. Quando a informação era cabível, ainda havia casos de duplicidade de categoria com linguagens diferentes ou os sinônimos. Atualmente são aplicadas políticas mais rigorosas à hierarquia de anotação, porém ainda existem problemas.

Vale também observar o ato generalizado de vandalismo que aconteceu em Julho de 2009. Sem o consentimento da comunidade, foi lançada uma nova versão do Wikimapia. A revolta dos colaboradores resultou em usuários banidos e tópicos apagados, o que foi considerado pela comunidade como censura. Na época também houve discussões a respeito do excesso de propagandas no sistema. Muitos usuários abandonaram o projeto e alguns vandalizaram a base de dados com objetos e anotações incoerentes.

2.3 ANOTAÇÕES EM DADOS GEOGRÁFICOS

A necessidade do atributo temático do dado geográfico é comum para as aplicações. O atributo temático é considerado uma anotação digital, uma informação sobreposta (*superimposed information*) (DELCAMBRE; MAIER, 1999), e também pode ser chamado de rótulo (*label*) ou etiqueta (*tag*). A abordagem de anotação facilita a administração das diferentes perspectivas e interpretações do dado espacial. O entendimento a respeito do objeto anotado pelo usuário depende do contexto, do domínio da aplicação e do ponto de vista do usuário.

Há diversos tipos de anotação. Uma das categorizações leva em conta como é o armazenamento da anotação, intrusivo ou não intrusivo. Anotações

[17]Disponível em: <`http://moscow.wikimapia.org/`>. Acesso em jul. 2013.
[18]Disponível em: <`http://new-york.wikimapia.org/`>. Acesso em jul. 2013.
[19]Mais informações no anexo A

intrusivas são acopladas ao documento ou objeto anotado, enquanto as não intrusivas são armazenadas separadamente e referenciam o documento ou objeto com auxílio de identificadores e endereçamentos. Outra categorização considera o modo de representação da anotação, podendo ser livre ou semântica. A anotação livre associa o objeto anotado com um texto ou outra informação escrita livremente. As anotações semânticas, por outro lado, associam o objeto anotado a descrições com semântica bem definida (*e.g.*, em uma ontologia), sendo características dos dados ligados (*linked data*).

2.3.1 Anotação Livre

A anotação livre é aquela que não apresenta nenhum tipo de estruturação de seu conteúdo e é informalmente composta de um texto ou outro tipo de informação livre. A liberdade da anotação livre torna o SIG que a utiliza mais versátil para a manipulação de novos dados geográficos.

Entretanto, a liberdade também apresenta problemas. Cada pesquisador, especialista, empresa ou país tem sua própria linguagem e métodos descritivos. A anotação livre herda os problemas da manipulação de textos livres: sinônimos, ambiguidades e erros ortográficos, tornando mais complexo interpretá-la e utilizá-la de forma automatizada em aplicações. Sistemas de anotações livres, como *Annotea* (KAHAN et al., 2001) e *GeoNote* (GIL; KOZIEVITCH; TORRES, 2011), e SIGs com anotações livres partilham estes mesmos problemas.

Sistemas sociais de anotação (*Social Tagging Systems* - STS) utilizam a anotação livre para agregar significado às informações ou objetos anotados e, posteriormente, recuperar a informação. Por exemplo, em sistemas de coleta colaborativa de dados geográficos (*e.g.*, OSM e Wikimapia) são anotados elementos espaciais, em Flickr são anotadas fotografias e imagens, em del.ici.ous são anotadas páginas da *web*. As estruturas conceituais deste tipo de anotação, emergentes de STSs, são chamadas de *folksonomias* (GARCíA-CASTRO; GARCíA, 2011). Os três princípios da *folksonomia*, fundamentais para desambiguação dos termos usados em anotações e para compreensão do objeto anotado (WAL, 2007), são:

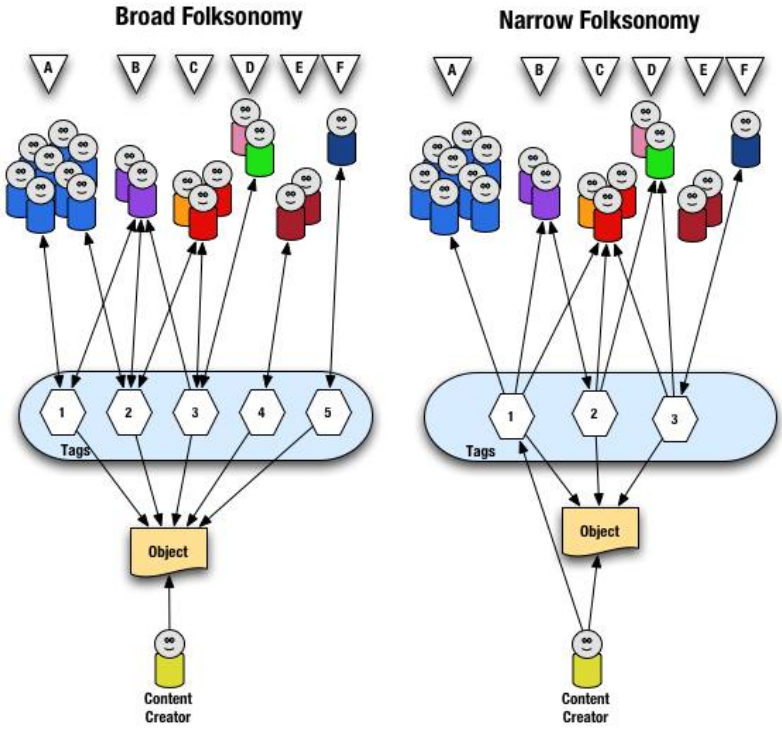
1. anotação (*tag*), *i.e.* termo(s), ou palavra(s) chave(s), utilizado(s) em anotações.

2. objeto sendo anotado, *e.g.*, elemento espacial, fotografia, página da *web*.

3. identidade, *i.e.* relações de sinônimo.

Uma *folksonomia* pode ser classificada como ampla (*broad*) e restrita (*narrow*). Uma *folksonomia* ampla apresenta objetos com múltiplas instâncias da mesma *tag* (*e.g.*, del.ici.ous), enquanto na *folksonomia* restrita o objeto é diretamente anotado por uma única instância de cada *tag* (*e.g.*, Flickr). A figura 5 apresenta um diagrama ilustrando o processo de anotação de e recuperação de informações ou objetos por diferentes grupos de usuários (A, B, C, D, E, F) do sistema social de anotação. Setas de *content creator* para objeto representam a relação de criação de conteúdo (objeto ou informação). Setas de *tag* para objeto significam que aquele objeto é anotado com aquela *tag*. Setas de *tag* para grupo de usuários indicam que aquele grupo utilizam aquela *tag* para recuperar o objeto anotado. Setas de grupo de usuários ou *content creator* para *tag* ilustra que o grupo ou o criador do conteúdo anotou o objeto em questão com aquela *tag*.

Figura 5 – ***Folksonomia* ampla e restrita (WAL, 2005)**



(a) *Folksonomia* Ampla

(b) *Folksonomia* Restrita

A *folksonomia* geralmente é representada por nuvem de etiquetas[20], onde a variação de tamanho de um termo é proporcional a sua frequência de uso em anotações do sistema.

Algumas normativas (OGC, 2005) e padrões, de anotação e simbologismo, foram propostos em diversos domínios diferentes para evitar esses erros, seguindo as especificações já definidas em enciclopédias (*thesaurus*), glossários, e dicionários geográficos (*gazetters*), *e.g.*, *Agi GIS Dictionary*[21], *Getty Thesaurus of Geographic Names Online*[22]. Mesmo assim, as convenções se tornam muito complexas de administrar. Não podem ser muito genéricas nem muito específicas. Uma definição mais rigorosa e formal é necessária para estabelecer qual conteúdo é esperado de um determinado atributo temático.

2.3.2 Anotação Semântica

A anotação semântica é uma marcação com sêmantica bem definida (BERNERS-LEE; HENDLER; LASSILA, 2001), usualmente correlacionada com uma ontologia (GUARINO, 1998), que explicita metadados importantes do objeto anotado para o mundo ao seu redor. Segundo o *World Wide Web Consortium*[23] (W3C), a ontologia define os conceitos e relacionamentos usados para descrever e representar uma área de interesse. Na prática, uma ontologia pode ser muito complexa (com centenas de conceitos e relacionamentos) ou muito simples (descrevendo apenas um ou dois conceitos). Ontologias mais simples também são chamadas de vocabulários.

As anotações semânticas surgiram com a necessidade de criar documentos inteligentes (FRAPPAOLO et al., 1994). Os documentos inteligentes são documentos que conhecem seu próprio conteúdo. Eles são capazes de informar aos programas que os manipulam o que fazer com eles e como. As tecnologias da *Web* Semântica, como a linguagem *Resource Description Framework* (RDF) para representar conhecimento e a linguagem *Web Ontology Language* (OWL) de descrição de ontologias, tornam possível novos meios de anotação semântica, recuperação, interpretação e agregação de informação. O detalhamento melhora a recuperação da informação (WELTY;

[20] Por exemplo, a nuvem de etiquetas do Flickr. Disponível em: <`http://www.flickr.com/photos/tags/`>. Acesso em jul. 2013.

[21] Disponível em: <`http://loi.sscc.ru/gis/defterm/agidict/welcome.html`>. Acesso em jul. 2013.

[22] Disponível em: <`http://www.getty.edu/research/tools/vocabularies/tgn/`>. Acesso em jul. 2013.

[23] Disponível em: <`http://www.w3.org/standards/semanticweb/ontology`>. Acesso em jul. 2013.

IDE, 1999) e resolve os problemas de ambiguidade e integração de dados heterogêneos.

A explicitação de metadados por meio da anotação semântica não facilita apenas a manipulação de documentos da *web* como também de imagens médicas, fotos e arquivos de texto. Como exemplo do uso de anotação semântica, podemos considerar aplicações de diversas áreas: o projeto *Medical Imaging and Advanced Knowledge Technologies* (MIAKT) (BONTCHEVA; WILKS, 2004), em notícias de tv e rádio (DOWMAN et al., 2005), no estudo do genoma (KIM et al., 2003), em páginas *web* para deficientes visuais (PLESSERS et al., 2005), na informação do trabalho e ocupação (MAYNARD et al., 2004), nas compras online (SVAB; LABSKY; SVATEK, 2004) e na descrição de artefatos culturais em museus (HUNTER et al., 2004).

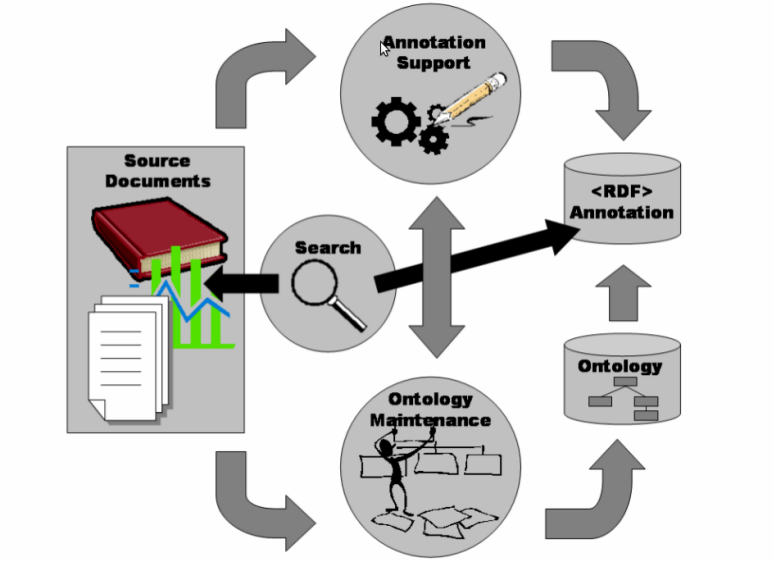
A utilização da anotação semântica não é trivial. É recomendado identificar o processo de gestão de conhecimento (*Knowledge Management - KM*) que considera todas as necessidades da anotação semântica para a aplicação. O processo de KM centrado em documentos inteligentes (UREN et al., 2006) necessita manipular três classes de dados: ontologias, documentos e anotações. Como ilustrado na figura 6, novos tipos de ferramentas são necessários para administrar o uso de anotações semânticas: ferramentas de anotação semântica de documentos, ferramentas de manutenção da ontologia e ferramentas de recuperação semântica.

A ferramenta de anotação semântica é responsável pela anotação dos documentos inteligentes, sendo geralmente utilizada por especialistas do domínio. A ferramenta de manutenção da ontologia possibilita a mudança e a evolução da semântica das anotações. A ferramenta de recuperação semântica usufrui dos metadados de semântica bem definida para conceder maior expressividade na pesquisa dos documentos, considerando de forma granular a semântica do objeto buscado.

Mesmo utilizando o processo de gestão de conhecimento centrado em documentos inteligentes e as anotações semânticas, nem todos os problemas da definição são solucionados. Criar anotações semânticas e ontologias não são tarefas simples. Tais anotações devem ser realizadas por meios semi-automáticos e/ou com apoio de especialistas de domínio, recursos geralmente escassos e também propensos a erros. Além disso, a anotação semântica se torna obsoleta sem a devida manutenção da ontologia.

Cada aplicação tem sua visão particular de um determinado domínio, necessitando de anotações que sigam sua ontologia particular. A reutilização de anotações semânticas não é trivial, devendo-se considerar tanto a ontologia da anotação quanto aquela da aplicação consumidora. A interoperabilidade dos dados anotados não é garantida apenas pelo processo de KM.

Figura 6 – **Processo de KM centrado em documentos inteligentes (UREN et al., 2006)**



2.3.3 Dados ligados

Os dados ligados (*linked data*) são uma coleção de bases de dados interrelacionadas, publicadas na *web* segundo um formato padrão alcançável e manipulável pelas ferramentas da *web* semântica, segundo a W3C[24]. Estas ferramentas possibilitam que máquinas e pessoas realizem consultas e integração de dados, garantindo assim a interoperabilidade. Os dados são interrelacionados de maneira similar às âncoras do hipertexto, formando uma estrutura de navegação não-linear. As ligações representam relacionamentos entre as características descritas nos RDFs, interligando informações de diferentes ontologias, aplicações e fontes. Assim como o documento *hypertext markup language* (HTML) está para a *web* de hipertexto, os documentos da *web* de dados são os dados ligados, propostos pela *web* semântica.

Há quatro regras (expectativas de comportamento) para a publicação de dados ligados (BIZER; HEATH; BERNERS-LEE, 2009):

[24]Disponível em: <`http://www.w3.org/standards/semanticweb/data`>. Acesso em jul. 2013.

- Utilizar *uniform resource identifier* (URI) para identificar recursos (objetos, documentos, eventos, pessoas, recursos da web e nomes).
- Utilizar *hypertext transfer protocol* (HTTP) URIs quando referenciar recursos, para torná-los acessíveis quando citados.
- Quando alguém acessa uma URI, fornecer informação útil sobre aquele recurso, utilizando padrões como RDF e SPA RDF *Query Language* (SPARQL).
- Incluir outras URIs nas descrições dos recursos, interligando informação.

A tecnologia de dados ligados utiliza meios de desambiguação semântica providos pelas ontologias e interliga dados independente de sua aplicação fonte. O projeto *Friend of a Friend* (FOAF) utiliza sua ontologia para construir redes sociais. *Opera Community* é um portal *web* que publica suas informações em formato RDF. Sendo assim, é possível incluir, por exemplo, na descrição da URI de Leigh Dodds que ele conhece Dan Brickley, utilizando uma referência ao URI de Dan publicada pelo *Opera Community*.

A figura 7 apresenta um RDF ligado a outro[25], contendo informação mais detalhada de Dan. Entretanto, os dados ligados do FOAF não respeitam a primeira regra: utilizar URIs como identificadores, apresentando referências internas[26] ao documento quando deveria apresentar referências externas[27], tornando desnecessária a repetição da descrição de Dan no documento de Leigh, como mostra figura 8.

O exemplo mais famoso de dados ligados é o projeto DBpedia. Este projeto é uma *wiki* semântica, *i.e.*, uma *wiki*, por exemplo a Wikipédia, que disponibiliza suas informações em um formato, em geral RDF, que possibilita realizar consultas. Estas consultas podem ser realizadas utilizando a linguagem SPARQL.

O *Linked Data Plataform* (LDP) *Working Group* da W3C pretende expandir o conceito do dado ligado em uma especificação. O modelo extendido do LDP fornece a definição formal[32] de uma arquitetura para acesso, de escrita e leitura, de dados ligados, publicando dados no formato RDF e

[25] http://rdfweb.org/people/danbri/foaf.rdf

[26] <rdf:resource="#dan">

[27] <rdf:resource="http://example.org/rdf_doc#dan">

[29] Disponível em: <`http://wiki.foaf-project.org/w/UsingFoafKnows`>. Acesso em jun. 2013.

[31] Disponível em: <`http://wiki.foaf-project.org/w/UsingFoafKnows`>. Acesso em jun. 2013.

[32] Disponível em: <`https://dvcs.w3.org/hg/ldpwg/raw-file/default/ldp.html`>. Acesso em jun. 2013.

Figura 7 – **RDF ligado segundo projeto FOAF (exemplo extraído da Wiki do projeto FOAF** [29]**)**

```
<rdf:RDF xmlns:rdf=
  "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <foaf:Person>
    <foaf:name>Leigh Dodds</foaf:name>
    <foaf:firstName>Leigh</foaf:firstName>
    <foaf:surname>Dodds</foaf:surname>
    <foaf:knows rdf:resource="#dan"/>
  </foaf:Person>

  <foaf:Person rdf:ID="dan">
    <foaf:name>Dan Brickley</foaf:name>
    <rdfs:seeAlso rdf:resource=
      "http://example.org/rdf_doc#dan"/>
  </foaf:Person>
</rdf:RDF>
```

Figura 8 – **RDF ligado (figura adaptada do exemplo extraído da Wiki do projeto FOAF**[31]**)**

```
<rdf:RDF xmlns:rdf=
  "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <foaf:Person>
    <foaf:name>Leigh Dodds</foaf:name>
    <foaf:firstName>Leigh</foaf:firstName>
    <foaf:surname>Dodds</foaf:surname>
    <foaf:knows rdf:resource=
      "http://example.org/rdf_doc#dan"/>
  </foaf:Person>
</rdf:RDF>
```

utilizando o protocolo padrão HTTP. A aplicação apresenta características do protocolo REST, utilizando requisições HTTP GET, PUT e DELETE e respondendo as requisições com RDF(s).

2.3.4 Semântica Geoespacial

Sistemas Geográficos baseados em Ontologias (SIG-O), do inglês *Ontology-Driven* GIS (ODGIS) (FONSECA; EGENHOFER, 1999), são aqueles que utilizam anotações semânticas para descrever dados geográficos. Os atributos temáticos são correlacionados a ontologias de domínio geográfico, garantido as vantagens de seu uso.

Entretanto, a grande maioria dos sistemas utilizam anotações livres, baseadas em *folksonomias*, como é o caso do Wikimapia e GeoNote. Há ainda alguns dados geográficos ligados: GeoNames, LGD e o Geo Linked Data. Consultas SPARQL podem ser realizadas em dados ligados espaço-temporais publicados em *end-points*, utilizando repositórios como Strabon (KYZIRAKOS; KARPATHIOTAKIS; KOUBARAKIS, 2012), Parliament[33] e Virtuoso[34].

[33]Disponível em: <`http://http://parliament.semwebcentral.org/`>. Acesso em nov. 2013.

[34]Disponível em: <`http://http://virtuoso.openlinksw.com/`>. Acesso em nov. 2013.

3 CONVERSÃO DE DADOS GEOGRÁFICOS ANOTADOS EM LINKED DATA

Este capítulo apresenta o problema de conversão de dados geográficos colaborativamente anotados em *Linked Data* e propõe um método para solucioná-lo. Tal método envolve extração, limpeza, transformação, enriquecimento semântico, ligação com ontologias existentes e triplificação dos dados.

3.1 DEFINIÇÃO DO PROBLEMA

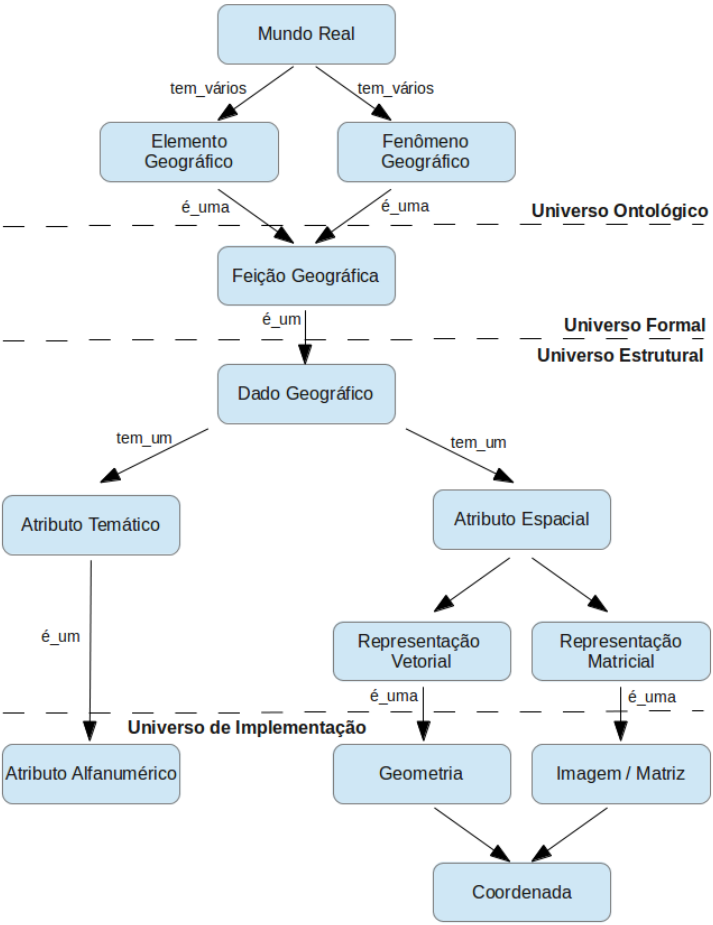
A interpretação do espaço para o homem apresenta inúmeras perspectivas importantes para as aplicações de dados geográficos. Há aplicações que apenas a exibição da imagem do mapa é suficiente. Outras necessitam em de um mapa interativo que possa executar operações espaciais (operações geométricas com os dados espaciais). E ainda há aquelas aplicações que necessitam de anotações para facilitar a interpretação do mapa para o usuário e, consequentemente, a interação do usuário com o mapa.

As diferentes perspectivas do espaço contribuem para o desafio de desenvolver uma aplicação geográfica. As quatro etapas do paradigma dos quatro universos e as diferentes etapas de implementação do SIG estão correlacionadas, como mostra a figura 9. A aplicação deve determinar quais e como elementos e fenômenos do espaço geográfico serão modelados em temas, quais os atributos temáticos de cada tema e se precisa de anotações geográficas para representá-los. A formalização da feição geográfica também depende de qual modelo de anotação foi escolhido. A implementação da aplicação deve considerar como manipular e gerenciar as anotações, e principalmente, como obter anotações. A coleta de dados é cara. A manutenção dos dados deve ser constante.

A coleta colaborativa de dados geográficos reduz os custos de desenvolvimento de aplicações. Entretanto, o sistema de coleta colaborativa deve garantir a qualidade e interoperabilidade de seus dados. A maioria dos sistemas colaborativos ainda falha em ambos os requisitos. Como tornar a utilização de dados colaborativamente coletados viável para as aplicações? Como integrar dados anotados de diferentes sistemas colaborativos?

Fontes de dados geográficos colaborativamente coletados e anotados com semântica bem definida por ontologias, com um sistema de manutenção de dados eficaz e publicado em formatos padrões, são ferramentas de grande utilidade para qualquer aplicação que precisa considerar o significado dos elementos espaciais. Estes dados proporcionam enriquecimento semântico

Figura 9 – **Diagrama explanativo do dado geográfico (Figura criada pelo autor, em 2013)**



em aplicações já existentes e incentivam a exploração de novas aplicações.

Para obter dados geográficos anotados semanticamente em sistemas colaborativos, duas abordagens foram encontradas na literatura. A primeira (GONZALEZ et al., 2013) considera que as anotações são restritas aos conceitos existentes na ontologia do sistema. A segunda (STADLER et al., 2012) realiza a ligação da anotação com a ontologia após a ação de anotar. Este trabalho explora a segunda abordagem, com o intuito de viabilizar coleções de

dados já coletadas para aplicações que exigem uma semântica bem definida. Um sistema colaborativo foi escolhido como caso de estudo, o Wikimapia.

A ligação da anotação com a ontologia ou a conversão da anotação em um conceito de uma ontologia são similares ao alinhamento e/ou mapeamento entre ontologias, e também são tarefas laboriosas que depende de especialistas de domínio. A total automatização deste processo não é viável, pois a interpretação de anotações livres é determinada pela comunidade que utiliza estas anotações. Abordagens semi-automatizadas de alinhamento e/ou mapeamento ontológico são mais populares pois agilizam e facilitam a interação do especialista.

Uma diferença importante da conversão de anotação em um conceito de uma ontologia para o alinhamento e/ou mapeamento entre ontologias é que o conjunto de anotações (*folksonomia*) não possui relações (*e.g.*, hierárquicas, sinônimo) necessariamente explicitas, sendo o rótulo (conteúdo da anotação) geralmente a única representação do conceito vinculado ao objeto anotado. Relações devem ser explicitadas durante o processo de conversão, considerando: características da folksonomia (valores de frequência de uso de anotações), os objetos anotados, os grupos de usuário que utilizam aquela anotação e as regras do domínio (DAMME; HEPP; SIORPAES, 2007).

O problema de correlacionar uma anotação livre com um conceito de uma ontologia já existente pode ser reduzido ao problema de correlacionar o rótulo da anotação livre com o rótulo do conceito da ontologia.

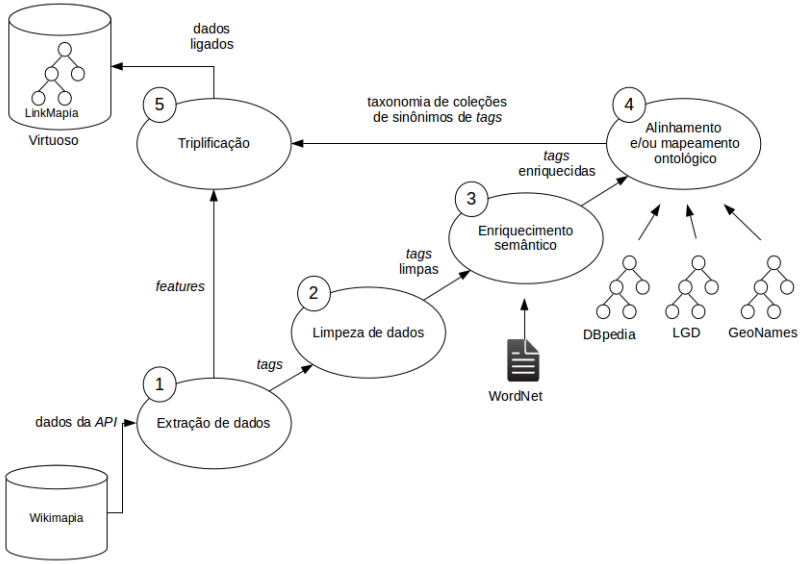
3.2 PROPOSTA

Este trabalho define um processo de conversão de dados geográficos anotados colaborativamente em dados ligados, ilustrado na figura 10. Neste trabalho, o sistema de coleta colaborativa de dados geográficos eleito para a conversão é o Wikimapia.

O processo se inicia com a extração de objetos geográficos (*features*) e anotações livres utilizadas na categorização (categorias) destes objetos no sistema de coleta colaborativa de dados geográficos. Os objetos geográficos geralmente apresentam atributos espaciais, matriciais ou vetorias: ponto, linha e polígono; e atributos descritivos (nome do local representado, descrição, categoria do local). As categorias geralmente apresentam nome e número de objetos geográficos anotados.

Na segunda etapa, a limpeza de dados, o atributo nome das categorias são tratados lexicamente (*e.g.*, *tokenização*, *stemming*, remoção de *stopwords*). A *tokenização* separa nomes compostos em palavras (termos). A remoção de *stopwords* exclui termos irrelevantes (*e.g.*, artigos, preposições)

Figura 10 – **Processo proposto - conversão de dados geográficos anotados em dados ligados (Figura criada pelo autor, em 2013)**



para a análise semântica. O passo de *stemming* reduz lexicamente a radical os termos restantes.

O enriquecimento semântico, a terceira etapa, expande os *tokens* de categorias (resultantes do passo de remoção de *stopwords*) em conjuntos de sinônimos, considerando os diferentes conceitos que o termo representa. A categoria não é mais representada por uma lista de *tokens*, mas por listas de conjuntos de sinônimos. Os conjuntos de sinônimos podem ser obtidos em base de dados léxicos, como o WordNet[1]. Por exemplo, o termo *service* é expandido em seu conjunto de sinômimos:

$$[service, work, assist, assistance, help, activity, care, maintenance, upkeep]$$

A quarta etapa, alinhamento e/ou mapeamento ontológico, relaciona as categorias com conceitos de uma dada ontologia geográfica (*e.g.*, GeoNa-

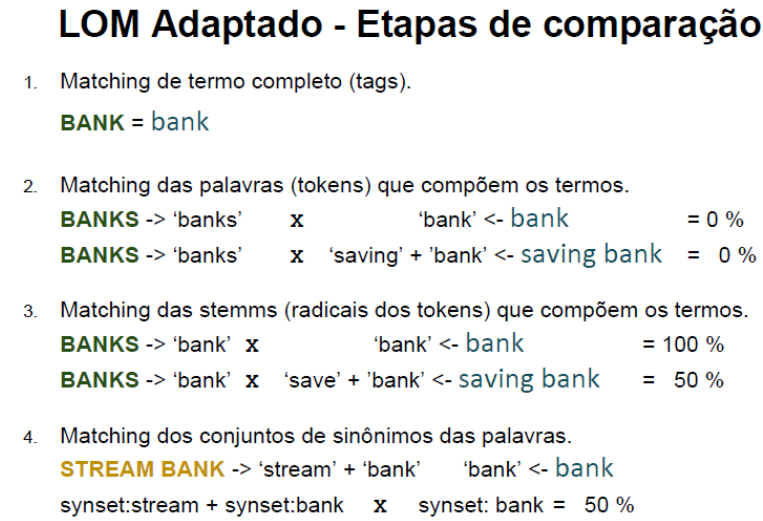
[1] O WordNet é uma grande base léxica da língua inglesa. A base contém sujeitos, adjetivos, verbos e advérbios agrupados em conjuntos de sinônimos que expressam conceitos distintos sendo assim uma ferramenta muito útil para processamento de linguagem natural.Disponível em: `http://wordnet.princeton.edu/`. Acesso em nov. 2013.

mes, LGD). Os rótulos (*label*) dos conceitos da ontologia são comparados com as listas de conjuntos de sinônimos das categorias, obtidas na etapa de enriquecimento semântico. Esta comparação é inspirada do processo de alinhamento ontológico de Li (2004), o LOM.

LOM é uma técnica de *matching* de ontologias que considera apenas a representação léxica (rótulos) de conceitos, desconsiderando a estrutura relacional da ontologia. Em virtude da *folksonomia* não possuir uma estrutura de relacionamentos comparável com a de uma ontologia, é possível adaptar LOM para alinhamento e/ou mapeamento de termos da *folksonomia* para conceitos de ontologias. Este processo pode, inclusive, obter relações adicionais, além da de conceitos equivalentes, que facilitam a conversão da *folksonomia* em uma nova ontologia.

LOM compara os conjuntos dos rótulos dos conceitos das duas ontologias que se pretende alinhar. Tendo em mãos os dois conjuntos, são aplicados quatro processos de pareamento de conceitos: pareamento de termo completo, pareamento de termos (*tokens*) que constituem os termos completos, pareamento de conjunto de sinônimos correspondentes ao termos completos e pareamento de tipos (não explorado nesse trabalho).

Figura 11 – **Fluxo de pareamento entre *folksonomia* e ontologia (Figura criada pelos autores, em 2013)**



Neste trabalho, o pareamento de categoria e conceitos é realizado por diferentes níveis: termo completo, *tokens*, *stemming tokens* e conjunto de

sinônimos (figura 11). O pareamento de termo completo é bastante simples e envolve a comparação do texto completo. Em caso de igualdade o alinhamento é positivo e os termos envolvidos removidos das listas de candidatos.

O pareamento de termos (*tokens*) é realizado utilizando os conjuntos de *tokens* produzidos a partir de cada conceito na etapa de limpeza. Em nossa adaptação, também é realizado o pareamento dos conjuntos de stems dos *tokens* de cada conceito. Essa etapa ocorre em virtude de que radicais das palavras possuem um maior grau de generalização. Em seguida ocorre o pareamento dos conjuntos de sinônimos de cada conceito, obtidos na etapa de enriquecimento semântico.

Ambas as etapas utilizam uma métrica para avaliar a similaridade entre categorias da *folksonomia* (A) e conceitos de ontologia (B), tal qual:

$$sim(A,B) = \frac{quantidade\_termos\_iguais}{TamanhoDoMaiorConjunto(A,B)}$$

A comparação léxica pode utilizar métricas de similaridade léxica (*e.g.*, *soft* TFIDF). Valores de similaridade acima de um determinado nível de confiança A categoria mais similar (categoria alinhada) a um conceito é associada a URI deste conceito para definir o mapeamento ontológico em etapa futura. As categorias menos similares àquele conceito, que respeitem um limite inferior de similaridade, são relacionadas com categoria alinhada como conceitos próximos (mais ou menos abrangentes).

Na quinta e última etapa, triplificação, tanto os objetos geográficos quanto as categorias são convertidas para o formato RDF. Primeiro, as categorias são convertidas gerando uma ontologia rasa (*i.e.*, uma árvore de altitude igual a 1 e raiz igual a *Place*). Segundo, conceitos de categorias alinhadas são mapeados para os respectivos conceitos da ontologia externa. Terceiro, conceitos próximos são relacionados por uma propriedade temporária. Finalmente, os objetos geográficos são convertidos na forma de RDF e associados aos conceitos (pelas URIs) da ontologia formada.

Os dados podem então ser publicados em *endpoint* SPARQL, possibilitando consultas integradas na coleção de dados ligados gerada e nas ontologias externas. O processo de conversão e integração ainda pode ser avaliado e aprimorado considerando as correlações entre as instâncias das coleções de dados ligados.

4 ESTUDO DE CASO: WIKIMAPIA

Neste capítulo é ilustrado como é possível implementar o processo proposto para a conversão de dados geográficos anotados colaborativamente em dados ligados. O sistema de coleta colaborativa de dados geográficos deste caso de estudo é o Wikimapia. Os objetos geográficos são restritos aos limites da cidade de Milão, Itália.

4.1 EXTRAÇÃO DE DADOS

A etapa de extração de dados pode ser dividida na extração dos objetos geográficos e das categorias utilizadas na classificação destes objetos. Este conjunto de categorias forma a *folksonomia* do Wikimapia. Detalhes da extração e caracterização dos dados no anexo A.

4.1.1 Objetos Geográficos

A extração obteve 1276 objetos geográficos anotados. Destes objetos, 6 apresentavam objetos internos (relação *children*), ao todo 42 objetos internos. Os objetos coletados apresentam 1513 relações de anotação, que correlacionam os objetos com um total de 243 categorias (*tags*). A tabela 1 apresenta as 30 categorias mais frequentes em anotações.

4.1.2 *Folksonomia* do Wikimapia

Uma segunda extração (figura 12), objetivando obter todas as categorias utilizadas na classificação de objetos geográficos, resultou em 8615 categorias. De todos os atributos que um *Object* apresenta, aquele que se refere a categoria (*tag*) é a anotação livre, que representa o tipo do lugar ou construção. Este atributo é utilizado para recuperação e restrição de consultas no Wikimapia. O conjunto de categorias forma a *folksonomia* do Wikimapia, criada colaborativamente pelos usuários e em constante evolução.

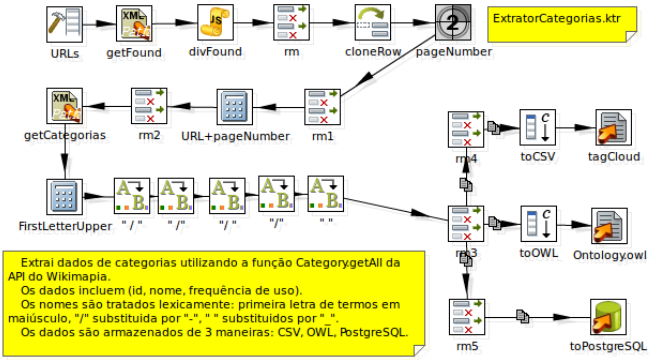
O número de objetos anotados (*amount*) por uma categoria também foi recuperado. A tabela 2 apresenta as 30 categorias mais utilizadas em anotações.

A *folksonomia* geralmente é representada por nuvem de etiquetas, onde a variação de tamanho de um termo é proporcional a sua frequência de uso

NOME	FREQUÊNCIA
DRAW ONLY BORDER	106
PARK	88
SQUARE	80
QUARTERS	63
METRO / SUBWAY / UNDERGROUND STATION	55
UNDERGROUND FACILITY	51
STORE / SHOP	42
CHURCH	41
HISTORICAL LAYER / DISAPPEARED OBJECT	38
PETROL / GAS STATION	34
SCHOOL	31
THIRD-LEVEL ADMINISTRATIVE DIVISION	30
VILLAGE	30
HOTEL	30
OFFICE BUILDING	30
TRAIN STATION	30
PRODUCTION	28
COMMUNE - ADMINISTRATIVE DIVISION	27
CASCINA	25
STADIUM	24
LAKE	21
UNDER CONSTRUCTION	20
UNIVERSITY	20
SUPERMARKET	19
BAR	17
SPORT VENUE	16
CEMETERY	14
APARTMENT BUILDING	14
SWIMMING POOL	14
HOUSE	13

Tabela 1 – **Tabela de categorias de Milão - 30 mais utilizadas em anotações (tabela criada pelo autor, em 2013)**

Figura 12 – **Extrator de categorias (figura criada pelo autor, em 2013)**



em anotações do sistema. A figura 13 foi gerada utilizando a ferramenta Wor-

NOME	FREQUÊNCIA
PLACE WITHOUT PHOTOS	22237340
PLACE WITHOUT DESCRIPTION	14170795
PLACE WITHOUT CATEGORY	13281368
BUILDING WITHOUT ADDRESS	6769359
PLACE WITHOUT POLYGON	3909786
BUILDING	1473509
DWELLING	1401579
HOME	1401579
HOUSE	1401579
RESIDENCE	1401579
VILLA	1401579
PLACE WITH TRIANGULAR POLYGON	1156392
VILLAGE	997893
APARTMENT BUILDING	969453
APARTMENTS	969453
BLOCK OF FLATS	969453
TENEMENT	969453
TOWER BLOCK	969453
SHOPPING AND SERVICES	843255
STORE / SHOP	835422
EDUCATION	680954
LEARNING	680954
LESSONS	680954
SCHOOL	680954
SCHOOLHOUSE	680954
SCHOOLING	680954
TEACHING	680954
DINING AND LEISURE	675989
DEITY	558168
FAITH	558168

Tabela 2 – **Tabela de categorias global - 30 mais utilizadas em anotações de todos os locais do mundo cadastrados no Wikimapia (tabela criada pelo autor, em 2013)**

dle[1], desconsiderando as 5 categorias mais frequentes (*place without photos, place without description, place without category, building without address, place without polygon*), por causa de valores discrepantes de utilização.

4.2 LIMPEZA DE DADOS

A etapa de limpeza de dados foi realizada aplicando a ferramenta de indexação Lucene da fundação Apache[2]. Essa ferramenta não só dispõe de um módulo indexador, mas fornece recursos de análise e tratamento de palavras comumente necessários em processos de limpeza de dados.

Entre os mecanismos de limpeza desejados, foram aplicados sobre os elementos da *folksonomia* e da ontologia destino três processos como vistos

[1]Disponível em: <`http://www.wordle.net`>. Acesso em out. 2013.

[2]Disponível em: `https://lucene.apache.org/`. Acesso em nov. 2013.

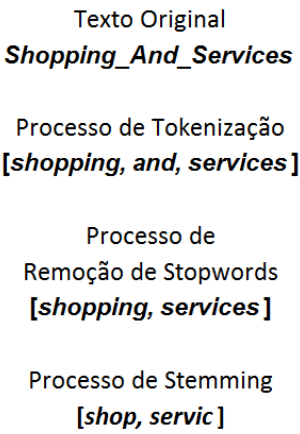
Figura 13 – **Nuvem de anotações da *folksonomia* global (figura criada pelo autor, em 2013)**

na figura 14: *tokenização*, remoção de *stopwords* e *stemming*.

4.3 ENRIQUECIMENTO SEMÂNTICO

A expansão de *tokens* em conjunto de sinônimos foi restrita a apenas termos da língua inglesa e utilizou a base léxica WordNet. Para melhores resultados nas comparações entre termos da ontologia e da *folksonomia*, os *tokens* resultantes do processo de limpeza, em anotações do sistema colaborativo, são submetidos como parâmetros de consulta à base do WordNet e assim são recuperados conjuntos de sinônimos para cada termo. A comunicação com a API do WordNet é realizada pelo conversor.

Figura 14 – **Processos de limpeza das categorias - *tokenização*, remoção de *stopwords* e *stemming* (figura criada por Koerich, W. V., em 2013)**



4.4 ALINHAMENTO E/OU MAPEAMENTO ONTOLÓGICO

O alinhamento das categorias do Wikimapia com conceitos de ontologias (*e.g.*, DBpedia, GeoNames e LGD) apresentada neste trabalho é uma adaptação das idéias apresentadas no artigo sobre LOM (LI, 2004). Foram selecionadas 101 categorias sob o critério de maior número de objetos anotados. Os conceitos de ontologias selecionados foram 143 conceitos do GeoNames[3], já mapeados para conceitos do DBpedia e LGD. A tabela 3 mostra o número de alinhamentos obtidos nos diferentes níveis de pareamento.

RESULTADOS	
NÚMERO DE CATEGORIAS UTILIZADAS DO WIKIMAPIA	101
NÚMERO DE CONCEITOS DE ONTOLOGIAS EXTERNAS	143
NÚMERO DE ALINHAMENTOS POR TEXTO ORIGINAL	15
NÚMERO DE ALINHAMENTO POR TOKENS	7
NÚMERO DE ALINHAMENTO POR STEMMING TOKENS	2
NÚMERO DE ALINHAMENTO POR SYMSET	0
TOTAL DE ALINHAMENTOS	**24**

Tabela 3 – **Tabela de resultados do alinhamento (tabela criada pelo autor, em 2013)**

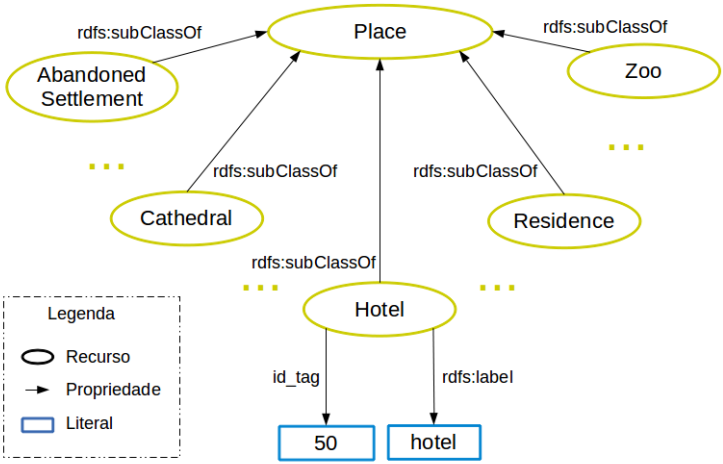
[3]Disponível em: `http://www.geonames.org/ontology/mappings_v3.01.rdf`. Acesso em nov. 2013.

A similaridade entre rótulos é um número real entre 1 (igual) e 0 (sem similaridade). As relações de equivalência e proximidade de conceitos são obtidas classificando os resultados, baseando-se no valor de similaridade dos rótulos (*e.g.*, resultados com valor de similaridade dentro do intervalo (1; 0,9) são considerados idênticos, no intervalo [0,9; 0,5) são considerados similares, e resultados menores que 0,5 não são considerados similares).

4.5 TRIPLIFICAÇÃO

O processo de transformação dos dados do Wikimapia para uma coleção de RDFs, chamada LinkMapia, iniciou-se com a conversão de sua *folksonomia* em uma ontologia rasa, *i.e.* uma árvore de altura 1 (figura 15). Foi escolhida a classe *Place* como derivada direta da classe *Thing*. Todas as categorias da *folksonomia* foram inicialmente classificadas como subclasses de *Place*. Esta etapa foi realizada de forma automatizada, dentro da transformação ExtratorCategorias (figura 12).

Figura 15 – **Ontologia rasa (figura criada pelo autor, em 2013)**



Os passos (*steps*) *FirstLetterUpper*, ”/ ”, ”/”, ”/ ”, ”/”e realizaram um tratamento léxico, caso a caso, em nomes de categorias, tornando-as uti-

lizáveis como nome de URIs. O passo *toOWL* estruturou a informação em triplas similares a particula OWL abaixo (figura 16), gerada automaticamente pela ferramenta Protege, versão 4.3.0, após a modelagem de uma classe exemplo.

Figura 16 – **OWL da categoria (figura criada pelo autor, em 2013)**

```
<owl:Class rdf:about="&om;Hotel">
    <rdfs:subClassOf rdf:resource="&om;Place"/>
    <om:id_tag rdf:datatype="&xsd;int">50</om:id_tag>
    <rdfs:label xml:lang="en">hotel</rdfs:label>
</owl:Class>
```

A coleção de dados resultante ainda não pode ser chamada de dados ligados, visto que a ontologia precisa ser correlacionada com ontologias externas como LGD e GeoNames. O mapeamento da ontologia gerada para as ontologias externas é realizado convertendo as relações de equivalência de rótulos, obtidas na etapa de alinhamento e/ou mapeamento ontológico, em propriedades *rdfs:equivalentClass*, enquanto que relações de proximidade de conceito são convertidas em propriedades temporárias *nearConcept* para, posteriormente, facilitar a estruturação da ontologia pela comunidade do sistema colaborativo, como o caso da identificação da correlação entre *Hotel* e *Motel*, ilustrado na figura 17.

A geração de indivíduos para esta ontologia foi restrita aos dados da cidade de Milão recolhidos na etapa de extração de dados para caracterização do Wikimapia. A transformação GerarIndividuos baseou-se também no RDF gerado pelo Protege resultante de uma modelagem de exemplo (figura 18). A figura 19 ilustra o RDF de exemplo.

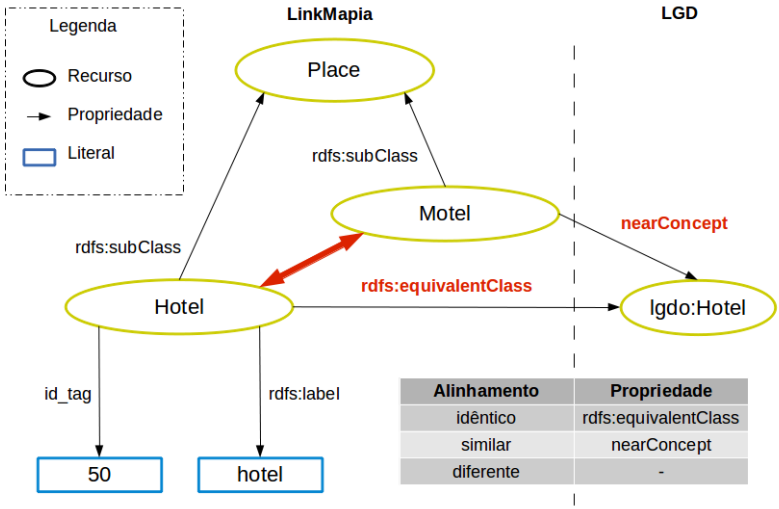
4.5.1 Publicação dos dados ligados

A publicação dos dados ligados foi realizada utilizando o servidor universal Virtuoso[4], que fornece um repositório de triplas RDF e um *endpoint* SPARQL. Virtuoso é um software proprietário, com possibilidade de quinze dias de experimentação. Ele foi escolhido por sua estabilidade e aceitação como repositório RDF, sendo utilizado por DBpedia e LGD.

Os arquivos RDF referentes à ontologia, ao mapeamento para outras

[4]Versão 06.04.3132 32-bits. Disponível em: <`http://virtuoso.openlinksw.com/`>. Acesso em nov. 2013.

Figura 17 – **Triplificação de categoria (figura criada pelo autor, em 2013)**



ontologias e aos dados ligados devem ser publicados na pasta *home* do usuário dentro da organização de arquivos do servidor, utilizando o comando a seguir:

```
curl -i -T <arquivoRDF>
    http://localhost:8890/DAV/home/<user>
          /rdf_sink/<arquivoRDF>
    -u <user>:<password>
```

Os arquivos estão disponíveis para consulta (figura 20) a partir do *endpoint* SPARQL do Virtuoso[5]. Basta endereçar o repositório de consulta para a pasta *home*[6].

4.6 RESULTADOS OBTIDOS

Com a nova coleção de dados LinkMapia, é possível responder perguntas como: **Quais são as opções de locais para dormir (como hotel,**

[5] <http://localhost/sparql/>
[6] <http://localhost:8890/DAV/home/<user>/rdf_sink/>

pensão, albergue) a até quinhentos metros do centro tecnológico da UFSC? Dependendo de como a ontologia foi estruturada pela comunidade, é possível utilizar termos mais abrangentes (locais para dormir) para considerar na consulta uma coleção de conceitos (hotel, pensão, albergue), possibilitando diferentes granularidades para a categoria do objeto espacial.

Consultas podem ser executadas no *endpoint* SPARQL criado na publicação dos dados. A ligação da ontologia de LinkMapia com GeoNames, DBpedia e LGD permite ainda realizar consultas como: **Qual o restaurante conhecido como "Universitário" a até 200 metros do Centro de Eventos?** (figura 21). Esta consulta considera não apenas objetos geográficos do LGD classificados como *lgdo:Restaurant* como também objetos geográficos da Wikimapia anotados como *restaurant*, ou *seafood restaurant*, *drive-in restaurant*, *Subway (restaurant)*, dentre outros.

Figura 18 – **RDF do objeto geográfico (figura criada pelo autor, em 2013)**

```
<owl:NamedIndividual rdf:about="&om;1304">
    <geom:geometry om_geom:1>
    <rdf:type rdf:resource="&om;Cathedral"/>
    <rdf:type rdf:resource="&om;Church"/>
    <rdfs:label>Milan Cathedral</rdfs:label>
    <distance rdf:datatype="&xsd;decimal">1743</distance>
    <south rdf:datatype="&xsd;decimal">45.4638</south>
    <lat rdf:datatype="&xsd;decimal">45.4642</lat>
    <north rdf:datatype="&xsd;decimal">45.4647</north>
    <west rdf:datatype="&xsd;decimal">9.19053</west>
    <lon rdf:datatype="&xsd;decimal">9.19156</lon>
    <east rdf:datatype="&xsd;decimal">9.1926</east>
    <country rdf:datatype="&xsd;string">Italy</country>
    <state rdf:datatype="&xsd;string">Lombardei</state>
    <place rdf:datatype="&xsd;string">Milan</place>
    <description>Il Duomo di Milano is the cathedral of
Milan in Lombardia, Italy. It is the seat of the
Archbishop of Milan, currently His Eminence Dionigi
Cardinal Tettamanzi. The Duomo is famous throughout the
world for its significance in the promulgation of the
Christian faith, for its role in the establishment of
Catholic traditions of worship, its outstanding musical
heritage and the splendour of its Gothic architecture.

It is the second largest Gothic Cathedral in the world,
second to Seville Cathedral in Seville, Spain, and the
second largest church in Italy after St. Peter&apos;s
Basilica in Rome. Rising to a height of 92 meters (35 ft),
it can seat 40,000 people.

http://binged.it/SWnbnU</description>
    <id>1304</id>
    <title>Milan Cathedral</title>
    <wikipedia>
        http://en.wikipedia.org/wiki/Milan_Cathedral
    </wikipedia>
    <child rdf:resource="&om;26494281"/>
    <child rdf:resource="&om;8831198"/>
</owl:NamedIndividual>
```

Figura 19 – **Triplificação de objeto (figura criada pelo autor, em 2013)**

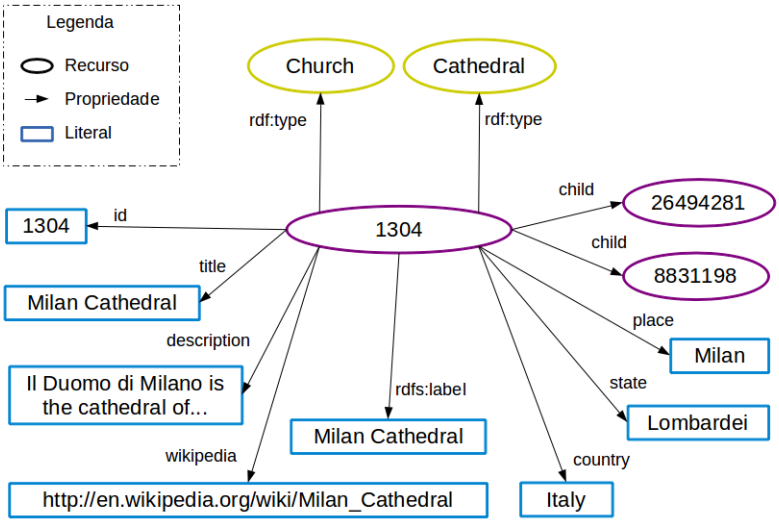


Figura 20 – **Exemplo consulta SPARQL 1 - consulta genérica**

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?subject ?object ?graph
WHERE {
    GRAPH ?graph {
        ?subject rdfs:subClassOf ?object
    }
}
```

Figura 21 – **Exemplo consulta SPARQL 2 - Qual o restaurante conhecido como ”Universitário” a até 200 metros do Centro de Eventos?**

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?local ?nomeLocal {
 ?local
    a lgdo:Restaurant;
    rdfs:label ?nomeLocal ;
    geom:geometry [ ogc:asWKT ?geo ] .
FILTER( regex(?nomeLocal, "Universitário") &&
        bif:st_intersects ( ?geo,
        bif:st_point (-48.52015,-27.60225), 0.2) ).
}
```

5 TRABALHOS CORRELATOS

Os principais trabalhos relacionados encontrados na literatura são o Projeto SEEK[1], o Linked Geo Data (STADLER et al., 2012), GeoLinkedData[2] (Espanha), o Linked Wikimapia[3], o *OurMap* (GONZALEZ et al., 2013), o FolksOntology (DAMME; HEPP; SIORPAES, 2007) e o LOM (LI, 2004).

A iniciativa deste trabalho foi a obtenção de dados geográficos anotados para realizar experimentos de busca por similaridade sêmantica. O enriquecimento semântico dos dados geográficos proporcionam melhor entendimento dos fenômenos e eventos que acontecem nele, como as trajetórias de objetos móveis. Pretendia-se com este trabalho dar apoio a trabalhos futuros do grupo de pesquisa, como o projeto *Semantic Enrichment of trajectory Knowledge discovery* (SEEK).

Este trabalho propôs o enriquecimento semântico de dados coletados colaborativamente. Este processo já havia sido realizado antes, com os dados do projeto OSM (STADLER et al., 2012). Outra coleção de dados geográficos ligados é o GeoLinkedData, uma iniciativa de *Ontology Engineering Group* (OEG) destinada ao enriquecimento da *Web* de dados com dados geoespaciais do território nacional da Espanha.

Existe um projeto de publicação dos dados do Wikimapia como dados ligados, o Linked Wikimapia. Entretanto, atualmente é apenas publicado dados ligados de objetos geográficos, sem nenhuma ontologia para representar a *folksonomia* da Wikimapia. Os dados são ligados com o DBpedia a partir da conversão dos *links* já existentes para páginas do Wikipedia.

Uma diferente abordagem é seguida no sistema de coleta colaborativa *OurMap* (GONZALEZ et al., 2013). Este sistema de coleta colaborativa de dados geográficos já considera o processo de gerenciamento de conhecimento em documentos inteligentes, adaptando-o para dados geográficos. Este sistema utiliza anotações semânticas e o usuário tem a opção de gerenciar as ontologias.

O processo de derivação de uma ontologia a partir de uma *folksonomia* é discutido por Damme, Hepp e Siorpaes (2007), que inspirou o processo proposto. O processo de alinhamento ontológico de Li (2004), LOM, foi adaptado para folksonomias neste trabalho.

[1]Disponível em: `http://www.seek-project.eu/`. Acesso em nov. 2013.
[2]Disponível em: `http://geo.linkeddata.es/web/guest/home`. Acesso em nov. 2013.
[3]Disponível em: `http://openeanwrap.appspot.com/`. Acesso em nov. 2013.

6 CONCLUSÕES

Este capítulo apresenta conclusões do Em. TCC trabalhos futuros, são apresentadas as propostas: a conversão integral dos dados do Wikimapia para dados ligados; a expansão da ontologia do Wikimapia utilizando a mineração de texto das anotações *description*; a análise de qualidade dos dados do Wikimapia segundo (MOONEY et al., 2011); o desenvolvimento de um *data warehouse* (DW) semântico cuja dimensão espacial utiliza os dados geográficos ligados do Wikimapia e do OSM; e a mineração de trajetórias de objetos móveis considerando o enriquecimento semântico da dimensão espacial.

Este trabalho inicialmente apresenta uma breve pesquisa sobre anotações colaborativas de dados geográficos, apresentando as limitações semânticas de anotações livres em comparação aos artefatos da *web* semântica. Em seguida, propõe um processo de conversão de *folksonomia* em ontologia, possibilitando a publicação de dados geográficos anotados em sistemas colaborativos da web como dados ligados, baseando-se em alguns trabalhos relacionados.

Posteriormente, exemplifica o processo proposto como parte dos dados de um sistema colaborativo sem suporte a dados ligados, o Wikimapia, contribuindo com a extensão da coleção de dados geográficos ligados. Este trabalho ainda apresenta uma análise da semântica presente nos dados acessíveis do sistema colaborativo escolhido, apresentando histogramas e algumas consultas para caracterização dos dados.

Entretanto, este trabalho apenas apresenta o assunto abordado, sem definir a melhor forma de converter folksonomia em ontologia. Pesquisas mais maduras, análises de outros sistemas colaborativos, elaboração e implementação de novas propostas de processo e análises de novos casos de estudo são fundamentais para aperfeiçoar o processo de converção. Ainda assim, a viabilidade da elaboração do processo proposto, implementação e seus resultados preliminares demonstram como o trabalho é promissor para refinamentos.

Este trabalho não teve pretensão de realizar a publicação total do Wikimapia, visto que seu principal objetivo foi de fundamentar trabalhos futuros realizados durante o mestrado.

6.1 TRABALHOS FUTUROS

Os resultados deste trabalho podem ser interpretados de diferentes maneiras: processo de conversão de *folksonomia* em ontologia, expansão do

sistema colaborativo Wikimapia, expansão da coleção de dados geográficos ligados.

Independente do trabalho futuro a ser seguido, uma fundamentação mais madura sobre sistemas de coleta colaborativa, *folksonomia* e como que artefatos da *web* semântica podem auxiliar no enriquecimento semântico de dados encontrados nestes sistemas é essencial, visto que são assuntos ainda não consolidados.

6.1.1 Processo de conversão de *folksonomia* em ontologia

O processo proposto deve ser aprimorado, considerando trabalhos relacionados ainda não abordados. É necessário realizar uma revisão sistemática sobre a conversão de *folksonomia* em ontologia. Na etapa de enriquecimento semântico, devem ser considerados também anotações em línguas diferentes de inglês, sendo necessário métodos de tradução ou bases léxicas em outras línguas. Também deve-se identificar, na etapa de alinhamento e/ou mapeamento ontológico, a métrica de similaridade mais apropriada para o processo e considerar comparações léxicas mais complexas como o *soft* TFIDF (COHEN; RAVIKUMAR; FIENBERG, 2003). A etapa semi-automátizada deve ser determinada, utilizando técnicas como *gameficação* (THALER; SIMPERL; SIORPAES, 2011) para atrair a participação dos usuários do sistema de coleta colaborativa.

6.1.2 Expansão do sistema de coleta colaborativa

A arquitetura de sistemas de coleta colaborativa deve ser expandida para inserção do processo de conversão de *folksonomia* em ontologia, visando possibilitar a coexistência de anotações livres e anotações semânticas como proposto por Christiaens (2006). Deste modo o usuário não estará restrito a utilizar a apenas conceitos da ontologia para realizar anotações, e mesmo assim o sistema de coleta colaborativa será capaz de gerar dados com anotações de semântica bem definida. É prevista a sincronização periódica da *folksonomia* com a ontologia e vice-versa. Esta nova arquitetura também deve considerar a evolução ontológica, moderação e privacidade dos dados anotados.

6.1.3 Expansão da coleção de dados geográficos ligados

O desenvolvimento de um DW semântico com os dados convertidos neste trabalho é essencial para a análise das informações fornecidas e obtenção de conhecimento. A anotação dos dados geográficos proporciona a expansão da granularidade da dimensão espacial (especificamente, a categoria do objeto espacial) ao nível de detalhamento das entidades da ontologia usada na anotação. Um DW oferece uma estrutura ideal para a aplicação de algoritmos de mineração por causa da utilização do modelo estrela. Diferente do modelo relacional, o modelo estrela valoriza mais as operações de consulta e recuperação de informação que as de criação, modificação e deleção. A anotação de dados geográfico proporciona novas oportunidades de aplicações de mineração de dados, como por exemplo a mineração de subestruturas frequentes, localização de subestruturas e e a busca por similaridade de entidades correlacionadas, aplicações conhecidas da mineração de grafos.

A mineração de trajetórias de objetos móveis estuda o comportamento de objetos móveis no espaço geográfico. O processo de descoberta de conhecimento em trajetórias de objetos móveis é similar ao processo de descoberta de conhecimento e mineração de dados (*Knowledge Discovery and Data mining* - KDD). O enriquecimento semântico da dimensão espacial pode contribuir para o enriquecimento semântico da trajetória dos objetos móveis. A motivação, ou *goal*, da trajetória pode estar presente na descrição dos lugares visitados pelo objeto móvel.

REFERÊNCIAS

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: Scientific American. **Scientific American**, maio 2001. Disponível em: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&pageNumber=1&catID=2>.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data - the story so far. **International Journal on Semantic Web and Information Systems**, v. 5, n. 3, p. 1–22, 2009. ISSN 1552-6283, 1552-6291. Disponível em: <http://www.igi-global.com/article/linked-data-story-far/37496>.

BONTCHEVA, K.; WILKS, Y. Automatic report generation from ontologies: The miakt approach. In: **NLDB**. [S.l.: s.n.], 2004. p. 324–335.

CâMARA, G. **Modelos, Linguagens e Arquiteturas para Bancos de Dados Geográficos**. Tese (Doutorado), 1995. Disponível em: <http://www.dpi.inpe.br/teses/gilberto/>. Acesso em: 13 de Maio de 2013.

CâMARA, G.; MONTEIRO, A. M. V.; MEDEIROS, J. S. d. (Ed.). **Introdução à Ciência de Geoprocessamento**. São José dos Campos: INPE, 2004. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/introd>. Acesso em: 17 de Abril de 2013.

CASANOVA, M. et al. (Ed.). **Bancos de Dados Geográficos**. MundoGEO, 2005. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/bdados/capitulos.html>. Acesso em: 17 de Abril de 2013.

CHRISTIAENS, S. Metadata mechanisms: From ontology to folksonomy... and back. In: **Lecture Notes in Computer Science**. [S.l.: s.n.], 2006. p. 199–207.

COHEN, W. W.; RAVIKUMAR, P. D.; FIENBERG, S. E. A comparison of string distance metrics for name-matching tasks. In: KAMBHAMPATI, S.; KNOBLOCK, C. A. (Ed.). **IIWeb**. [S.l.: s.n.], 2003. p. 73–78.

DAMME, C. V.; HEPP, M.; SIORPAES, K. Folksontology: An integrated approach for turning folksonomies into ontologies. In: **IN PROCEEDINGS OF THE ESWC WORKSHOP BRIDGING THE GAP BETWEEN SEMANTIC WEB AND WEB 2.0**. [S.l.]: Springer, 2007.

DELCAMBRE, L. M. L.; MAIER, D. Models for superimposed information. In: **Proceedings of the Workshops on Evolution and Change in Data Management, Reverse Engineering in Information Systems, and the World Wide Web and Conceptual Modeling**. London, UK, UK: Springer-Verlag, 1999. (ER '99), p. 264–280. ISBN 3-540-66653-2. Disponível em: <http://dl.acm.org/citation.cfm?id=647523.728336>.

DOWMAN, M. et al. Web-assisted annotation, semantic indexing and search of television and radio news. In: **Proceedings of the 14th International World Wide Web Conference**. [S.l.]: ACM Press, 2005. p. 05.

FONSECA, F. T.; EGENHOFER, M. J. **Ontology-Driven Geographic Information Systems**. 1999.

FRAPPAOLO, C. et al. **The document is the process**. [S.l.], 06 1994. Disponível em: <http://www.delphigroup.com/whitepapers/pdf/DocIsProcess.pdf>. Acesso em: 27 de Abril de 2013.

FRITZ, S. et al. Geo-wiki.org: The use of crowdsourcing to improve global land cover. **Remote Sensing**, v. 1, n. 3, p. 345–354, 2009. ISSN 2072-4292. Disponível em: <http://www.mdpi.com/2072-4292/1/3/345>.

GARCíA-CASTRO, L. J.; GARCíA, E. **Folksonomies behind the scenes**. 2011.

GIL, F. B.; KOZIEVITCH, N. P.; TORRES, R. da S. Geonote: A web service for geographic data annotation in biodiversity information systems. **JIDM**, v. 2, n. 2, p. 195–210, 2011. Disponível em: <http://dblp.uni-trier.de/db/journals/jidm/jidm2.html#GilKT11>.

GONZALEZ, A. et al. Representação Aberta e Semântica de Anotações de Incidentes em Mapas Web. In: **Simpósio Brasileiro de Sistemas Multimídia e Web, 2013, João Pessoa. Anais do IX Simpósio Brasileiro de Sistemas Multimídia e Web (por aparecer)**. [S.l.: s.n.], 2013. v. 1, p. 1–12.

GOODCHILD, M. F. Citizens as voluntary sensors: spatial data infrastructure in the world of web 2.0. **International Journal of Spatial Data Infrastructures Research**, v. 2, p. 24–32, 2007.

GUARINO, N. Formal ontology and information systems. In: . [S.l.]: IOS Press, 1998. p. 3–15.

HUNTER, J. et al. Using the semantic grid to build bridges between museums and indigenous communities. In: **Proceedings of the GGF11-Semantic Grid Applications Workshop**. [s.n.], 2004. p. 46–61. Disponível em: <http://metadata.net/filmed/pub/GGF11%5FSemanticGrid.pdf>.

JOHNSON, R. **GIS Technology for Disasters and Emergency Management**. [S.l.], 05 2000.

KAHAN, J. et al. **Annotea: An Open RDF Infrastructure for Shared Web Annotations**. 2001.

KIM, J. dong et al. GENIA corpus - a semantically annotated corpus for bio-textmining. In: **Intelligent Systems in Molecular Biology**. [S.l.: s.n.], 2003. v. 19, p. 180–182.

KYZIRAKOS, K.; KARPATHIOTAKIS, M.; KOUBARAKIS, M. Strabon: A semantic geospatial dbms. In: CUDRé-MAUROUX, P. et al. (Ed.). **International Semantic Web Conference (1)**. [S.l.]: Springer, 2012. (Lecture Notes in Computer Science, v. 7649), p. 295–311. ISBN 978-3-642-35175-4.

LI, J. Lom: A lexicon-based ontology mapping tool. In: **Proceedings of the Performance Metrics for Intelligent Systems (PerMIS**. [S.l.: s.n.], 2004. p. 2004.

MAGALHÃES, W. G. et al. **Noções Básicas de Cartografia**. 1998. Disponível em: <http://www.ibge.gov.br/home/geociencias/cartografia/manual_nocoes/representacao.html>. Acesso em: 17 de Abril de 2013.

MALINOWSKI, E.; ZIMNYI, E. **Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications (Data-Centric Systems and Applications)**. 1. ed. [S.l.]: Springer Publishing Company, Incorporated, 2008. ISBN 3540744045, 9783540744047.

MAYNARD, D. et al. **Automatic Creation and Monitoring of Semantic Metadata in a Dynamic Knowledge Portal**. [S.l.: s.n.], 2004. 65–74 p.

MOONEY, P. et al. Citizen-generated spatial data and information: Risks and opportunities. In: **Intelligence and Security Informatics**. [S.l.: s.n.], 2011.

OGC. Go-1 application objects. **Open Geospatial Consortium Inc.**, p. 111–115, 2005. Disponível em: <http://portal.opengeospatial.org/files/?artifact_id=10378>. Acesso em: 13 de Maio de 2013.

OGC. Opengis® implementation standard for geographic information - simple feature access - part 1: Common architecture. **Open Geospatial Consortium Inc.**, p. 9, 2011. Disponível em: <http://portal.opengeospatial.org/files/?artifact_id=25355>. Acesso em: 13 de Maio de 2013.

O'REILLY, T. What is web 2.0: Design patterns and business models for the next generation of software. 2005. Disponível em: <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>.

PLESSERS, P. et al. Accessibility: a web engineering approach. In: **WWW '05: Proceedings of the 14th international conference on World Wide Web**. [S.l.]: ACM Press, 2005. p. 353–362.

RIGAUX, P.; SCHOLL, M.; VOISARD, A. **Introduction to Spatial Databases: Applications to GIS**. [S.l.]: Morgan Kaufmann, 2000. ISBN 1-55860-689-0.

SCHARL, A.; TOCHTERMANN, K. (Ed.). **The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society (Advanced Information and Knowledge Processing)**. 1. ed. Springer, 2007. Hardcover. ISBN 1846288266. Disponível em: <http://www.worldcat.org/isbn/1846288266>.

SIMON, R.; JUNG, J.; HASLHOFER, B. The yuma media annotation framework. In: GRADMANN, S. et al. (Ed.). **TPDL**. Springer, 2011. (Lecture Notes in Computer Science, v. 6966), p. 434–437. ISBN 978-3-642-24468-1. Disponível em: <http://dblp.uni-trier.de/db/conf/ercimdl/tpdl2011.html#SimonJH11>.

SMITH, J. **Collective Intelligence in Computer-Based Collaboration**. LAWRENCE ERLBAUM ASSOC Incorporated, 1994. (Computers, Cognition, and Work). ISBN 9780805813197. Disponível em: <http://books.google.com.br/books?id=Mx0MuBf0jG0C>.

STADLER, C. et al. Linkedgeodata: A core for a web of spatial open data. 2012. Disponível em: <http://svn.aksw.org/papers/2011/SWJ_LinkedGeoData/public.pdf>. Acesso em: 13 de Novembro de 2013.

SVAB, O.; LABSKY, M.; SVATEK, V. **RDF-Based Retrieval of Information Extracted from Web Product Catalogues**. 2004.

THALER, S.; SIMPERL, E. P. B.; SIORPAES, K. Spotthelink: A game for ontology alignment. In: MAIER, R. (Ed.). **Wissensmanagement**. GI, 2011. (LNI, v. 182), p. 246–253. ISBN 978-3-88579-276-5. Disponível em: <http://dblp.uni-trier.de/db/conf/wm/wm2011.html#ThalerSS11>.

TURNER, A. **Introduction to neogeography**. First. [S.l.]: O'Reilly, 2006. ISBN 0596529953.

UREN, V. et al. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. **Web Semant.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 4, n. 1, p. 14–28, jan. 2006. ISSN 1570-8268. Disponível em: <http://dx.doi.org/10.1016/j.websem.2005.10.002>.

WAL, T. V. **Explain and Showing Broad and Narrow Folksonomies**. 2005. Disponível em: <http://vanderwal.net/random/entrysel.php?blog=1635>. Acesso em: 13 de Novembro de 2013.

WAL, T. V. **Folksonomy**. 2007. Disponível em: <http://vanderwal.net/folksonomy.html>. Acesso em: 13 de Novembro de 2013.

WELTY, C.; IDE, N. Using the right tools: Enhancing retrieval from marked-up documents. **Computers and the Humanities**, Kluwer Academic Publishers, v. 33, n. 1-2, p. 59–84, 1999. ISSN 0010-4817. Disponível em: <http://dx.doi.org/10.1023/A%3A1001800717376>.

APÊNDICE A – LinkMapia: An approach to convert volunteer geographic information to linked geographic data collection

LinkMapia: An approach to convert volunteer geographic information to linked geographic data collection

Juarez A. P. Sacenti[1], **Willian Ventura Koerich**[1]

[1]Departamento de Informática e Estatística – Universidade Federal do Santa Catarina (UFSC)
Caixa Postal 476 – 88040-900 – Florianópolis – SC – Brazil

{juarezsacenti,willian.vkoerich}@inf.ufsc.br

Abstract. *The collection of geographic data using satellites and global positioning system (GPS) is insufficient to gather information on how man interprets the space. Several approaches have been proposed, including patterns of annotation and geographic information crowdsourcing. However, the data collected in this way do not have enough formal semantics for use computationally and sufficiently detailed to meet the needs of applications. The semantic web solutions, between them semantic annotations, can contribute to solving the problems of ambiguity and interoperability of collaborative annotations. This paper propose an approach to convert volunteer geographic information from Wikimapia project, a collaborative system which was not yet connected to web data, to linked geographic data collection.*

Resumo. *A coleta de dados geográficos utilizando satélites e dispositivos com sistema de posicionamento global (GPS) é insuficiente para levantar informações de como o homem interpreta o espaço. Diversas abordagens têm sido propostas, incluindo padrões de anotação e coleta colaborativa (crowdsourcing) de informação geográfica. Contudo, os dados coletados desta forma não apresentam semântica suficientemente formal para ser utilizada computacionalmente e suficientemente detalhada para suprir as necessidades das aplicações. As soluções da web semântica, como as anotações semânticas, ontologias e dados ligados (linked data), podem contribuir para resolver os problemas de ambiguidade e interoperabilidade das anotações colaborativas. Este trabalho propõe uma abordagem para conversão de dados geográficos coletados voluntariamente do projeto Wikimapia, um sistema colaborativo ainda não conectado a web de dados, em uma coleção de dados geográficos ligados.*

1. Introdução

Os dados associados a caracteríticas espaciais, que referenciam uma localização na superfície da Terra, são chamados de dados geográficos. O espaço ocupado por elementos (*e. g.*, construções, rios e estradas) e fenômenos (*e. g.*, massas de ar, dissiminação de doenças e temperatura) são representados computacionalmente por esses dados, respectivamente por formas vetoriais (*e. g.*, pontos, retas, polígonos) e matriciais (*e. g.*, imagens *raster*). Estas representações são consideradas os atributos espaciais dos dados geográficos.

Porém, dados geográficos possuem informações alfanuméricas, tão importantes quanto os atributos espaciais: os atributos descritivos. A descrição do dado geográfico

depende da interpretação humana do elemento ou fenômeno representado. A coleta de dados geográficos utilizando satélites e dispositivos com sistema de posicionamento global (GPS) é insuficiente para levantar estas informações.

Sistemas de coleta colaborativa de dados geográficos (*e. g.*, Wikimapia, *Open Street Map* - OSM) são providos de recursos para permitir a coleta de dados geográficos, principalmente atributos descritivos, geralmente de forma voluntária e via *web*. Entretanto, dados anotados desta forma (*i. e.*, *volunteer geographic information* - VGI) geralmente não apresentam semântica suficientemente formal para ser utilizada computacionalmente e suficientemente detalhada para suprir as necessidades das aplicações. As anotações livres, *i. e.*, anotações que não apresentam nenhum tipo de estruturação explicita de seu conteúdo (*e. g.*, vocabulário comum, glossário), herdam os problemas da manipulação de textos livres: sinônimos, ambiguidades e erros ortográficos.

Nestes sistemas sociais de anotação (*Social Tagging Systems* - STS), algumas anotações livres são utilizadas para posteriormente recuperar os objetos anotados. A estrutura conceitual destes tipos de anotação, emergente de STSs, é chamada de folksonomia.

A adoção de convenções de anotação, *e. g.*, enciclopédias (*thesaurus*), glossários, dicionários geográficos (*gazetters*), é dificil de ser mantida em sistemas com grande número de voluntários. As soluções da *web* semântica, como as anotações semânticas, ontologias e dados ligados (*linked data*), podem contribuir para resolver os problemas de ambiguidade e interoperabilidade das anotações colaborativas.

Para obter dados ligados de um sistema colaborativo, é necessário converter a folksonomia em ontologia. Este trabalho propõe uma abordagem para conversão de dados geográficos coletados voluntariamente do projeto Wikimapia, um sistema colaborativo ainda não conectado a *web* de dados, em uma coleção de dados geográficos ligados.

2. Proposta

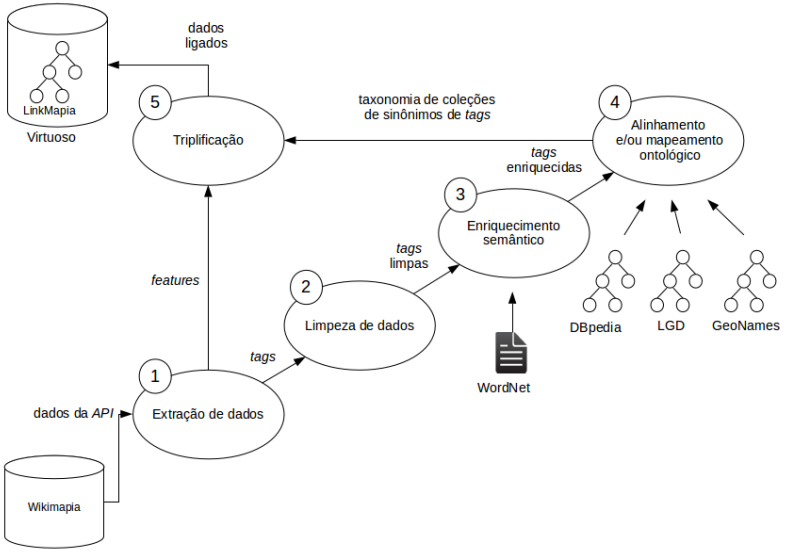
Este trabalho define um processo de conversão de dados geográficos anotados colaborativamente em dados ligados, ilustrado na figura 1. Neste trabalho, o sistema de coleta colaborativa de dados geográficos eleito para a conversão é o Wikimapia.

O processo se inicia com a extração de objetos geográficos (*features*) e anotações livres utilizadas na categorização (categorias) destes objetos no sistema de coleta colaborativa de dados geográficos. Os objetos geográficos geralmente apresentam atributos espaciais, matriciais ou vetorias: ponto, linha e polígono; e atributos descritivos (nome do local representado, descrição, categoria do local). As categorias geralmente apresentam nome e número de objetos geográficos anotados.

Na segunda etapa, a limpeza de dados, o atributo nome das categorias são tratados lexicamente (*e.g.*, *tokenização*, *stemming*, remoção de *stopwords*). A *tokenização* separa nomes compostos em palavras (termos). A remoção de *stopwords* exclui termos irrelevantes (*e.g.*, artigos, preposições) para a análise semântica. O passo de *stemming* reduz lexicamente a radical os termos restantes.

O enriquecimento semântico, a terceira etapa, expande os *tokens* de categorias (resultantes do passo de remoção de *stopwords*) em conjuntos de sinônimos, considerando os diferentes conceitos que o termo representa. A categoria não é mais representada por

Figure 1. Processo proposto - conversão de dados geográficos anotados em dados ligados (Figura criada pelo autor, em 2013)



uma lista de *tokens*, mas por listas de conjuntos de sinônimos. Os conjuntos de sinônimos podem ser obtidos em base de dados léxicos, como o WordNet[1]. Por exemplo, o termo *service* é expandido em seu conjunto de sinômimos:

$$[service, work, assist, assistance, help, activity, care, maintenance, upkeep]$$

A quarta etapa, alinhamento e/ou mapeamento ontológico, relaciona as categorias com conceitos de uma dada ontologia geográfica (*e.g.*, GeoNames, LGD). Os rótulos (*label*) dos conceitos da ontologia são comparados com as listas de conjuntos de sinônimos das categorias, obtidas na etapa de enriquecimento semântico. Esta comparação é inspirada do processo de alinhamento ontológico LOM [Li 2004].

LOM é uma técnica de *matching* de ontologias que considera apenas a representação léxica (rótulos) de conceitos, desconsiderando a estrutura relacional da ontologia. Em virtude da *folksonomia* não possuir uma estrutura de relacionamentos comparável com a de uma ontologia, é possível adaptar LOM para alinhamento e/ou mapeamento de termos da *folksonomia* para conceitos de ontologias. Este processo pode,

[1]O WordNet é uma grande base léxica da língua inglesa. A base contém sujeitos, adjetivos, verbos e advérbios agrupados em conjuntos de sinônimos que expressam conceitos distintos sendo assim uma ferramenta muito útil para processamento de linguagem natural.Disponível em: `http://wordnet.princeton.edu/`. Acesso em nov. 2013.

inclusive, obter relações adicionais, além da de conceitos equivalentes, que facilitam a conversão da *folksonomia* em uma nova ontologia.

LOM compara os conjuntos dos rótulos dos conceitos das duas ontologias que se pretende alinhar. Tendo em mãos os dois conjuntos, são aplicados quatro processos de pareamento de conceitos: pareamento de termo completo, pareamento de termos (*tokens*) que constituem os termos completos, pareamento de conjunto de sinônimos correspondentes ao termos completos e pareamento de tipos (não explorado nesse trabalho).

Figure 2. Fluxo de pareamento entre *folksonomia* e ontologia (Figura criada pelos autores, em 2013)

Neste trabalho, o pareamento de categoria e conceitos é realizado por diferentes níveis: termo completo, *tokens*, *stemming tokens* e conjunto de sinônimos (figura 2). O pareamento de termo completo é bastante simples e envolve a comparação do texto completo. Em caso de igualdade o alinhamento é positivo e os termos envolvidos removidos das listas de candidatos.

O pareamento de termos (*tokens*) é realizado utilizando os conjuntos de *tokens* produzidos a partir de cada conceito na etapa de limpeza. Em nossa adaptação, também é realizado o pareamento dos conjuntos de stems dos *tokens* de cada conceito. Essa etapa ocorre em virtude de que radicais das palavras possuem um maior grau de generalização. Em seguida ocorre o pareamento dos conjuntos de sinônimos de cada conceito, obtidos na etapa de enriquecimento semântico.

Ambas as etapas utilizam uma métrica para avaliar a similaridade entre categorias da *folksonomia* (A) e conceitos de ontologia (B), tal qual:

$$sim(A,B) = \frac{quantidade\_termos\_iguais}{TamanhoDoMaiorConjunto(A,B)}$$

A comparação léxica pode utilizar métricas de similaridade léxica (*e.g.*, *soft* TFIDF). Valores de similaridade acima de um determinado nível de confiança A cate-

goria mais similar (categoria alinhada) a um conceito é associada a URI deste conceito para definir o mapeamento ontológico em etapa futura. As categorias menos similares àquele conceito, que respeitem um limite inferior de similaridade, são relacionadas com categoria alinhada como conceitos próximos (mais ou menos abrangentes).

Na quinta e última etapa, triplificação, tanto os objetos geográficos quanto as categorias são convertidas para o formato RDF. Primeiro, as categorias são convertidas gerando uma ontologia rasa (*i.e.*, uma árvore de altitude igual a 1 e raiz igual a *Place*). Segundo, conceitos de categorias alinhadas são mapeados para os respectivos conceitos da ontologia externa. Terceiro, conceitos próximos são relacionados por uma propriedade temporária. Finalmente, os objetos geográficos são convertidos na forma de RDF e associados aos conceitos (pelas URIs) da ontologia formada.

Os dados podem então ser publicados em *endpoint* SPARQL, possibilitando consultas integradas na coleção de dados ligados gerada e nas ontologias externas. O processo de conversão e integração ainda pode ser avaliado e aprimorado considerando as correlações entre as instâncias das coleções de dados ligados.

3. Desenvolvimento

Nesta seção é ilustrado como é possível implementar o processo proposto para a conversão de dados geográficos anotados colaborativamente em dados ligados. O sistema de coleta colaborativa de dados geográficos deste caso de estudo é o Wikimapia. Os objetos geográficos deste caso de estudo são restritos aos limites da cidade de Milão, Itália.

3.1. Extração de dados

A etapa de extração de dados pode ser dividida na extração dos objetos geográficos e das categorias utilizadas na classificação destes objetos. Este conjunto de categorias forma a *folksonomia* do Wikimapia.

3.1.1. Objetos Geográficos

Como referido anteriormente, é possivel obter dados da base do Wikimapia a partir de sua API. Para fins de praticidade, a caracterização da base foi limitada espacialmente a um estudo de caso: a cidade de Milão. O menor retângulo que contém a cidade de Milão (*minimal bounding rectangle*) é definido pelos pontos (45.388039, 9.043907) e (45.536266, 9.278963), de latitude e longitude respectiva, segundo o polígono dos limites políticos da cidade na base de dados *Database of Global Administrative Areas* (GADM).

A extração de objetos geográficos (features) utiliza as funções *box*[2] e *object*[3] da API da Wikimapia. A função *box*, com esses parâmetros, resultou em uma lista 1276 ids de objetos, separados por páginas, sendo a váriável *count* o limite de resultados por página, em formato XML ou JSON. A função *object* obteve a descrição detalhada de cada objeto, dado seu *id*. Os dados extraídos apresentaram 243 categorias e 1513 relações de

[2]Disponível em: <http://api.wikimapia.org/?function=box&lon_min=9.043907&lat_min=45.388039&lon_max=9.278963&lat_max=45.536266&rsaquo&key=example&page=1&count=100&disable=location,polygon>. Acesso em jan. 2013.

[3]Disponível em: <http://api.wikimapia.org/?function=object&key=example&id=ID_DO_OBJETO>. Acesso em jan. 2013.

anotação. O SGBD utilizado para o armazenamento dos dados extraídos foi o PostgreSQL 9.2.4 com a extensão geográfica PostGIS 2.0.3-1[4].

3.1.2. Folksonomia do Wikimapia

A função *Category.GetAll*[5] retorna todas as categorias do Wikimapia, compondo a *folksonomia* do sistema colabborativo. A extração obteve 8615 categorias: *ids*, nomes e frequência de uso (tabela 1).

ELEMENTO	DESCRIÇÃO	TIPO	RESTRIÇÃO	CLASSE
ID	IDENTIFICADOR	NUM.	NÃO NULO	SIMPLES
AMOUNT	NÚMERO DE LUGARES CLASSIFICADOS POR ESTA CATEGORIA.	NUM.	NÃO NULO	SIMPLES
ICON	ENDEREÇO DA IMAGEM DA CATEGORIA.	URL	-	SIMPLES
NAME	NOME DA CATEGORIA.	TEXTO	NÃO NULO	SIMPLES

Table 1. Dicionário de dados - Elemento retornado pela função Categoria: Categoria utilizada para classificar objetos criados por usuários do Wikimapia (tabela criada pelo autor, em 2013)

A *folksonomia* geralmente é representada por nuvem de palavras, onde a variação de tamanho de um termo é proporcional a sua frequência de uso em anotações do sistema. A figura 3 foi gerada utilizando a ferramenta Wordle[6], desconsiderando as 5 categorias mais frequentes (*place without photos, place without description, place without category, building without address, place without polygon*).

Figure 3. Nuvem de anotações da folksonomia (figura criada pelo autor, em 2013)

[4] *dump* do banco em anexo

[5] Disponível em: <http://api.wikimapia.org/?key=example&function=category.getall&format=&pack=&language=en&name=&page=1&count=50>. Acesso em set. 2013.

[6] Disponível em: <http://www.wordle.net>. Acesso em out. 2013.

3.2. Limpeza de dados

A etapa de limpeza de dados foi realizada aplicando a ferramenta de indexação Lucene da fundação Apache[7]. Essa ferramenta não só dispõe de um módulo indexador, mas fornece recursos de análise e tratamento de palavras comumente necessários em processos de limpeza de dados.

Entre os mecanismos de limpeza desejados, foram aplicados sobre os elementos da folksonomia e da ontologia destino três processos: *tokenização*, remoção de *stopwords* e *stemming*.

3.3. Enriquecimento semântico

A expansão de *tokens* em conjunto de sinônimos foi restrita a apenas termos da língua inglesa e utilizou a base léxica WordNet. Para melhores resultados nas comparações entre termos da ontologia e da *folksonomia*, os *tokens* resultantes do processo de limpeza, em anotações do sistema colaborativo, são submetidos como parâmetros de consulta à base do WordNet e assim são recuperados conjuntos de sinônimos para cada termo. A comunicação com a API do WordNet é realizada pelo conversor.

3.4. Alinhamento e/ou Mapeamento Ontológico

O alinhamento das categorias do Wikimapia com conceitos de ontologias (*e.g.*, DBpedia, GeoNames e LGD) apresentada neste trabalho é uma adaptação das idéias apresentadas no artigo sobre LOM [Li 2004]. Foram selecionadas 101 categorias sob o critério de maior número de objetos anotados. Os conceitos de ontologias selecionados foram 143 conceitos do GeoNames[8], já mapeados para conceitos do DBpedia e LGD. A tabela 2 mostra o número de alinhamentos obtidos nos diferentes níveis de pareamento.

RESULTADOS	
NÚMERO DE CATEGORIAS UTILIZADAS DO WIKIMAPIA	101
NÚMERO DE CONCEITOS DE ONTOLOGIAS EXTERNAS	143
NÚMERO DE ALINHAMENTOS POR TEXTO ORIGINAL	15
NÚMERO DE ALINHAMENTO POR TOKENS	7
NÚMERO DE ALINHAMENTO POR STEMMING TOKENS	2
NÚMERO DE ALINHAMENTO POR SYMSET	0
TOTAL DE ALINHAMENTOS	**24**

Table 2. Tabela de resultados do alinhamento (tabela criada pelo autor, em 2013)

A similaridade entre rótulos é um número real entre 1 (igual) e 0 (sem similaridade). As relações de equivalência e proximidade de conceitos são obtidas classificando os resultados, baseando-se no valor de similaridade dos rótulos (*e.g.*, resultados com valor de similaridade dentro do intervalo (1; 0,9) são considerados idênticos, no intervalo [0,9; 0,5) são considerados similares, e resultados menores que 0,5 não são considerados similares).

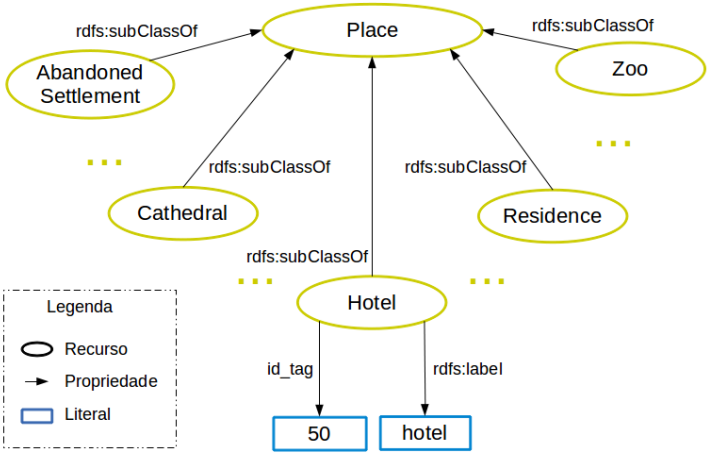
[7]Disponível em: `https://lucene.apache.org/`. Acesso em nov. 2013.

[8]Disponível em: `http://www.geonames.org/ontology/mappings_v3.01.rdf`. Acesso em nov. 2013.

3.5. Triplificação

O processo de transformação dos dados do Wikimapia para uma coleção de RDFs, chamada LinkMapia, iniciou-se com a conversão de sua *folksonomia* em uma ontologia rasa, *i.e.* uma árvore de altura 1 (figura 4). Foi escolhida a classe *Place* como derivada direta da classe *Thing*. Todas as categorias da *folksonomia* foram inicialmente classificadas como subclasses de *Place*. O esquema OWL foi baseado em um protótipo gerado automaticamente pela ferramenta Protege, versão 4.3.0.

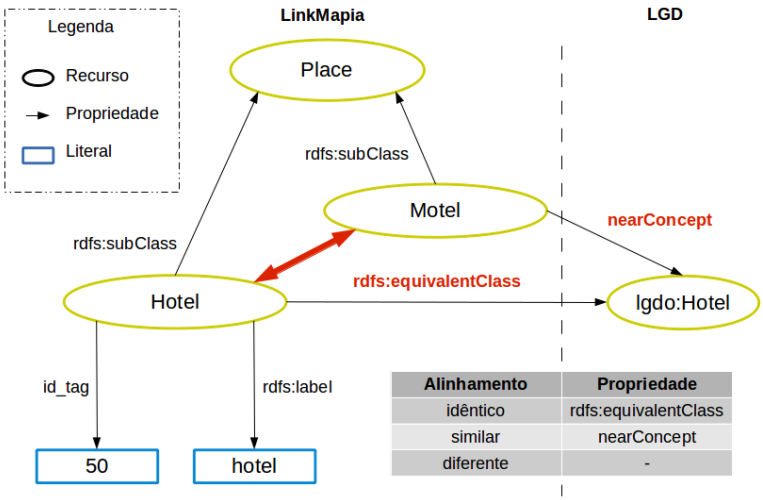
Figure 4. Ontologia rasa (figura criada pelo autor, em 2013)



A coleção de dados resultante ainda não pode ser chamada de dados ligados, visto que a ontologia precisa ser correlacionada com ontologias externas como LGD e GeoNames. O mapeamento da ontologia gerada para as ontologias externas é realizado convertendo as relações de equivalência de rótulos, obtidas na etapa de alinhamento e/ou mapeamento ontológico, em propriedades *rdfs:equivalentClass*, enquanto que relações de proximidade de conceito são convertidas em propriedades temporárias *nearConcept* para, posteriormente, facilitar a estruturação da ontologia pela comunidade do sistema colaborativo, como o caso da identificação da correlação entre *Hotel* e *Motel*, ilustrado na figura 5.

A geração de indivíduos para esta ontologia foi restrita aos dados da cidade de Milão recolhidos na etapa de extração de dados para caracterização do Wikimapia. A figura 6 ilustra o RDF de exemplo gerado pelo Protege, utilizado como esquema para a conversão.

Figure 5. Triplificação de categoria (figura criada pelo autor, em 2013)



4. Resultados obtidos

Com a nova coleção de dados LinkMapia, é possível responder perguntas como: **Quais são as opções de locais para dormir (como hotel, pensão, albergue) a até quinhentos metros do centro tecnológico da UFSC?** Dependendo de como a ontologia foi estruturada pela comunidade, é possível utilizar termos mais abrangentes (locais para dormir) para considerar na consulta uma coleção de conceitos (hotel, pensão, albergue), possibilitando diferentes granularidades para a categoria do objeto espacial.

Consultas podem ser executadas no *endpoint* SPARQL criado na publicação dos dados. A ligação da ontologia de LinkMapia com GeoNames, DBpedia e LGD permite ainda realizar consultas como: **Qual o restaurante conhecido como "Universitário" a até 200m do Centro de Eventos?** (figura 7). Esta consulta considera não apenas objetos geográficos do LGD classificados como *lgdo:Restaurant* como também objetos geográficos da Wikimapia anotados como *restaurant*, ou *seafood restaurant*, *drive-in restaurant*, *Subway (restaurant)*, dentre outros.

5. Trabalhos Correlatos

Os principais trabalhos relacionados encontrados na literatura são o Projeto SEEK[9], o Linked Geo Data [Stadler et al. 2012], GeoLinkedData[10] (Espanha), o Linked

[9]Disponível em: `http://www.seek-project.eu/`. Acesso em nov. 2013.
[10]Disponível em: `http://geo.linkeddata.es/web/guest/home`. Acesso em nov. 2013.

Figure 6. Triplificação de objeto (figura criada pelo autor, em 2013)

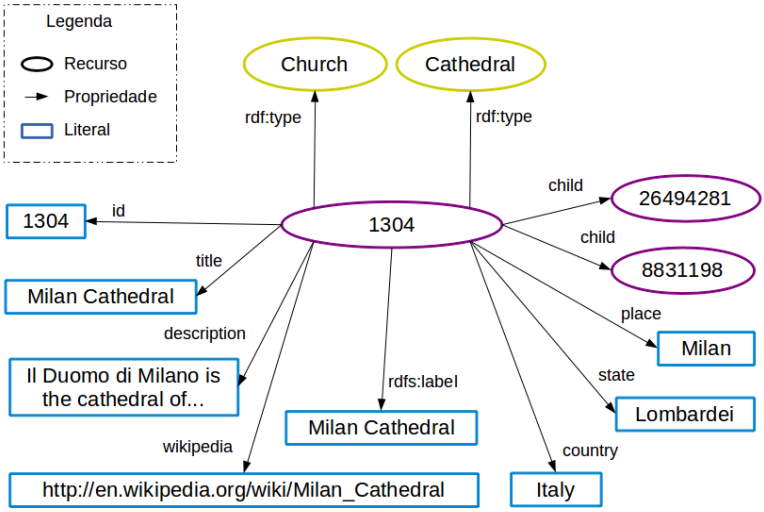


Figure 7. Exemplo consulta SPARQL 2 - Qual o restaurante conhecido como "Universitário" a até 200m do Centro de Eventos?

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?local ?nomeLocal {
 ?local
    a lgdo:Restaurant;
    rdfs:label ?nomeLocal ;
    geom:geometry [ ogc:asWKT ?geo ] .
FILTER( regex(?nomeLocal, "Universitário") &&
        bif:st_intersects ( ?geo,
        bif:st_point (-48.52015,-27.60225), 0.2) ).
}
```

Wikimapia[11], o *OurMap* [Gonzalez et al. 2013], o FolksOntology [Damme et al. 2007] e o LOM [Li 2004].

[11]Disponível em: `http://openeanwrap.appspot.com/`. Acesso em nov. 2013.

A iniciativa deste trabalho foi a obtenção de dados geográficos anotados para realizar experimentos de busca por similaridade sêmantica. O enriquecimento semântico dos dados geográficos proporcionam melhor entendimento dos fenômenos e eventos que acontecem nele, como as trajetórias de objetos móveis. Pretendia-se com este trabalho dar apoio a trabalhos futuros do grupo de pesquisa, como o projeto *Semantic Enrichment of trajectory Knowledge discovery* (SEEK).

Este trabalho propôs o enriquecimento semântico de dados coletatos colaborativamente. Este processo já havia sido realizado antes, com os dados do projeto OSM [Stadler et al. 2012]. Outra coleção de dados geográficos ligados é o GeoLinkedData, uma iniciativa de *Ontology Engineering Group* (OEG) destinada ao enriquecimento da *Web* de dados com dados geoespaciais do território nacional da Espanha.

Existe um projeto de publicação dos dados do Wikimapia como dados ligados, o Linked Wikimapia. Entretanto, atualmente é apenas publicado dados lidados de objetos geográficos, sem nenhuma ontologia para representar a *folksonomia* da Wikimapia. Os dados são ligados com o DBpedia a partir da conversão dos *links* já existentes para páginas do Wikipedia.

Uma diferente abordagem é seguida no sistema de coleta colaborativa *OurMap* [Gonzalez et al. 2013]. Este sistema de coleta colaborativa de dados geográficos já considera o processo de gerenciamento de conhecimento em documentos inteligentes, adaptanto-o para dados geográficos. Este sistema utiliza anotações semânticas e o usuário tem a opção de gerenciar as ontologias.

O processo de derivação de uma ontologia a partir de uma *folksonomia* é discutido por [Damme et al. 2007]. Este trabalho inspirou o processo proposto. Outro trabalho que ajudou na definição e implementação do processo foi o de [Li 2004].

6. Conclusões

Este capítulo apresenta a conclusões do trabalho. Em trabalhos futuros, são apresentadas as propostas: a conversão integral dos dados do Wikimapia para dados ligados; a expansão da ontologia do Wikimapia utilizando a mineração de texto das anotações *description*; a análise de qualidade dos dados do Wikimapia segundo [Mooney et al. 2011]; o desenvolvimento de um *data warehouse* (DW) semântico cuja dimensão espacial utiliza os dados geográficos ligados do Wikimapia e do OSM; e a mineração de trajetórias de objetos móveis considerando o enriquecimento semântico da dimensão espacial.

Este trabalho inicialmente apresenta uma breve pesquisa sobre anotações colaborativas de dados geográficos, apresentando as limitações semânticas de anotações livres em comparação aos artefatos da *web* semântica. Em seguida, propõe um processo de conversão de *folksonomia* em ontologia, possibilitando a publicação de dados geográficos anotados em sistemas colaborativos da web como dados ligados, baseando-se em alguns trabalhos relacionados.

Posteriormente, exemplifica o processo proposto com parte dos dados de um sistema colaborativo sem suporte a dados ligados, o Wikimapia, contribuindo com a extensão da coleção de dados geográficos ligados. Este trabalho ainda apresenta uma análise da semântica presente nos dados acessíveis do sistema colaborativo escolhido, apresentando histogramas e algumas consultas para caracterização dos dados.

Entretanto, este trabalho apenas apresenta o assunto abordado, sem definir com exatidão o processo. Pesquisas mais maduras, análises de outros sistemas colaborativos, elaboração e implementação de novas propostas de processo e análises de novos casos de estudo são fundamentais para determinar a melhor maneira de converter *folksonomias* em ontologias. Ainda assim, a viabilidade da elaboração do processo proposto, implementação e seus resultados preliminares demonstram como o trabalho é promissor para refinamentos.

Este trabalho não teve pretensão de realizar a publicação total do Wikimapia, visto que seu principal objetivo foi de fundamentar trabalhos futuros realizados durante o mestrado.

6.1. Trabalhos Futuros

Os resultados deste trabalho podem ser interpretados de diferentes maneiras: processo de conversão de *folksonomia* em ontologia, expansão do sistema colaborativo Wikimapia, expansão da coleção de dados geográficos ligados.

Independente do trabalho futuro a ser seguido, uma fundamentação mais madura sobre sistemas de coleta colaborativa, *folksonomia* e como que artefatos da *web* semântica podem auxiliar no enriquecimento semântico de dados encontrados nestes sistemas é essencial, visto que são assuntos ainda não consolidados.

6.1.1. Processo de conversão de *folksonomia* em ontologia

O processo proposto deve considerar trabalhos relacionados não abordados neste trabalho. É necessário realizar uma revisão sistemática sobre a conversão de *folksonomia* em ontologia. Na etapa de enriquecimento semântico, devem ser considerados também anotações em línguas diferentes de inglês, sendo necessário métodos de tradução ou bases léxicas em outras línguas. Também deve-se identificar, na etapa de alinhamento e/ou mapeamento ontológico, a métrica de similaridade mais apropriada para o processo e considerar comparações léxicas mais complexas como o *soft* TFIDF [Cohen et al. 2003]. A etapa semi-automátizada deve ser determinada, utilizando técnicas como *gameficação* [Thaler et al. 2011] para atrair a participação dos usuários do sistema de coleta colaborativa.

6.1.2. Expansão do sistema de coleta colaborativa

A arquitetura de sistemas de coleta colaborativa deve ser expandida para inserção do processo de conversão de *folksonomia* em ontologia, visando possibilitar a coexistência de anotações livres e anotações semânticas como proposto por [Christiaens 2006]. Deste modo o usuário não estará restrito a utilizar a apenas conceitos da ontologia para realizar anotações, e mesmo assim o sistema de coleta colaborativa será capaz de gerar dados com anotações de semântica bem definida. É prevista a sincronização periódica da *folksonomia* com a ontologia e vice-versa. Esta nova arquitetura também deve considerar a evolução ontológica, moderação e privacidade dos dados anotados.

6.1.3. Expansão da coleção de dados geográficos ligados

O desenvolvimento de um DW semântico com os dados convertidos neste trabalho é essencial para a análise das informações fornecidas e obtenção de conhecimento. A anotação dos dados geográficos proporciona a expansão da granularidade da dimensão espacial (especificamente, a categoria do objeto espacial) ao nível de detalhamento das entidades da ontologia usada na anotação. Um DW oferece uma estrutura ideal para a aplicação de algoritmos de mineração por causa da utilização do modelo estrela. Diferente do modelo relacional, o modelo estrela valoriza mais as operações de consulta e recuperação de informação que as de criação, modificação e deleção. A anotação de dados geográfico proporciona novas oportunidades de aplicações de mineração de dados, como por exemplo a mineração de subestruturas frequentes, localização de subestruturas e e a busca por similaridade de entidades correlacionadas, aplicações conhecidas da mineração de grafos.

A mineração de trajetórias de objetos móveis estuda o comportamento de objetos móveis no espaço geográfico. O processo de descoberta de conhecimento em trajetórias de objetos móveis é similar ao processo de descoberta de conhecimento e mineração de dados (*Knowledge Discovery and Data mining* - KDD). O enriquecimento semântico da dimensão espacial pode contribuir para o enriquecimento semântico da trajetória dos objetos móveis. A motivação, ou *goal*, da trajetória pode estar presente na descrição dos lugares visitados pelo objeto móvel.

References

Christiaens, S. (2006). Metadata mechanisms: From ontology to folksonomy... and back. In *Lecture Notes in Computer Science*, pages 199–207.

Cohen, W. W., Ravikumar, P. D., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *IIWeb*, pages 73–78.

Damme, C. V., Hepp, M., and Siorpaes, K. (2007). Folksontology: An integrated approach for turning folksonomies into ontologies. In *IN PROCEEDINGS OF THE ESWC WORKSHOP BRIDGING THE GAP BETWEEN SEMANTIC WEB AND WEB 2.0*. Springer.

Gonzalez, A., Izidoro, D., Willrich, R., and Santos, C. (2013). Representação Aberta e Semântica de Anotações de Incidentes em Mapas Web. In *Simpósio Brasileiro de Sistemas Multimídia e Web, 2013, João Pessoa. Anais do IX Simpósio Brasileiro de Sistemas Multimídia e Web (por aparecer)*, volume 1, pages 1–12.

Li, J. (2004). Lom: A lexicon-based ontology mapping tool. In *Proceedings of the Performance Metrics for Intelligent Systems (PerMIS)*, page 2004.

Mooney, P., Sun, H., Corcoran, P., and Yan, L. (2011). Citizen-generated spatial data and information: Risks and opportunities. In *Intelligence and Security Informatics*.

Stadler, C., Lehmann, J., Höffner, K., and Auer, S. (2012). Linkedgeodata: A core for a web of spatial open data.

Thaler, S., Simperl, E. P. B., and Siorpaes, K. (2011). Spotthelink: A game for ontology alignment. In Maier, R., editor, *Wissensmanagement*, volume 182 of *LNI*, pages 246–253. GI.

ANEXO A – Caracterização do Wikimapia

O Wikimapia é um projeto com o objetivo de criar e manter um mapa atualizado, completo, gratuito e multilíngue de todo o mundo. Seu conteúdo é inteiramente criado por voluntários da *Internet*, não havendo restrições para colaborar. Os colaboradores do Wikimapia usufruem de um SIG para realizar consultas, criar, editar e visualizar os dados. Após a seleção ou inserção da geometria de um elemento geográfico, a ferramenta disponibiliza formulários *web* para coletar atributos descritivos. A descrição pode incluir *links* de vídeos, imagens e páginas na *web*, em especial para a Wikipédia. O usuário é incentivado a classificar uma ou mais vezes o local segundo uma hierarquia de classes não fixa (aberta).

Além da ferramenta para visualização e edição dos dados geográficos e anotações, o Wikimapia disponibiliza também uma *wiki* com documentação e orientações para os usuários, um *blog* com as últimas notícias do projeto e um fórum para discussão sobre novas *features* e correções de problemas e *bugs*.

O projeto disponibiliza uma API, desenvolvida como um *web service*, para a consulta de seu banco de dados desde 2010. Esta API utiliza o protocolo REST, aceitando requisições em HTTP GET e respondendo em formato *Extensible Markup Language* (XML), *JavaScript Object Notation* (JSON), JSON *with padding* (JSONP), *Keyhole Markup Language* (KML) ou binário, com opção de compressão gzip. Para utilizar a API, basta ter uma conta na ferramenta e solicitar uma chave. Os desenvolvedores podem integrar os dados geográficos do Wikimapia em suas aplicações usando as 3 funções descritas a seguir.

- *box(pt1, pt2)*: dados dois pontos pt1 e pt2, a função *box* retorna a localização, o polígono e as informações básicas de todos os objetos dentro do retângulo formado por esses 2 pontos. Um polígono da Wikimapia (*e.g.*, representando um país, estado, ou cidade) também pode ser utilizado ao invés dos pontos. Esta função retorna dados armazenados em *cache* e, por essa razão, tais dados podem ser desatualizados.

- *object(id)*: a função *object* retorna informações mais completas do objeto cujo *id* foi parametrizado.

- *search(pt, keyword)*: dado um ponto pt, a função *search* busca as construções mais próximas (sem informações de como são determinadas) cujo nome possua a palavra-chave *keyword*. O algoritmo de busca utilizado não é informado. É possivel escolher uma quantidade limite de resultados.

A partir de 2013, a API do Wikimapia passou a oferecer novas funções: *Place.Getbyid*, *Place.Getbyarea*, *Place.Getnearest*, *Place.Search*, *Street.Get-*

byid, *Category.Getbyid*, *Category.GetAll* e *Api.Getlanguages*. Estas novas funções oferecem mais informações que as funções antigas: A função *Place.Getbyid* prove fotos e comentários dos usuários a respeito dos lugares pesquisados. As funções *Category* permitem a busca das *tags*, agora chamadas de categorias, oferecendo informações sobre a quantidade de dados anotados (*amount*), descrição (*description*) e categorias sinônimas (*synonyms*).

Entretanto, a documentação é focada na manipulação da interface. Informações quanto a representação dos dados e arquitetura do projeto são escassas.

A.1 EXTRAÇÃO DE DADOS

Cada função da API retorna parte dos dados do Wikimapia.

A.1.1 Objetos Geográficos

A extração da base relativa a objetos geográficos foi limitada espacialmente a um estudo de caso: a cidade de Milão. O menor retângulo que contém a cidade de Milão (*minimal bounding rectangle*) é definido pelos pontos (45.388039, 9.043907) e (45.536266, 9.278963), de latitude e longitude respectiva, segundo o polígono dos limites políticos da cidade na base de dados *Database of Global Administrative Areas* (GADM).

A função *box* da API da Wikimapia apresenta uma limitação. Para ser devidamente utilizada, suas entradas devem respeitar as condições:

```
1. (lon\_max - lon\_min) * 10000000 > 12100
2. |(lon\_max - lon\_min) * (lat\_max - lat\_min)|
                          * 10000000 < 14641000000
```

Aplicando as formulas ao *minimal bounding rectangle* de Milão, obtemos:

```
1. (9.278963 - 9.043907) * 10000000 > 12100
                  0.235056          > 12100
                   2350560          > 12100

2. 0.235056 * (45.536266 - 45.388039) * 10000000
       0.235056 * 0.148227 * 10000000       < 14641000000
              348416.45712                  < 14641000000
```

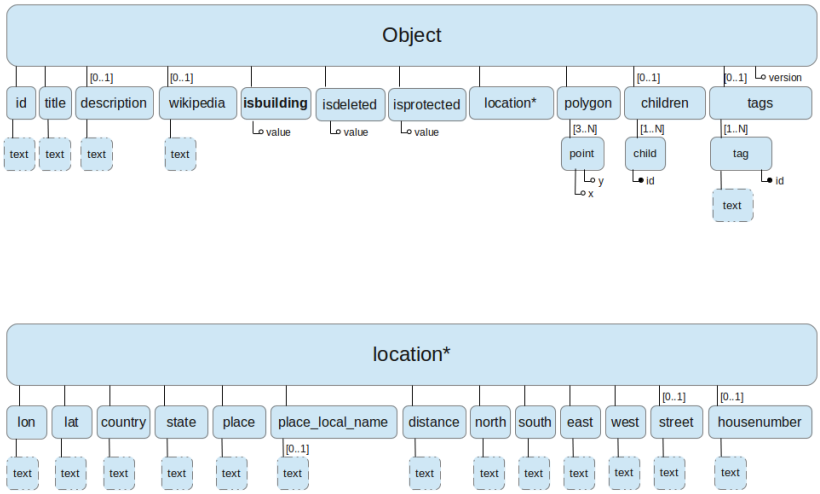
A função *box*[1], com esses parâmetros, resulta em uma lista com 1276

[1]Disponível em: <http://api.wikimapia.org/?function=box&lon_min=9.

ids de objetos, separados por páginas, sendo a váriável *count* o limite de resultados por página, em formato XML ou JSON. Deve-se utilizar a função *object*[2] para obter a descrição detalhada de cada objeto.

Para extrair o conteúdo do XML retornado pela função *object* é importante considerar seu modelo lógico. Contudo, a API não oferece nem esta informação nem um dicionário de dados apropriado, mas pode-se inferir ambos a partir dos resultados.

Figura 22 – **Modelo lógico 1 - inferido da função *object* da API do Wikimapia (figura criada pelo autor, em 2012)**



Com o modelo lógico (figura 22) e o dicionário (tabelas 4 e 5) inferidos é possível construir um modelo conceitual e o modelo lógico do banco de dados geográfico provisório para realizar a caracterização dos dados extraídos.

A figura 23 apresenta o modelo conceitual que, por sua vez, favorece o desenvolvimento do modelo lógico da base de dados intermediária:

Object(**id**, title, description, wikipedia, isBuilding, isDeleted, isPro-

043907&lat_min=45.388039&lon_max=9.278963&lat_max=45.536266&rsaquo&key=example&page=1&count=100&disable=location,polygon>. Acesso em jan. 2013.

[2]Disponível em: <http://api.wikimapia.org/?function=object&key=example&id=ID_DO_OBJETO>. Acesso em jan. 2013.

[3]Numérico.

[4]Determinante.

[5]Booleano.

ELEMENTO	DESCRIÇÃO	TIPO	RESTRIÇÃO	CLASSE
ID	IDENTIFICADOR ÚNICO.	NUM.[3]	NÃO NULO	DETERM.[4]
TITLE	NOME DO LUGAR QUE O OBJETO REPRESENTA.	TEXTO	NÃO NULO	SIMPLES
DESCRIPTION	DESCRIÇÃO DO OBJETO, O QUE ELE REPRESENTA PRA SOCIEDADE E SUAS FORMAS DE UTILIZAÇÃO.	TEXTO	-	SIMPLES
WIKIPEDIA	ENDEREÇO ELETRÔNICO DA PÁGINA DA WIKIPÉDIA SOBRE ESSE OBJETO.	TEXTO	-	SIMPLES
ISBUILDING	INDICA SE O OBJETO É UMA CONSTRUÇÃO.	BOOL.[5]	NÃO NULO	SIMPLES
ISDELETED	INDICA SE O OBJETO FOI MARCADO COMO DELETADO NA BASE, DEVIDO A DADOS ERRADOS OU DESATUALIZADOS.	BOOL.	NÃO NULO	SIMPLES
ISPROTECTED	INDICA SE O DADO É PROTEGIDO, NO CASO DE SER UMA RESIDÊNCIA OU ÁREA RESTRITA.	BOOL.	NÃO NULO	SIMPLES
LOCATION	INFORMAÇÕES SOBRE A LOCALIDADE.	XML	NÃO NULO	SIMPLES
POLYGON	POLÍGONO DO OBJETO.	XML	NÃO NULO	SIMPLES
POINT	PONTO NO ESPAÇO ONDE X E Y SÃO LATITUDE E LONGITUDE.	TEXTO	NÃO NULO	SIMPLES
CHILDREN	LISTA DE IDS DE OBJETOS INTERNOS A ESSE OBJETO.	XML	-	SIMPLES
CHILD	ID DO OBJETO QUE É INTERNO A ESSE OBJETO.	TEXTO	NÃO NULO	SIMPLES
TAGS	LISTA DAS CATEGORIAS QUE O OBJETO FOI CLASSIFICADO.	XML	-	SIMPLES
TAG	DESCRIÇÃO DE QUAL LUGAR QUE O OBJETO REPRESENTA.	TEXTO	NÃO NULO	SIMPLES

Tabela 4 – **Dicionário de dados - Elemento Object: O objeto criado pelos usuários da Wikimapia (tabela criada pelo autor, em 2012)**

tected, id_plg): Armazena os objetos(lugares) criados pelos usuários da Wikimapia e suas informações;

Tagged(**id_obj**, **id_tag**): Representa a classificação de um objeto a determinada categoria;

Tag(**id**, name): Armazena as categorias usadas na classificação dos objetos;

Child(**id_obj**, **id_obj_child**): Representa a relação de um objeto dentro de outro objeto. Por exemplo, uma loja dentro de um centro comercial;

Polygon(**id_plg**, points): Armazena os polígonos dos objetos criados pelos usuários da Wikimapia. Esta entidade representa as tabelas do *framework* escolhido para suporte aos dados geográficos, como, por exemplo,

ELEMENTO	DESCRIÇÃO	TIPO	RESTRIÇÃO	CLASSE
LON	LONGITUDE DO PONTO CENTRAL DO POLÍGONO DO OBJETO.	NUM.	NÃO NULO	SIMPLES
LAT	LATITUDE DO PONTO CENTRAL DO POLÍGONO DO OBJETO.	NUM.	NÃO NULO	SIMPLES
COUNTRY	PAÍS DO LUGAR REPRESENTADO PELO OBJETO.	TEXTO	NÃO NULO	SIMPLES
STATE	ESTADO OU PROVÍNCIA OU REGIÃO DO LUGAR REPRESENTADO PELO OBJETO.	TEXTO	NÃO NULO	SIMPLES
PLACE	NOME DO LUGAR QUE O OBJETO REPRESENTA.	TEXTO	NÃO NULO	SIMPLES
PLACE_LOCAL _NAME	NOME LOCAL DO LUGAR QUE O OBJETO REPRESENTA.	TEXTO	NÃO NULO	SIMPLES
DISTANCE	-	NUM.	NÃO NULO	SIMPLES
NORTH	O PONTO DO POLÍGONO DO OBJETO QUE ESTÁ MAIS AO NORTE.	NUM.	NÃO NULO	SIMPLES
SOUTH	O PONTO DO POLÍGONO DO OBJETO QUE ESTÁ MAIS AO SUL.	NUM.	NÃO NULO	SIMPLES
EAST	O PONTO DO POLÍGONO DO OBJETO QUE ESTÁ MAIS AO LESTE.	NUM.	NÃO NULO	SIMPLES
WEST	O PONTO DO POLÍGONO DO OBJETO QUE ESTÁ MAIS AO OESTE.	NUM.	NÃO NULO	SIMPLES
STREET	NOME DA RUA QUE O OBJETO ESTÁ LOCALIZADO.	TEXTO	-	SIMPLES
HOUSE-NUMBER	NÚMERO DA CONSTRUÇÃO REPRESENTADA PELO OBJETO.	NUM.	-	SIMPLES

Tabela 5 – **Dicionário de dados - Elemento Location: Informações sobre o local do objeto e metadados para a ferramenta (tabela criada pelo autor, em 2012)**

o Postgis;

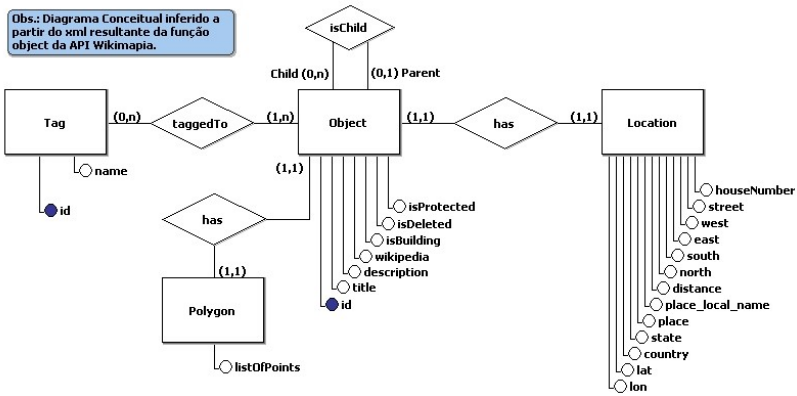
Location(id_obj, lon, lat, country, state, place, place_local_name, distance, north, south, east, west, street, housenumber): Armazena as informações dos objetos referentes ao lugar que representam e metadados para o software de interface da Wikimapia;

Obs.: As chaves primárias estão em negrito e as chaves estrangeiras sublinhadas;

O extrator foi modelado utilizando o GeoKettle 3.2.0, uma ferramenta de extração transformação e carga, ou *extract, transform and load* (ETL), de dados dirigida aos dados geográficos. O GeoKettle é uma versão do Kettle (*Pentaho Data Integration*) que suporta dados espaciais e é disponível sob a lincensa *GNU Lesser General Public License* (LGPL).

Além dos 1276 objetos, foram extraídos também 42 relações de *children*, 243 categorias (a tabela 6 e a figura 24 apresentam as 30 categorias

Figura 23 – **Modelo conceitual (figura criada pelo autor, em 2012)**



mais frequentes em anotações) e 1513 relações de anotação. O SGBD utilizado para o armazenamento dos dados extraídos foi o PostgreSQL 9.2.4 com a extensão geográfica PostGIS 2.0.3-1[6].

A.1.2 *Folksonomia* do Wikimapia

De todos os atributos que um *Object* apresenta, aquele que se refere a categoria (*tag*) é a anotação livre, que representa o tipo do lugar ou construção. Este atributo é utilizado para recuperação e restrição de consultas no Wikimapia. Em Outubro de 2013, uma pesquisa utilizando a função *Category.GetAll*[7] revela a existência de 8615 categorias. Este conjunto de categorias forma a *folksonomia* do Wikimapia, criada colaborativamente pelos usuários e em constante evolução.

O retorno da função disponibiliza as informações apresentadas no modelo lógico inferido (figura 25) e descritas pelo dicionário de dados (tabela 7).

A extração da *folksonomia* foi realizada por transformações de ETL. ExtratorCategorias (figura 26) utilizou a função da API *Category.GetAll* para obter *id*, nome e frequência de uso de todas as categorias. A tabela 8 e a figura 27 apresentam as 30 categorias mais utilizadas em anotações.

[6] *dump* do banco em anexo

[7] Disponível em: <`http://api.wikimapia.org/?key=example&function=category.getall&format=&pack=&language=en&name=&page=1&count=50`>. Acesso em set. 2013.

NOME	FREQUÊNCIA
DRAW ONLY BORDER	106
PARK	88
SQUARE	80
QUARTERS	63
METRO / SUBWAY / UNDERGROUND STATION	55
UNDERGROUND FACILITY	51
STORE / SHOP	42
CHURCH	41
HISTORICAL LAYER / DISAPPEARED OBJECT	38
PETROL / GAS STATION	34
SCHOOL	31
THIRD-LEVEL ADMINISTRATIVE DIVISION	30
VILLAGE	30
HOTEL	30
OFFICE BUILDING	30
TRAIN STATION	30
PRODUCTION	28
COMMUNE - ADMINISTRATIVE DIVISION	27
CASCINA	25
STADIUM	24
LAKE	21
UNDER CONSTRUCTION	20
UNIVERSITY	20
SUPERMARKET	19
BAR	17
SPORT VENUE	16
CEMETERY	14
APARTMENT BUILDING	14
SWIMMING POOL	14
HOUSE	13

Tabela 6 – **Tabela de categorias de Milão - 30 mais utilizadas em anotações (tabela criada pelo autor, em 2013)**

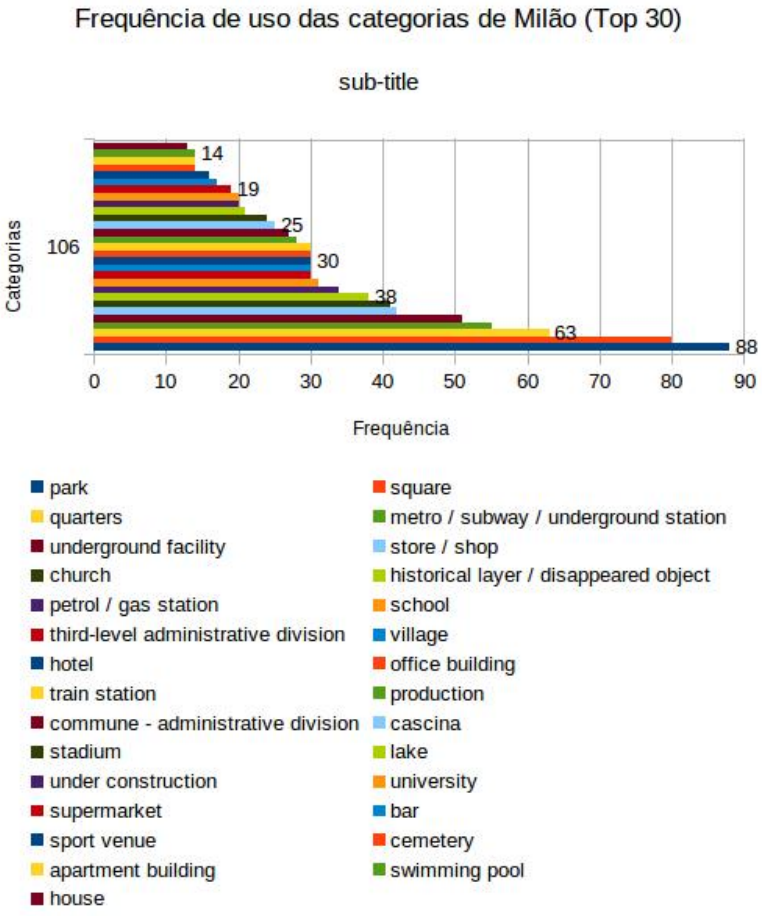
ELEMENTO	DESCRIÇÃO	TIPO	RESTRIÇÃO	CLASSE
ID	IDENTIFICADOR	NUM.	NÃO NULO	SIMPLES
AMOUNT	NÚMERO DE LUGARES CLASSIFICADOS POR ESTA CATEGORIA.	NUM.	NÃO NULO	SIMPLES
ICON	ENDEREÇO DA IMAGEM DA CATEGORIA.	URL	-	SIMPLES
NAME	NOME DA CATEGORIA.	TEXTO	NÃO NULO	SIMPLES

Tabela 7 – **Dicionário de dados - Elemento retornado pela função Categoria: Categoria utilizada para classificar objetos criados por usuários do Wikimapia (tabela criada pelo autor, em 2013)**

Para fins de ilustração da folksonomia extraída, sua nuvem de etiquetas (figura 13) foi gerada utilizando a ferramenta Wordle[8], desconsiderando as 5 categorias mais frequentes (*place without photos, place without description, place without category, building without address, place without polygon*).

[8]Disponível em: <`http://www.wordle.net`>. Acesso em out. 2013.

Figura 24 – **Gráfico Categorias de Milão - 30 mais utilizadas em anotações (figura criada pelo autor, em 2013)**



A.1.3 Caracterização do Conteúdo de Dados do Wikimapia

As categorias são analisadas utilizando histogramas e consultas em *Structured Query Language* (SQL). O SGBD utilizado para a análise foi o PostgreSQL com a extensão PostGIS.

Figura 25 – **Modelo lógico 2 - inferido da função *Categoria.GetAll* da API do Wikimapia (figura criada pelo autor, em 2013)**

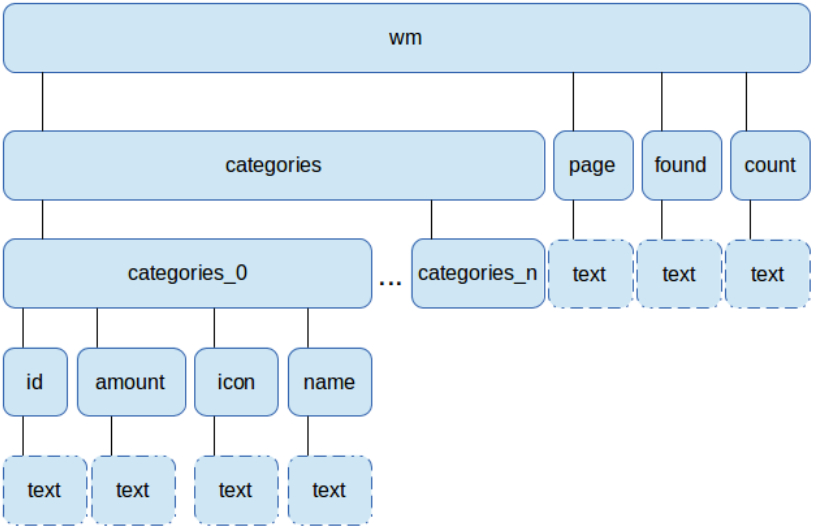
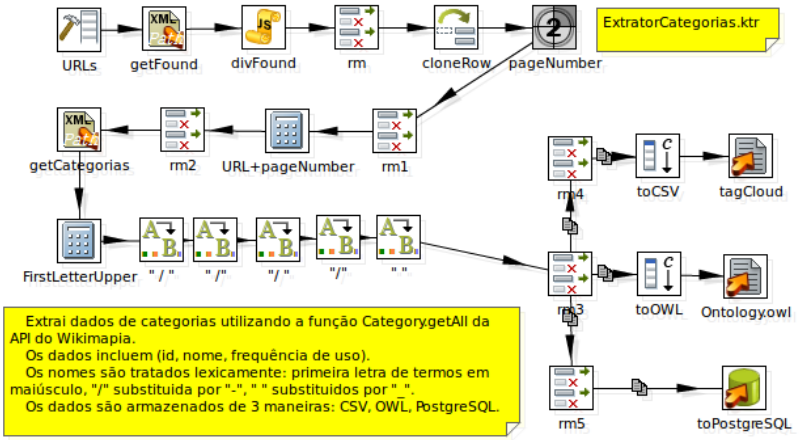


Figura 26 – **Extrator de categorias (figura criada pelo autor, em 2013)**



NOME	FREQUÊNCIA
PLACE WITHOUT PHOTOS	22237340
PLACE WITHOUT DESCRIPTION	14170795
PLACE WITHOUT CATEGORY	13281368
BUILDING WITHOUT ADDRESS	6769359
PLACE WITHOUT POLYGON	3909786
BUILDING	1473509
DWELLING	1401579
HOME	1401579
HOUSE	1401579
RESIDENCE	1401579
VILLA	1401579
PLACE WITH TRIANGULAR POLYGON	1156392
VILLAGE	997893
APARTMENT BUILDING	969453
APARTMENTS	969453
BLOCK OF FLATS	969453
TENEMENT	969453
TOWER BLOCK	969453
SHOPPING AND SERVICES	843255
STORE / SHOP	835422
EDUCATION	680954
LEARNING	680954
LESSONS	680954
SCHOOL	680954
SCHOOLHOUSE	680954
SCHOOLING	680954
TEACHING	680954
DINING AND LEISURE	675989
DEITY	558168
FAITH	558168

Tabela 8 – **Tabela de categorias global - 30 mais utilizadas em anotações de todos os locais do mundo cadastrados no Wikimapia (tabela criada pelo autor, em 2013)**

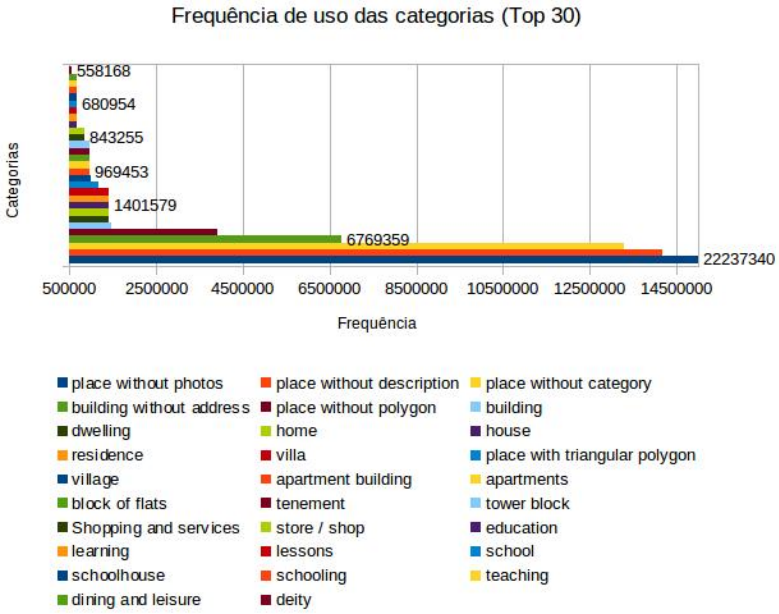
A caracterização dos objetos espaciais extraídos, armazenados em uma base de experimentos, fundamenta-se em tabelas de avaliação da esparcialidade de atributos (tabela 9) e histogramas ilustrando os relacionamentos *children* (figura 29) e categorização (*tag*), na figura 30 e 31.

A tabela de esparcialidade apresenta o número de objetos que possuem cada atributo opcional (exceto por *title*, que embora obrigatório, geralmente apresenta valor vazio).

Um histograma é uma relação entre classes de frequência (número de filhos ou objetos anotados) e quantidade de elementos (objetos ou categorias) daquela classe, onde a frequência é o número de filhos daqueles objetos, ou o número de objetos anotados por aquelas categorias. Obteve-se o mínimo e o máximo das frequências de anotação das categorias coletadas na extração de dados.

Os histogramas da relação *children* e dos objetos anotados de Milão apresentam classes intervaladas linearmente (de 1 em 1), enquanto o histo-

Figura 27 – **Frequência das categorias global - 30 mais utilizadas em anotações de todos os locais do mundo cadastrados no Wikimapia (figura criada pelo autor, em 2013)**



ATRIBUTO	NÚMERO DE OBJETOS
TITLE	306
DESCRIPTION	183
WIKIPEDIA	82
DESCRIPTION AND WIKIPEDIA	70
ISBUILDING	396
ISDELETED	0
ISPROTECTED	0
STREET	175
HOUSE-NUMBER	35
STREET AND HOUSE-NUMBER	35
PLACE_LOCAL_NAME	0
PLACE = ”MILAN”	420
CHILDREN	19
TAGS	1009
TOTAL	1276

Tabela 9 – **Tabela de esparcialidade de dados de Milão - Base Experimental (tabela criada pelo autor, em 2013)**

Figura 28 – **Nuvem de anotações da *folksonomia* global (figura criada pelo autor, em 2013)**

grama das categorias de Milão possui classes em intervalos distribuídos por logarítmo de base 2. A linha vermelha mostra a frequência acumulada de anotações já representadas pelas classes, medida em porcentagem.

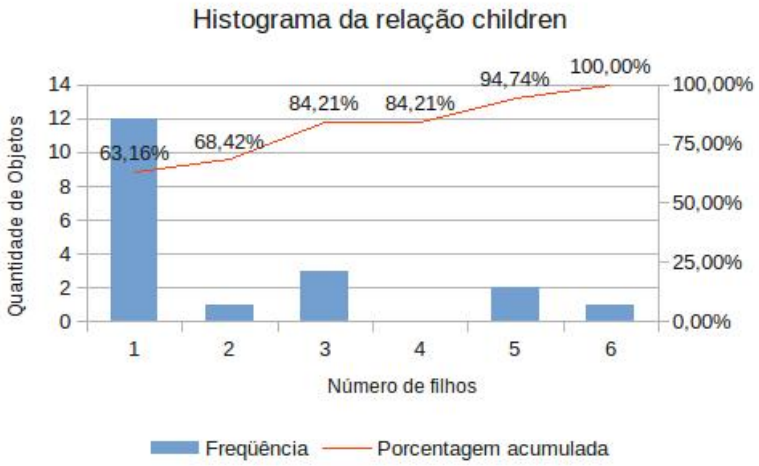
As análises da *folksonomia* identificaram a distribuição das anotações (figuras 32, 33, 34) e características da organização dos dados. A distribuição das anotações foi analisada com o auxílio de histogramas.

Os histogramas da *folksonomia* se diferem apenas nos tamanhos dos intervalos de classe (número de objetos anotados). Três funções foram utilizadas: crescimento linear (figura 32), resultando em 90 classes de tamanho igual (247082); crescimento quadrático (figura 33), 25 classes cujos tamanhos sempre são duas vezes maior que a classe anterior e o tamanho da primeira classe é igual a 2; crescimento cúbico (figura 34), 16 classes cujos tamanhos sempre são três vezes maior que a classe anterior e o tamanho da primeira classe é igual a 3.

Dentre as 8615 categorias, 1398 compartilham entre si 541 *ids*. Estes resultados podem ser obtidos pelas seguintes consultas:

```
SELECT * FROM categorias AS a JOIN (
```

Figura 29 – **Histograma da relação *children* dos dados de Milão - distribuição linear nas classes de amostragem (figura criada pelo autor, em 2013)**




```
    SELECT id, count(id) FROM categorias GROUP BY id
    ) AS b ON a.id = b.id WHERE count > 1 ORDER BY count;

SELECT DISTINCT a.id FROM categorias AS a JOIN (
    SELECT id, count(id) FROM categorias GROUP BY id
    ) AS b ON a.id = b.id WHERE count > 1;
```

O *id* 1663 (categoria *religion*), por exemplo, é compartilhado por 10 categorias. Isso ocorre devido a forma escolhida para a implementação da relação de sinônimo entre anotações, onde o conceito mais genérico é *religion* e seus sinônimos são *god* (1815), *holy* (2528), *religious* (3073), *faith* (3092), *sacred* (3232), *spirituality* (9274), *spiritual* (11949), *goddess* (15687), *deity* (34874)[9]. Neste caso, a frequência de uso é contabilizada em conjunto. Entretanto existem outras anotações, como *church* (122), *temple* (46), *assembly of god* (45349), que não estão correlacionadas com a relação de sinônimo ou de hierarquia.

[9]Relação também identificável utilizando a função *Categories.Getbyid*. Disponível em: <`http://api.wikimapia.org/?key=example&function=category.getbyid&id=1663&format=&pack=&language=en`>. Acesso em out. 2013.

Figura 30 – **Histograma dos objetos anotados de Milão - distribuição logarítmica de base 2 nas classes de amostragem (figura criada pelo autor, em 2013)**

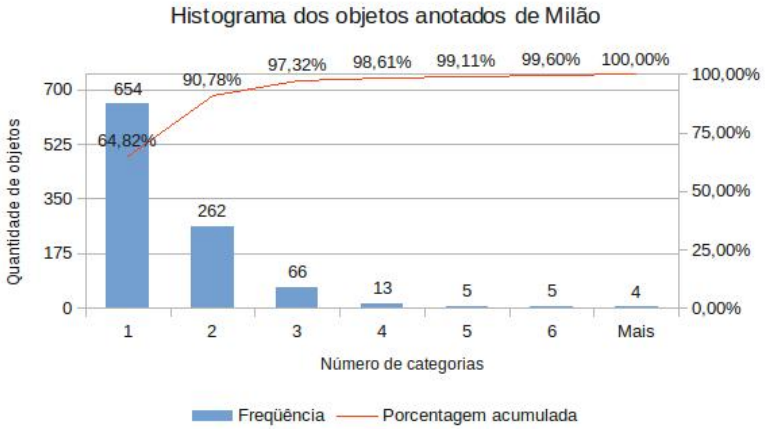


Figura 31 – **Histograma das categorias de Milão - distribuição logarítmica de base 2 nas classes de amostragem (figura criada pelo autor, em 2013)**

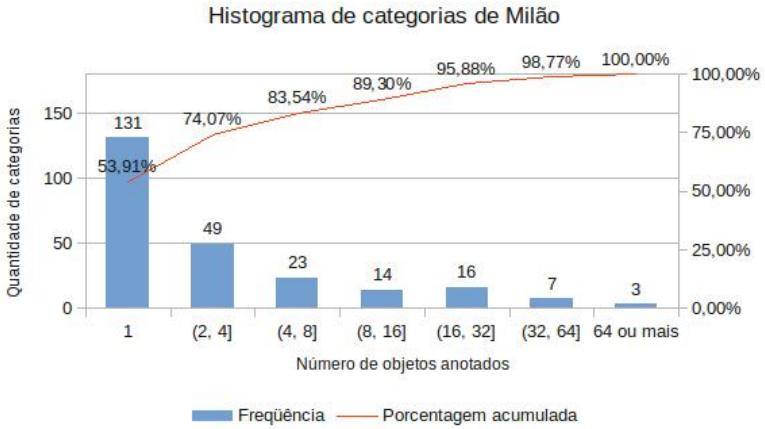


Figura 32 – **Histograma da *folksonomia* global 1 - distribuição linear nas classes de amostragem (figura criada pelo autor, em 2013)**

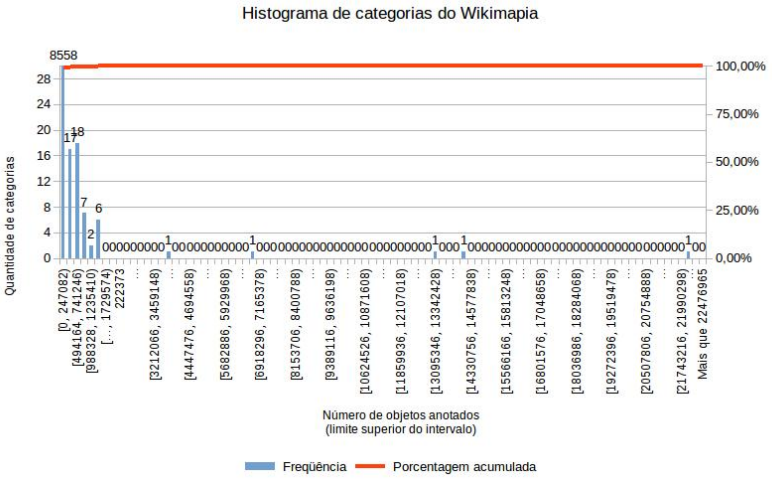


Figura 33 – **Histograma da *folksonomia* global 2 - distribuição logarítma de base 2 nas classes de amostragem (figura criada pelo autor, em 2013)**

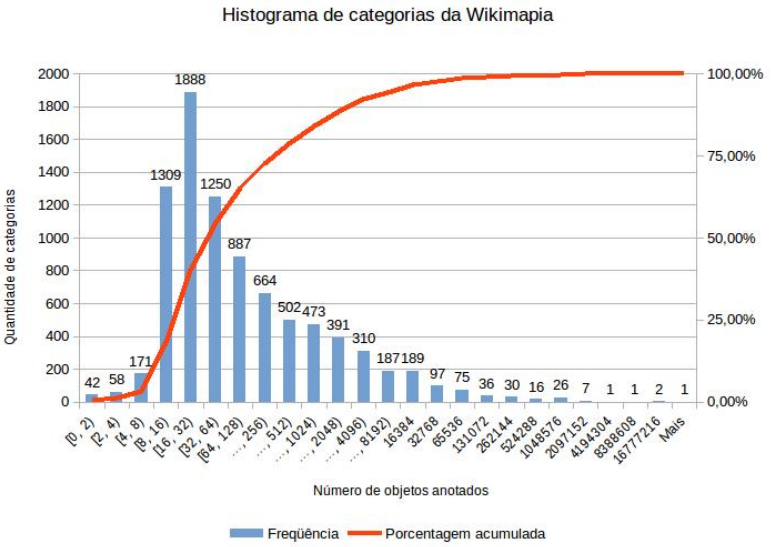


Figura 34 – **Histograma da *folksonomia* global 3 - distribuição logarítma de base 3 nas classes de amostragem (figura criada pelo autor, em 2013)**