Introducing spatial coverage in a semantic repository model

Tardy, Camille

This publication URL: https://archive-ouverte.unige.ch/unige:110652
Publication DOI: 10.13097/archive-ouverte/unige:110652

Introducing Spatial Coverage in a Semantic Repository Model

THESIS

presented to the Faculty of Economics and Management
of the University of Geneva
by

Camille Tardy

Under the direction of

Dr. Laurent Moccozet and
Prof. Gilles Falquet

to obtain the title of

Docteur ès systèmes d'information

Jury members:
Prof. Dimitri Konstantas, University of Geneva,
President of the jury
Prof. Javier Nogueras-Iso, University of Zaragoza, Spain
Prof. Giovanna Di Marzo Serugendo, University of Geneva
Dr. Claudine Métral, University of Geneva

Thèse n° 40

Geneva, 30 Janvier 2017

La Faculté d'Economies et Management, sur préavis du jury, a autorisé l'impression de la présente thèse, sans entendre, par là, émettre aucune opinion sur les propositions qui s'y trouvent énoncées et qui n'engagent que la responsabilité de leur auteur.

Genève, le 30 Janvier 2017

La doyenne

Maria-Pia Victoria-Feser

Impression d'après le manuscrit de l'auteur

Table of Contents

List of Figures

Acronyms and Abbreviations

3DCM	3D City Model
CG	Cumulated Gain
DC	Dublin Core
DCG	Discounted Cumulated Gain
DL	Digital Library/ies
DLMS	Digital Library Management System
EXIF	EXchangeable Image file Format
GIR	Geographic Information Retrieval
GIS	Geographic Information System
IPTC	International Press Telecommunications Council
IR	Information Retrieval
LOD	Level of Detail
MBR	Minimum Bounding Rectangle
NER	Name Entity Recognition
NERC	Name Entity Recognition and Classification
OWL	Web Ontology Language
OGC	Open Geospatial Consortium
POI	Point of Interest
POS	Part-of-Speech
PPGIS	Public Participation Geographic Information System
RDF	Resource Description Framework
VGI	Volunteered Geographic Information
W3C	World Wide Web Consortium
WOEID	Where on Earth Identifier
WS	Word Sense
WSD	Word Sense Disambiguation
XML	Extensible Markup Language

Résumé

La prise de décision est une tâche universelle qui demande souvent de rassembler des données hétérogènes provenant de différents domaines d'activité. Les bibliothèques numériques sont des outils renommés pour le stockage, la gestion et la manipulation de données hétérogènes. Elles sont souvent utilisées pour aider et faciliter les tâches de collaboration et de prise de décision. Pour exécuter ces tâches, les utilisateurs-trices de ces bibliothèques devraient se voir présenter les informations nécessaires contextualisées, sans qu'il/elle n'ait besoin de les traiter afin de les lire de façon compréhensible. Toutes les relations entre les ressources contenues dans la bibliothèque numérique doivent apparaître clairement à l'utilisateur-trice afin que celui/celle-ci puisse facilement identifier les ressources complémentaires.

La connaissance peut varier temporellement et spatialement, il est donc pertinent de prendre en considération la couverture (ou la portée) spatio-temporelle des entités qui composent les ressources de connaissance. Dans le contexte des bibliothèques numériques, les ressources de connaissance comme les ontologies ou les thésaurus sont utilisées pour annoter les ressources gérées par la bibliothèque. Donc la couverture spatio-temporelle peut aussi être calculée pour les ressources annotées par les ressources de connaissance.

Dans cette thèse nous proposons un modèle de bibliothèque numérique sémantique avec un contexte géo-spatial et une définition de couverture comme concept clé. Nous présentons les modèles de documents et de ressources spatiales. Nous définissons le modèle d'annotation et plus particulièrement la couverture géographique qui détaille et définit la localisation de chaque ressource en prenant en compte son type. Finalement, nous détaillons le modèle de requête et le procédé de correspondance où le contexte géo-spatial est une caractéristique clé.

Afin de valider ce modèle, nous développons des cas d'utilisation et des parties d'implémentation. Pour commencer, nous nous concentrons sur l'annotation de documents et précisément sur la localisation de documents au sein d'une ressource spatiale. Pour cela nous décrivons l'implémentation du modèle d'annotation, présenté dans le modèle de la bibliothèque numérique, et spécialement l'alignement des ressources de connaissances géo-sémantiques.

Puis nous présentons la méthodologie et l'implémentation d'une nouvelle technique d'extraction d'informations géographiques et de sémantique des lieux, depuis des tags issus de sources VGI (information géographique volontaire). Cette technique est basée sur un système de catégorisation avec une approche non statistique, basée sur la connaissance. Cette extraction peut partiellement

automatiser la création des couvertures géographiques pour les ressources des bibliothèques numériques, ou être utilisée pour enrichir sémantiquement ou compléter des modèles 3D ou des services géographiques.

Abstract

Decision-making is a universal task that often calls for the gathering of heterogeneous data from different domains of activity. Digital libraries are a famous tool for storing, managing and handling heterogeneous data. They are often used to support collaboration and decision-making task. In order to complete those tasks, the digital library user should be presented with all the necessary information contextualised, with no need for him to process them in order to understand them. All the relations between resources in the digital library should appear clearly for the user to easily identify complementary resources.

Knowledge can vary temporally and spatially, thus it is pertinent to take into account the spatio-temporal coverage (or scope) of entities that compose the knowledge resources. In the context of digital libraries, knowledge resources such as ontologies or thesauri are used to annotate the resources managed by the library. So the spatio-temporal coverage can also be computed for the resources annotated by the knowledge resources.

In this thesis, we propose a model for semantic digital libraries with a geospatial context and a definition of coverage as key concept. We present the document and spatial resource model. We define the annotation model and more particularly the geographic coverage that detail and define the location of each resource taking into account its type. Finally, we present the query model and matching process where the geospatial context is an essential feature.

To validate this model, we develop some use cases and implementation. We first focus on annotating documents and precisely locating the documents within the spatial resource. To do so, we describe the implementation of the annotation model, presented in the digital library model, especially the geo- semantic knowledge resources alignment.

Then we present the methodology and implementation of a new technique to extract geographic information and place semantic from tags issued of volunteered geographic information (VGI) sources. This technique is based on a categorisation system, with a non-statistical knowledge-based approach. This extraction can partly automate the definition of the geographic coverage for the digital library resources, or be used to enhance semantically or complete 3D models and geo services.

Remerciements

Je tiens tout d'abord à remercier Philippe sans qui je n'aurais rien commencé ni fini. Merci pour ton soutien et ta patience au jour le jour pendant toutes ces années.

Merci aussi à mes parents, ma sœur et ma grand-mère pour leur soutien et leurs encouragements. Merci à mon oncle pour ses conseils et nos échanges.

Merci à mes collègues pour tous les moments partagés, les échanges, l'entraide et les conseils.

Merci à toute l'équipe du CUI pour leur aide infaillible et grâce à qui la vie d'un chercheur est bien plus facile. Merci tout spécialement à Marie-France Culebras, Nicolas Mayencourt, Daniel Aguillero et Elie Zagury.

Merci à tous les membres de mon jury pour vos conseils et pour avoir lu et corrigé ce travail : Les professeurs Dimitri Konstantas, Giovanna Di Marzo Serugendo, Claudine Metral et spécialement Javier Nogueras-Iso de l'université de Zaragoza.

Je tiens aussi à remercier spécialement mes directeurs, Laurent Moccozet et Gilles Falquet. Merci de m'avoir permis de participer à cette aventure et de m'avoir fait confiance. Merci pour votre accompagnement, votre disponibilité, pour nos moments d'échanges et pour m'avoir permis malgré mes doutes d'en arriver là. Merci aussi à vous pour m'avoir permis d'ouvrir mes horizons en participants à des projets très enrichissants. Laurent pour m'avoir permis de collaborer sur les projets e-learning avec toi, et Gilles pour notre collaboration en formation continue.

Chapter 1
Introduction

1.1 Collaboration and Resource Sharing within a Spatial Context

Nowadays, more and more tasks rely on collaboration and more importantly in collaboration between different domains of activity. For example, in the urban or construction domain, different professions must access the same resources in order to take a decision, like politicians, electricians, builders, solicitors... Sharing knowledge has always been a key issue in research. With the democratisation of digital tools, more possibilities have emerged to solve this issue.

In a large-scale context, digital libraries (DL) are a renowned tool for storing and managing documents and resources. Candela *et al.* have defined the main concepts and foundation of digital libraries. They describe digital libraries "as a tool at the centre of intellectual activity having no logical, conceptual, physical, temporal, or personal borders or barriers to information" [1]. Today digital libraries are capable of handling a wide range of resources, and can manipulate them within complex processes. They also handle user management and collaboration through digital library management system (DLMS). Many digital libraries system (DLS) and infrastructure have been presented in recent research such as BRICKS [2], with the introduction of semantic as seen in 5S framework [3] or Inspire [4]. Example of cross-domain semantic DLS implementation can be seen in the Papyrus [5] project which gathers the history and news domain in a news archive library.

Digital libraries can be enhanced with geographic information system (GIS), to handle the storage and usage of spatial data. In the context of decision-making and knowledge sharing, GIS allows to spatially contextualise numerous types of information such as images, web pages, text documents, etc. The spatial contextualisation enables collaboration between different domains of activity. The geographic axis is a transversal aspect unrelated to any domain but used by many. As a common ground for visualisation and browsing, the spatial context is often translated in existing GIS through a cartographic 2D map of the world.

The combination of GIS and digital library is a pertinent solution for the design of a collaborative cross-domain tool that can manage heterogeneous resources with a spatial axis.

1.2 Motivation and Problem Statement

Nowadays the task of decision taking is often related to many domains of activities and often needs to handle heterogeneous corpora of resources. Decision-making is a key issue in our society, with concepts like smart cities that aims at finding ways to manage current challenges such as ecology, governance, technology, data accessibility, economy, and social implication. Albino *et al.* [6] present the different existing definitions of smart cities in the literature, and some existing projects. The authors output the main characteristics of smart cities as: an infrastructure that enables social development and political efficiency; the urban development through business or creative activity; the social inclusion of resident and the social capital of the city; and the environment as strategic component for the future. Figure 1 shows the infographic representation of smart cities domain of applications, from the World Smart City[1] community built by ISO[2], ITU[3] and IEC[4], the three global standard institutions.

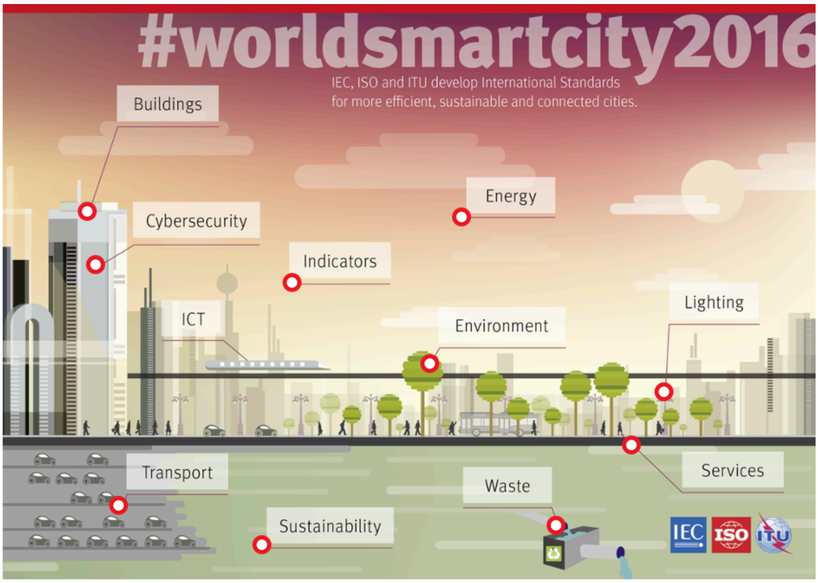


Figure 1 Smart city infographic

In this context we have seen the emergence of initiatives to share data openly. In May 2015, the European Council voted the open access to all scientific papers by

[1] http://www.worldsmartcity.org
[2] http://www.iso.org/iso/home.html
[3] http://www.itu.int/en/Pages/default.aspx
[4] http://www.iec.ch

2020[5]. They aim that open access to scientific publications will lead to optimal re-use of research data, and that embedding open science in society will make science more responsive to societal and economic expectations. The open access policy supports the open science[6] initiative.

More globally, the European Union legislates since 2003 on open data in Europe and proposes an open data portal[7], since 2012, to freely access datasets from the EU institutions. It also lists a series of applications, third parties or not, that uses the available datasets. Those applications cover a large range of application domain: space research, pharmaceutical, marine biology, quality of life, law, integrity and corruption watch, *etc.* Each government member of the EU also provides its own open data portal. The open data portal in the US was open in 2009, following the signature of President Obama of the "Memorandum on Transparency and Open Government"[8]. Open data platforms propose datasets in structured format such as json, csv, xml, kml, geojson... or unformatted such as pdf. The users of such systems, must then process the raw data to render it readable, *e.g.* turning a dataset into a graphic or integrating the set to a map when the data contains explicit geographic information.

As we have seen the data is available, through digital libraries, freely for anyone to use. Another aspect of sharing and openness is the open consultation where citizens are openly offered an input in the process of decision-making. For example in the European Union platform, at this date three open consultations are currently running[9] on diverse European research programs. In Europe, the British government is very active in this area, and proposes at this date 94 open consultations on diverse laws, regulation or other government decisions. In the same context, the IEEE[10] organisation launched the smart city challenge in 2014 where users could propose solutions for any cities on four domains: energy, communication infrastructure, traffic systems or buildings. For those consultations, the available additional resources are always of the textual form like pdf or links to the web.

Some cities and territories have also implemented a GIS services for users to visualise geo-localised public information. For example land registry services like the SITG [11] (Système d'Information du Territoire à Genève) in Geneva,

[5] http://www.consilium.europa.eu/en/meetings/compet/2016/05/26-27/

[6] https://ec.europa.eu/digital-single-market/en/open-science

[7] https://data.europa.eu/euodp/en/data

[8] https://www.whitehouse.gov/the-press-office/transparency-and-open-government

[9] https://ec.europa.eu/research/consultations/index.cfm?pg=list

[10] https://www.ieee.org/index.html

[11] http://ge.ch/sitg/

Switzerland, or the IGN (institut national de l'information geographique et forestière) in France with their geoportail[12], offer access to open data and proprietary data through a cartographic service. They processed the datasets into cartographic layers and enable users to create their own map by matching different layers. Both systems also provide a 3D view of their territory. However, those services are usually national services and so only propose certain dataset and not all of them are accessible to everyone, but some are reserved as a paid service.

Complementary to the previously described systems, there are also systems called Public Participation GIS or PPGIS as detailed in [7]. Those systems are a geospatial tool to inform planning processes with public knowledge, as it asks users to provide geographic information about their perception of a place. They are great systems to involve users in decision-making process, such as the open consultations.

In the context of decision-making, the user should be presented with all the necessary information contextualised, with no need for him to process them in order to understand them. Furthermore, if we carry on the example of the open consultation, citizens should be presented other law texts in relation with the one currently discussed. The links and relations to other text or complementary resources should be made clear in a way that contextualise the text currently examined by citizens, in order for their decision-making to be simplified.

In Figure 2, we depict an example of stakeholders and resources for the process of decision-making for a given task in a urban context. As we can see in our example, four stakeholders are implicated: a citizen, a politician, a urban planner and an ecologist. In order to make their decision, they need information contained in three corpora: institutions, transportation and ecology, each of them holds heterogeneous resources such as maps, text documents, images, videos, *etc.*

[12] https://www.geoportail.gouv.fr

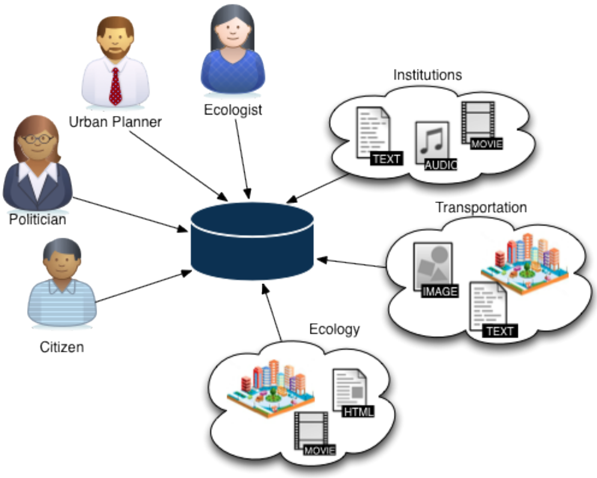


Figure 2 Illustration of decision taking in an urban context

This example depicts clearly the need for such systems to be able to handle multi-domain knowledge and heterogeneous types of resources. It is rare that a decision can be made efficiently by only taking into account the domain in question and not the surrounding implications. The stakeholders will also benefit to be able to filter within those resources those, which are relevant for their task according to their geographic and temporal validity.

Knowledge can vary temporally and spatially; a term or its definition can be valid in a given geographic area and in a given temporal range. For example, the term “trunk” defines a car boot in the U.S. whereas in the U.K. it defines a luggage. It is thus important to take into account this variation of scope in knowledge to be efficient. In digital libraries, domain is represented by its knowledge under the form of ontologies or thesauri for example. The handling of multiple domains in digital libraries can be done through the knowledge of each domain. Using ontologies as knowledge base, their alignment links the library’s domains together.

We will use in this thesis the urban domain as context for our research and example as it reflects the use of spatial axis and is a multi-disciplinary domain. As in our example in Figure 2, a multi-disciplinary domain implies a heterogeneous public that must be able to handle interaction of different vocabularies and knowledge.

In this work, we seek to establish precise links between a document and city objects that are either directly referenced in the document or considered as relevant. We aim to be able to group those relevant city objects using logic paradigm and semantic entities.

We propose a model of semantic digital library supported by GIS to answer the following research questions:

1. How to present, qualify and define the geo-spatial context of any resource: text, image, dataset, 3D models *etc.* in digital libraries?

Our contribution, presented in Chapter 3, is a proposed solution for this question. It defines a model of digital library with a semantic core, based on geographic semantics. We present the document and spatial resource model. We define the annotation model and more particularly the geographic coverage that details and defines the location of each resource taking into account its type. Finally, we detail the query model and matching process where the geo-spatial context is a key feature. They are both part of the information retrieval process, the later to define the interest of the research, and the former to filter the available resources according to the query.

2. How to validate this model? How to use this model to localise documents in 3D scenes and to semantically enhance 3D models?

We developed different use cases and implementations to demonstrate the feasibility of our model. In Chapter 4, we focus on annotating documents and precisely locating the documents within the spatial resources. We describe the implementation of the annotation model and notably the ontologies alignment. In Chapter 5, we present the methodology and implementation of a new technique to extract geographic information and places semantics from tags issued from volunteered geographic information (VGI) sources. This extraction can then be used to enhance semantically or complete 3D models and geo services.

In the following chapter, we present the state of research on GIS, geographic information retrieval, and the spatial visualisation of documents in 3D environments.

Chapter 2

State of the Art and Related Work

The geospatial and temporal contextualisation of information has emerged as an important research and development direction to increase the quality of search engines. Well-known web search engines such as Yahoo![13] or Google[14] have already implemented techniques to take the user location into account when processing queries. In fact, there exist many systems that enable users to query a repository of documents according to spatial and/or temporal contexts. Those systems are also known as Geographic Information System (GIS) and Geographic Information Retrieval Systems (GIR). Jones *et al.* [8] briefly present the current issues and state of GIR domain. They highlight the following key issues: the detection and disambiguation of geographic references, the interpretation of fuzzy geographic terminology, spatial and textual indexing, geographic relevance ranking and interfaces. We will address in this work the issues of the detection and disambiguation of geographic references and the interpretation of fuzzy geographic terminology.

GIR systems merge traditional contextualisation information retrieval (IR) issues with the one brought by the geographic and/or temporal dimension. The result ranking in GIR is not only influenced by the keyword search, but also by the matching of the query and the document geographic and temporal scope. This brings forward the need to define a matching algorithm that aggregates the keyword matching result with the scope matching results. The aggregation of both those results must take into account the fact that they are both issued from different scales, and so they cannot simply be added together.

Traditional IR is based on the appearance of the query keywords in the document. However, the temporal and geographic scopes need to be extracted as annotation or metadata of the document. The identification of the scopes can be done in different ways. In the case of web resources, as cited in the SPIRIT project [9], the web scope is defined as "*the geographic area the creator of the web resource intends to reach*". The authors propose to use the link/URL structure of the resource and describe two steps to identify the resource location. First, the *Power* value is the fraction of web pages from the location that should contain a link to the resource. So *Power* reflects the interest for the resource in a given location. Then, the *Spread* value measures the distribution of the *Power*

[13] http://local.yahoo.com

[14] http://google.com

for the resource. A location is in the scope of the resource if the *Spread* is over a given threshold and if the location ancestor's *Spread* value is below this threshold. For example, if for a given resource, the *Spread* value is high for Geneva city but low for Geneva Canton, then only Geneva City should be considered in the scope of the resource. The authors also propose to extract geographical scope from the resource content. Such methods can be applied to any type of resources and are detailed below.

The easiest and simplest way to identify the geographic and temporal scopes is to extract geographic and/or temporal entities from the document metadata, as used in the GeoTracker, VisGet and World Explorer projects; or extract the information from its content via identification techniques from the named entity recognition and classification (NERC) field [10] such as named entity recognition (NER) or part-of-speech (POS). The NER technique can identify entities like persons, organisations, locations, expressions of times, *etc,* and the POS identifies similar grammatical and syntax entities such as noun, verb, adjective, adverb, pronoun, *etc.* However, these techniques can be too vague on their own, as it will gather all geographic/temporal entities mentioned in the document and not only those specific to its scope. A way to make up for this flow is to combine them with manual annotation, or using ontologies, thesauri or databases built by experts. The entities extracted using NER are filtered using the ontology or thesaurus, as seen in the SPIRIT, STEWARD and GIPSY projects. Lastly, the NER technique can be associated with a spatial expression interpretation process to decode expressions such as "south of", or "adjacent", *etc.* as used in the GeoSem, PIV and SINAI projects.

The last important point is the question of the documents' presentation. In GIS the geographic and/or temporal aspect must be explicit in the query results display. Likewise, the interface for browsing and querying the repository should be adapted. A common way to explicit the geographic and or temporal dimension is for systems to display the documents on 2D maps of the world according to their scope, as it is done in the GeoTracker, STEWARD, PIV and GeoWorlds projects. Most use a point fix in space, others a polygon as a visual footprint for the document like it is done in the PIV project. Another often-used, complementary or stand-alone, way of displaying those aspects of the repository is to show them through the querying interface. As seen in the World Explorer and more particularly in the VisGet project, a time slider and an interactive map are used to define the query scope. The user navigates the map and uses the zoom to display and select the geo scope in the map window. The selected documents are then often simply listed and not integrated into the map. However, other less common visualisation techniques can be seen such as the GIPSY index result. This project uses a 3D map canvas of the document-wide scope and highlight through elevation picks the precise entities composing the document scope. An example is shown in Figure 4.

We present below in more detail, the list of GIR systems and projects we have previously cited.

2.1 Geographic Information System and Geographic Information Retrieval Example

The SPIRIT project [11], [12] proposes an information retrieval system that takes into account the geographic semantics of queries and documents. For instance, a query about towns in Switzerland should retrieve documents about Geneva, Zurich, *etc*. The system is based on geographic parsing and indexing of documents that assign a spatial footprint to each one of them. The document footprint is computed according to the place names extracted from the document content and matched with the geo-ontology. The query is given in the text, in three parts: subjects, place name and relation, by the user. The system then translates the place name and spatial relations into a geometric footprint for the query based on a spatial ontology. The query-document matching process combines textual matching with geographical matching. The SPIRIT project introduces the notion of query footprint to indicate the portion of the map relevant to the user's query. This work also deals extensively with the resolution of the frequent ambiguities that arise in the naming of geographic entities.

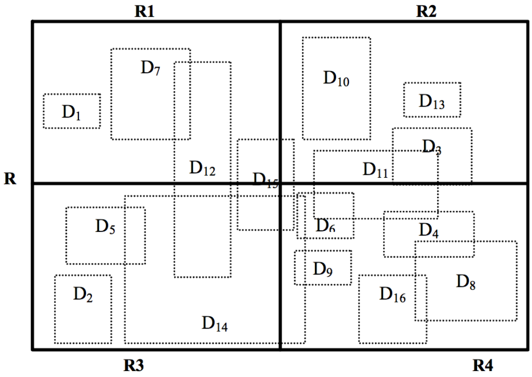


Figure 3 SPIRIT document search space

The GIPSY [13], Georeferenced Information Processing System, is a non-domain specific system that geographically indexes full-text documents. This system does not provide a querying engine or browsing interface, as it is solely designed for indexing. The algorithms determine coordinates from place names in text, with the use of a geographic thesaurus. The thesaurus contains 200'000 entries such as place names, feature types or land use types. The algorithm can interpret relations to place names such as *south of*, *adjacent*... The system returns a 3D grid, composed of a superposition of polygons, of the general location with picks where areas are identified in the text. An example of the output is shown in Figure 4 from [13] and labelled as: "*Surface plot produced from the State Water Project text which talks about Santa Barbara County, San Luis Obispo, and the Santa Ynez Valley area at some length*".

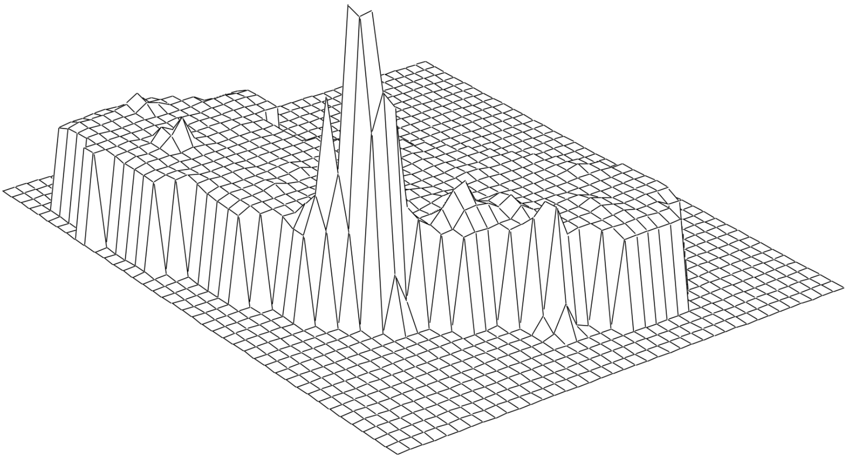
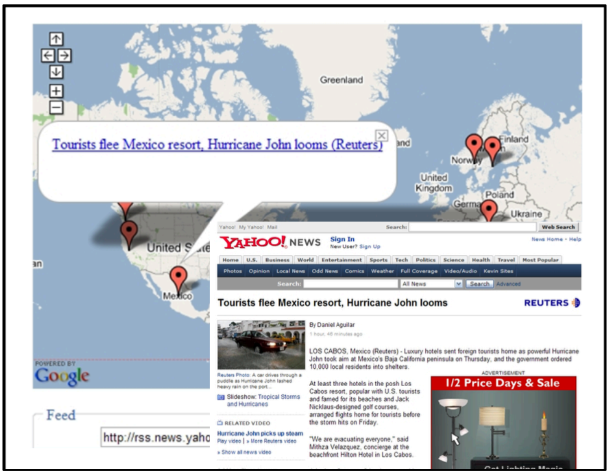


Figure 4 Example of GIPSY index output for a document.

The GeoSem system [14] extracts and interprets the spatial expressions in documents and document passages, and allows the query and ranking of those documents and passages. The spatial expressions are more complex than simple place names. The system can interpret spatial operations such as "north of", and spatial expressions such as "all the canton of Switzerland". The system was implemented using the LinguaStream[15] platform [15] to perform the semantic analysis of the geographic expressions. GeoSem returns extracts of documents in response to a geographic query, as each extract can be assigned a specific footprint.

GeoTracker [16] is a middleware system for RSS feed aggregator and browser. It enables users to query on a geographic and temporal axis by presenting the feed items on a world map. There is no query properly speaking but each user defines its profile with its interest. They make the assumption that the location information for each RSS feed item is explicitly given in the item. If many locations are present in the item, each will receive a pin linking to the item. The output is shown in Figure 5 as presented in [16].

[15] Available at: http//users.info.unicaen.fr/fbilhaut/linguastream.htm



Tourists flee Mexico resort, Hurricane John looms (Reuters)

Tourists flee Mexico resort, Hurricane John looms

By Daniel Aguilar

1 hour, 46 minutes ago

LOS CABOS, Mexico (Reuters) - Luxury hotels sent foreign tourists home as powerful Hurricane John took aim at Mexico's Baja California peninsula on Thursday, and the government ordered 10,000 local residents into shelters.

At least three hotels in the posh Los Cabos resort, popular with U.S. tourists and famed for its beaches and Jack Nicklaus-designed golf courses, arranged flights home for tourists before the storm hits on Friday.

"We are evacuating everyone," said Mithza Velazquez, concierge at the beachfront Hilton Hotel in Los Cabos.

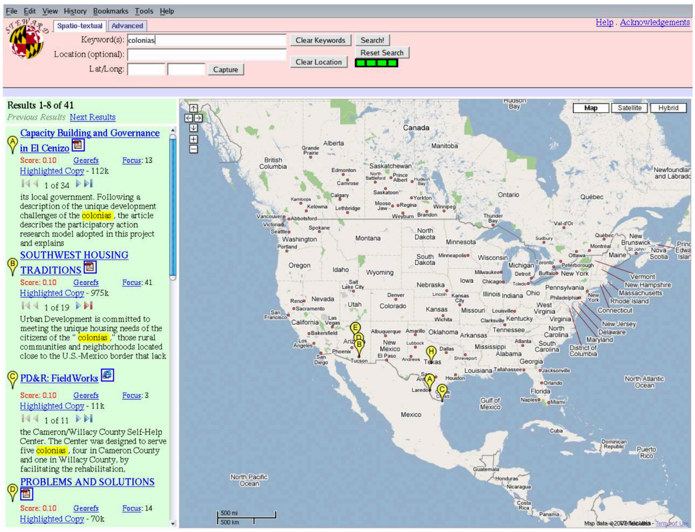
Reuters Photo: A car drives through a puddle as Hurricane John lashed heavy rain on the port...

Feed: http://rss.news.yaho

Greenland
Finland
Norway
United Kingdom
Poland
Germa
Ukraine
United S
Mexico
POWERED BY Google
Yahoo! My Yahoo! Mail
Search:
Web Search
YAHOO! NEWS
Sign In
New User? Sign Up
News Home - Help
Home U.S. Business World Entertainment Sports Tech Politics Science Health Travel Most Popular
Photos Opinion Local News Odd News Comics Weather Full Coverage Video/Audio Kevin Sites
All News
Search
Advanced
REUTERS
Slideshow: Tropical Storms and Hurricanes
RELATED VIDEO
Hurricane John picks up steam
Play video | » More Reuters video
» Show all news video
ADVERTISEMENT
1/2 Price Days & Sale

Figure 5 GeoTracker user interface

STEWARD [17] is a spatio-textual search engine for unstructured text documents, particularly web pages. Contrary to other similar search engines it does not assign the same scope to web pages according to their link structure, but fetch in every document references to geographic locations and register them as their scope. Each georeference registered as the scope is assigned a weight. The system uses a hybrid approach of part-of-speech and named-entity recognition techniques to identify the georeferences in the document. The process also contains a semantic disambiguation algorithm. The user can query the system using both location and keywords. An example of STEWARD user interface is shown in Figure 6.



Keyword(s): colonias

Location (optional):

Lat/Long:

Results 1-8 of 41

Previous Results Next Results

A Capacity Building and Governance in El Cenizo
Score: 0.10 Georefs Focus: 13
Highlighted Copy - 112k
1 of 34
its local government. Following a description of the unique development challenges of the colonias , the article describes the participatory action research model adopted in this project and explains

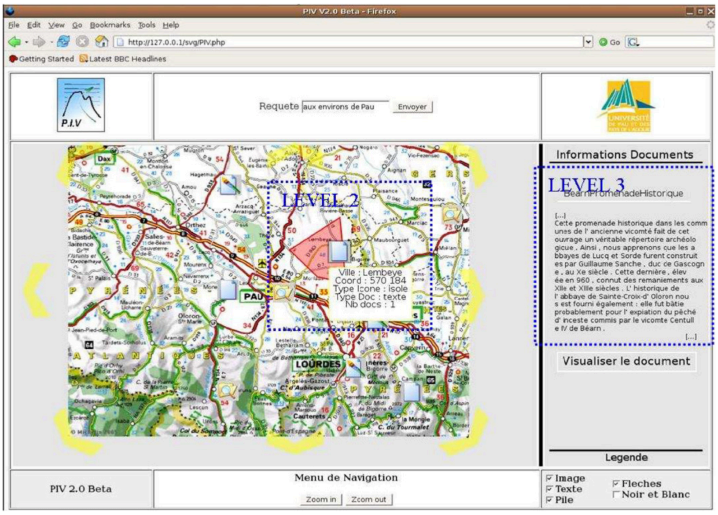
B SOUTHWEST HOUSING TRADITIONS
Score: 0.10 Georefs Focus: 41
Highlighted Copy - 975k
1 of 19
Urban Development is committed to meeting the unique housing needs of the citizens of the " colonias ," those rural communities and neighborhoods located close to the U.S.-Mexico border that lack

C PD&R: FieldWorks
Score: 0.10 Georefs Focus: 3
Highlighted Copy - 11k
1 of 11
the Cameron/Willacy County Self-Help Center. The Center was designed to serve five colonias , four in Cameron County and one in Willacy County, by facilitating the rehabilitation,

D PROBLEMS AND SOLUTIONS
Score: 0.10 Georefs Focus: 14
Highlighted Copy - 70k

Figure 6 STEWARD user interface

In the cultural heritage domain, the PIV project [18] proposes a repository of spatio-temporalised contents that gathers heterogeneous types of resources that represent human modes of expression. The system builds a semantic tag system to associate direct and indirect locations to the documents as well as their evolution in time or temporal references. Only one location is kept for each document scope. Gaio *et al.* have developed a process to interpret spatial expressions and translate them in a polygon. The PIV project also have implemented the LinguaStream platform [15] for their geo-semantic textual data process. Gaio *et al.* have developed a query engine to allow users to define a location of interest when creating a query. The system only manages geographic content and so does not include content and domain search, and relies on other library management system to do so.



http://127.0.0.1/svg/PIV.php

Requete aux environs de Pau

Ville : Lembeye
Coord : 570 184
Type Icone : isole
Type Doc : texte
Nb docs : 1

BearnPromenadeHistorique

[...]
Cette promenade historique dans les communes de l' ancienne vicomté fait de cet ouvrage un véritable répertoire archéologique . Ainsi , nous apprenons que les abbayes de Lucq et Sorde furent construites par Guillaume Sanche , duc de Gascogne , au Xe siècle . Cette dernière , élevée en 960 , connut des remaniements aux XIIe et XIIIe siècles . L' historique de l' abbaye de Sainte-Croix-d' Oloron nous est fourni également : elle fut bâtie probablement pour l' expiation du péché d' inceste commis par le vicomte Centulle IV de Béarn .
[...]

☑ Image ☑ Texte ☑ Pile ☑ Fleches □ Noir et Blanc

Figure 7 PIV user interface

Geooreka [19] is a web search engine integrated with a GIS database. Users select an area on a map to query the system. The system uses the zoom level chosen by the user to determine the type of place to use as the query scope (*i.e.* country, region, city, *etc.*). The higher the zoom level, the fewer toponyms are selected. The selected toponyms are then extracted and associated with the query keywords as a pair (Theme - Toponym). The pairs are then filtered and compared to each other processing a probability weight. For the 20 best-scored pair, they compute a Borda count [20] to determine which pair will be used to query the web search engine Google or Yahoo!.

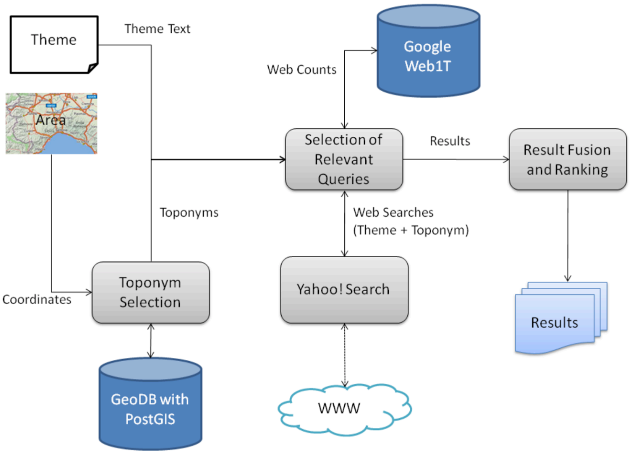


Figure 8 Geooreka architecture

The SINAI GIR system [21] uses GATE [22] to detect geo-entities via NER, verified using Geonames. To complete the NER results, the SINAI system has developed a process to detect and recognise topological spatial relationships, and uses Lemur[16] as an IR engine and to build a document index. They re-rank the IR results using their document index, and Geographical index built using GATE NER and validated by Geonames. By taking into account and interpreting the spatial expression in the query and the type of the locations in the query, the filtering process calculates the coordinates of the corresponding bounding box and filters the pre-selected documents.

GeoWorlds [23] is a collaborative GIR system. It allows users to select a geographic region on a GIS display and returns to the user a list of documents associated with the chosen region. Inversely, once a document is selected, the region attached to it is highlighted in the GIS. The documents are harvested from information spaces maintained by specialised groups and data warehouses. The document manager module of GeoWorlds allows users to organise, and annotate the documents. GeoWorlds focuses on the disaster domain and provides a data analysing module. In Figure 9, we can see the process to generate an analysis report with GeoWorlds.

[16] https://www.lemurproject.org/lemur.php

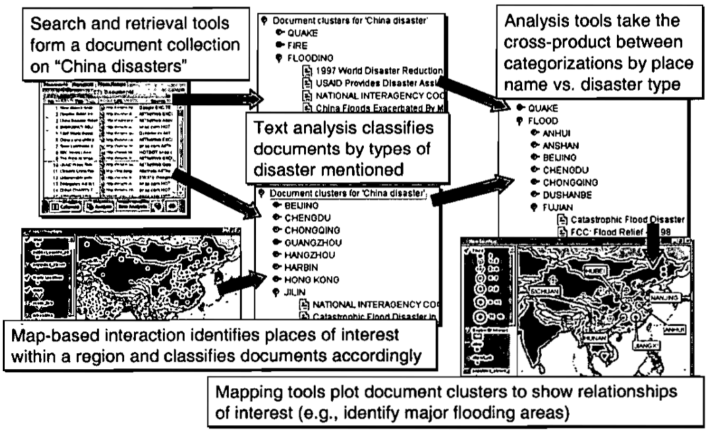
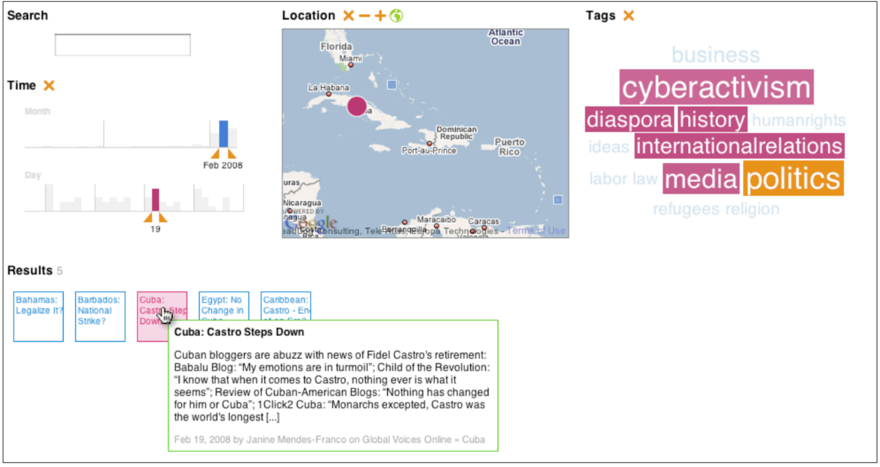

Search and retrieval tools form a document collection on “China disasters”
Document clusters for 'China disaster'
QUAKE
FIRE
FLOODING
1997 World Disaster Reduction
USAID Provides Disaster Assi
NATIONAL INTERAGENCY COO
China Floods Exacerbated By M
Analysis tools take the cross-product between categorizations by place name vs. disaster type
Text analysis classifies documents by types of disaster mentioned
Document clusters for 'China disaster'
BEIJING
CHENGDU
CHONGQING
GUANGZHOU
HANGZHOU
HARBIN
HONG KONG
JILIN
NATIONAL INTERAGENCY COO
Catastrophic Flood Disaster in
QUAKE
FLOOD
ANHUI
ANSHAN
BEIJING
CHENGDU
CHONGQING
DUSHANBE
FUJIAN
Catastrophic Flood Disaster
FCC: Flood Relief 98
HUBEI
SICHUAN
NANJING
HUNAN
JIANGXI
ANHUI
Map-based interaction identifies places of interest within a region and classifies documents accordingly
Mapping tools plot document clusters to show relationships of interest (e.g., identify major flooding areas)

FIG. 11: *An example of geo-spatial information management: identify recurring disaster areas in China*

Figure 9 GeoWorlds process example

The VisGet project [24], seves to browse news item from RSS feed. The system is based on the three following axis: temporal, geographic and topic. To add geographic information to an RSS item not specifying any, the project uses location information from the textual part of the item and identifies the corresponding coordinates with Geonames. The query time dimension is defined using a bar chart. The query's spatial dimension is defined zooming on a 2D map of the world. Finally, the topic dimension is defined using a tag cloud. The query's dimensions are displayed in Figure 10.



Cuba: Castro Steps Down

Cuban bloggers are abuzz with news of Fidel Castro's retirement: Babalu Blog: “My emotions are in turmoil”; Child of the Revolution: “I know that when it comes to Castro, nothing ever is what it seems”; Review of Cuban-American Blogs: “Nothing has changed for him or Cuba”; 1Click2 Cuba: “Monarchs excepted, Castro was the world's longest [...]

Feb 19, 2008 by Janine Mendes-Franco on Global Voices Online » Cuba

Figure 10 VisGets user interface

The World Explorer project [25] allows users to geographically query the Flickr photo database as shown in Figure 11. A visualisation tool shows the high-scored tags on a world map according to the zoom level. The tags are chosen to identify a region according to a statistical process, taking into consideration the number of unique tag's creator as well as tags repetition. A user can then retrieve the pictures related to a tag and a place. This project uses localisation and geographic information about concepts to retrieve geo-localised resources. Users can only browse the map, not input search keywords.

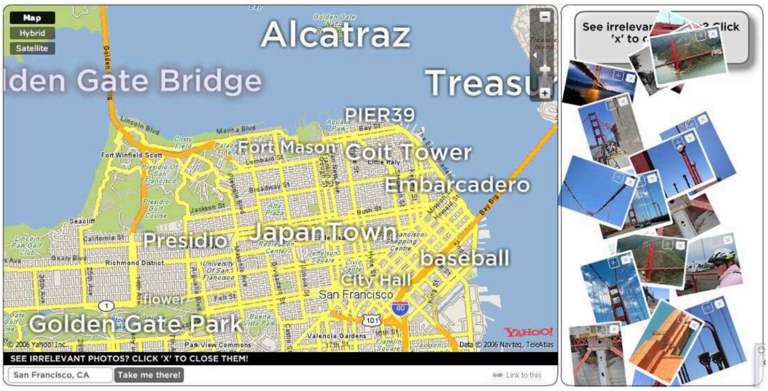


Figure 11 World Explorer user interface

As we have seen, GISs are used to contextualise spatially the resources and information. In GIS the geocoding process essentially consists, as described in the system presented before, in finding and extracting place names, geo entities or simply coordinates from textual data, such as street addresses or building names from the document content. The geographic footprint of the resource is then constructed by grouping the found entities. The problem we wish to address here is similar, but instead of finding place names in a document, we seek to establish precise links between a document and city objects that are either directly referenced in the document or considered as relevant. We aim to be able to group those relevant city objects using logic paradigm and semantic entities.

Each of the previously presented systems is either centred on a specific domain such as history or news, or not related to any domain. In this research we aim at enabling semantic cross-domain collaboration, to allow users from different expertise domains to actively and efficiently share knowledge. We will present in the following chapters a multi-domain integration model through ontology alignment.

Finally, contrary to the majority of the systems presented here that handles a single format of documents; we propose a model that can handle heterogeneous documents.

2.2 Indexing, Matching and Ranking

In this section we review in more detail the indexing, matching and ranking for the systems presented in section 2.1.

The SPIRIT project runs two indexes: a textual index that processes both spatial and non-spatial terms, and a spatio-textual index that combines a text and spatial indexing according to the documents footprint. In the first index, the matching depends on an exact match between the query terms and the document, whereas the former index uses geometric footprint for matching. The text relevance is based on the BM25 algorithm [26], and the spatial relevance on the distance between the query and the documents footprints. The two relevance scores are combined to form the final relevance score. The ranking is done using the geo-ontology to retrieve geometric footprint of places and comparing them to the query's footprint geometry.

The GeoSem project uses a linguistic analyser of spatial expression and text analysis tool to extract geographic location. It then translates the found location into a semantic representation, to use as indexes for the documents and documents' passages. The relevance is calculated depending on quantification, whether the query mentions quantity concept, or granularity. The quantification relevance weight is processed using probability such as "a quarter" will be 25% relevance. The granularity relevance is processed using the hierarchy between the geo entities from the queries and the document index.

The STEWARD project handles web documents. The indexing process stores them in a database with its URL, metadata, the ASCII and HTML version of the document. The geo-location of the resource is extracted and selected using TF-IDF statistic along with NER and POS techniques and compared with a geodatabase to differentiate the geo entities from the other extracted entities. The ranking is calculated according to the frequency and distribution of the keywords and references to geolocation in the document.

The PIV project gathers the extracted spatial feature in an index. Each feature is stored with its name, its interpretation and its geometric shape. The geometric shape of each spatial feature is recovered using GISs as a point for a building, a line for a road, *etc.* Then its shape is simplified as a minimum-bounding rectangle (MBR). They developed an algorithm to determine how to transform the spatial feature MBR to comply/translate the spatial relations found in the document. The retrieval process selects the documents or paragraphs to return according to the mapping result between the query geo features to the document's geo scope. There is no ranking of the selected valid document or paragraphs.

The SINAI project uses GATE to detect geo entities in documents and validates the identification with Geonames and manual rules for spatial relationship identification. The project implemented the Lemur system as an information retrieval engine. The project holds two indexes: a document and a geographic index. The document index stores stem words for each document. A stem is the root or roots of a word, together with any derivational affixes, to which inflectional affixes are added. The geographic index stores the list of all locations detected in the collection. The documents returned by the lemur engine are re-ranked according to filtering rules. The new rank is influenced by the weight of the corresponding filtering rule.

The GeoWorld project returns results imported from web search engines. The results are indexed in a table with their URL, their source, and their title. Rows are sorted according to the search engine ranking. The system processes two classifications: a textual and a place name classification. The textual classification is done using keyword extraction. The place name classification is done from place names extraction in the document comparing them with the one extracted from the map (query). The final ranking is processed doing a cross product of the two classifications.

The VisGet project extraction process assumes that each RSS item contains a title, a description, tags, a date and time of publication and a geo-location. If no location is explicit, the system uses Geonames web service to extract geo information from the RSS and transforms it in GeoRSS. There is no ranking in the result display, and each RSS item is pinned to the map according to its location.

The World Explorer project uses photos, users and tags as its dataset. The indexing of the tags is done for each photo in the dataset. The tags are associated with the coordinates of the picture they were extracted from. The world is then divided into tiles on different granularity (zoom). For each tile the system retrieves a geographic cluster of photo and their tags. Each tag receives a score within each cluster. The score is computed using a combination of terms frequency, using TF-IDF, the number of times a tag is used in the cluster, and user frequency, the percentage of photographers in the cluster that uses the tag. If the tag's score is above a given threshold, the tag is selected to appear in the corresponding tile. The user queries the system by zooming on the map. The system retrieves between 1 and 4 tiles that fit the display area and shows the corresponding tag cloud on the map.

The GIPSY projects only handle the indexation process. The system extracts geo locations from the resource content and retrieves their corresponding coordinates from a thesaurus. The system generates a matrix as the index of the resource.

The matrix represents the geographic scope of the document as a 3D grid. Each extracted geo-location appears as a peak on the grid according to their coordinates.

The Geooreka project did not involve ranking or information retrieval processes.

2.3 Summary table

Below is summary table of all the previously presented systems.

	Geo tools	**Indexing, matching and ranking process**	**Document format handling**
SPIRIT [11,12]	Spatial ontology	Text and spatial indexing according to the document footprint. Two relevance's score (text and spatial) that are combined for the final ranking. Text matching is based on exact match. Spatial matching is based on inclusion according to a spatial ontology.	Web pages with footprints
GIPSY [13]	Geographic thesaurus	No query engine, the system handles the geographic indexing only. Extract the place names in the text and determined the coordinates using the geographic thesaurus.	Text documents
GEOSEM SYSTEM [14]	LinguaStream, linguistic analyser of spatial expressions.	Extracts spatial expressions from the documents, to index them. The system can be queried using geographic queries. Relevance is calculated using quantification.	Text documents

GEO TRACKER [16]	Embedded map for the user to query the system.	Middleware system that allows to query on spatial and temporal axis. Compares the location information embedded in the RSS items to the map selected by the user.	RSS feed
STEWARD [17]	Semantic disambiguation algorithm. Geo database	Spatio textual search engine. Generates documents scope from the body of the document using POS and NER. Each scope gets assigned a weight. The ranking is computed taking into account the frequency and distribution of the keywords.	Unstructured text documents (Web pages)
PIV [18]	LinguaStream	Builds a semantic tag system. Translate the document scope into an MBR. Retrieval process returns the matching between the document scope and the query geo features. There is no ranking of the results.	Heterogeneous type of resources
GEOOREKA [19]	GIS database	The query scope is determined by the zoom level set by the user. The higher the zoom the fewer the toponyms selected. The system associates the toponyms with the selected theme as pairs to query web search engines. Only the best pair (determined using a Borda count) is used for	No resources are directly handled.

		the final query.	
SINAI GIR [21]	GATE to detect geo-entites and verified by Geonames. Lemur system as an IR engine.	Each document gets a geographical and a document index. The system interprets spatial expression in the query to determine the query bounding box to filter the pre-selected documents.	Text documents
GEO WORLDS [23]	GIS display	Users can annotate and organise the documents. The system generates two indexes for each document using keywords extraction: textual and place names. The ranking is the result of a cross product using the two indexes.	Documents are web pages harvested from information spaces.
VISGET [24]	Geonames	Define the geographic scope of an item by extracting place names from its content and identifying them with Geonames. There is no ranking of the results.	News item in RSS feeds
WORLD EXPLORER [25]	Yahoo Where On earth ID (WOEID)	Tags are indexed on a map using the photo coordinates. The tags identify a region using a statistic approach, taking into account the number of unique tag's creator and the tag repetition. Users can retrieve photos using the tags on the map. The results are not ranked.	Photos and tags form Flickr

Most of the presented projects use as index a kind of geographic database that contains all the location associated with a document. The retrieval process is then either an exact matching or a testing for an inclusion with the query footprint. The query footprint is of textual or geometric form, *i.e.* a polygon of coordinates.

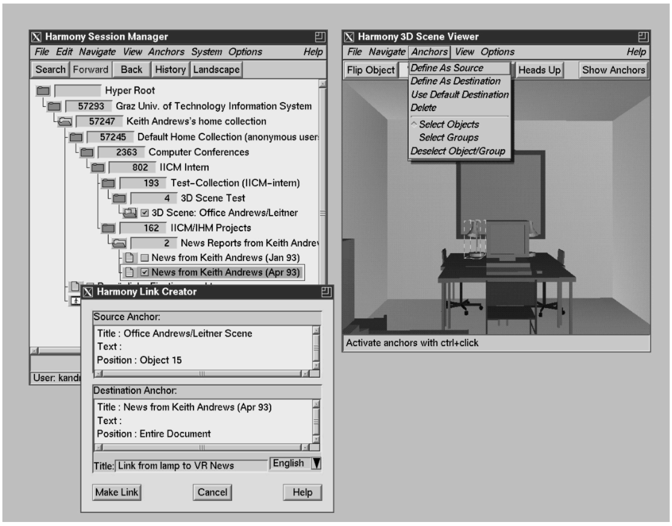
We propose a geographic indexing and geographic ranking process based on the semantic and geography of places. We wish to allow a finer querying mechanism that allows more complex geographic queries than an inclusion testing from a footprint polygon as most of the presented systems do. To do so, we introduce a query and indexing engine based on a semantic tool that combines a geo gazetteer and a knowledge base of building and place semantics.

2.4 Spatial Visualisation of Documents

As we have seen in the previously described projects, the most common way of displaying the geographic dimension of a repository is through 2D maps. However, the most natural visualisation environment is a 3D or at least 2,5D environment as the world is in three dimensions. 3D environments have been used to organise documents such as in the Bead system [27], where Chalmers build a 3D landscape from the similarities and dissimilarities of documents from a corpus. The assumption is that a retrieval task is better achieved if the relationships within a corpus are visible.

Annotation and more precisely resource implantation in 2D scenes are done using single points in space or flat polygons on the map, whereas within a 3D scene it can be attached to 3D objects or parts of objects. There is non-negligible research in the area of annotation and data integration of 3D models. We show through the example presented below the use of 3D models and particularly 3D city models as tools for data visualisation and interpretation.

The Harmony project [28], proposes an anchoring solution to link a document repository to objects in a 3D scene.


Harmony Session Manager
File Edit Navigate View Anchors System Options Help
Search Forward Back History Landscape
Hyper Root
57293 Graz Univ. of Technology Information System
57247 Keith Andrews's home collection
57245 Default Home Collection (anonymous user
2363 Computer Conferences
802 IICM Intern
193 Test-Collection (IICM-intern)
4 3D Scene Test
3D Scene: Office Andrews/Leitner
162 IICM/IHM Projects
2 News Reports from Keith Andre
News from Keith Andrews (Jan 93)
News from Keith Andrews (Apr 93)
User: kand
Harmony 3D Scene Viewer
File Navigate Anchors View Options Help
Flip Object Heads Up Show Anchors
Define As Source
Define As Destination
Use Default Destination
Delete
Select Objects
Select Groups
Deselect Object/Group
Activate anchors with ctrl+click
Harmony Link Creator
English
Make Link Cancel Help

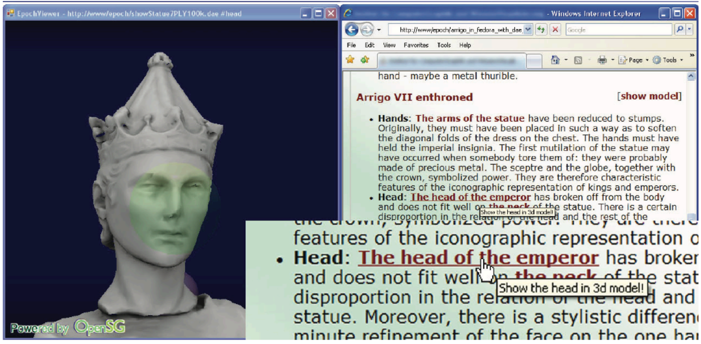
Source Anchor:
Title : Office Andrews/Leitner Scene
Text :
Position : Object 15

Destination Anchor:
Title : News from Keith Andrews (Apr 93)
Text :
Position : Entire Document

Title: Link from lamp to VR News

Figure 12 Harmony's anchoring example

More precisely, research on 3D annotation has brought forward ways of linking heterogeneous and dynamic information to 3D objects. Havemann *et al.* [29] describes a markup method which attaches information to parts of 3D models. This provides the possibility to associate hyperlinks and links to parts of 3D objects to web documents, as shown in Figure 13.


EpochViewer - http://www/epoch/showStatue?PLY100k.dae #head
Powered by OpenSG
- Windows Internet Explorer
http://www/epoch/arrigo_in_fedora_with_dae
Google
File Edit View Favorites Tools Help
Page Tools
Show the head in 3d model!

hand - maybe a metal thurible.

Arrigo VII enthroned [**show model**]

- **Hands**: **The arms of the statue** have been reduced to stumps. Originally, they must have been placed in such a way as to soften the diagonal folds of the dress on the chest. The hands must have held the imperial insignia. The first mutilation of the statue may have occurred when somebody tore them of: they were probably made of precious metal. The sceptre and the globe, together with the crown, symbolized power. They are therefore characteristic features of the iconographic representation of kings and emperors.
- **Head**: **The head of the emperor** has broken off from the body and does not fit well on **the neck** of the statue. There is a certain disproportion in the relation of the head and the rest of the

features of the iconographic representation o
• **Head**: **The head of the emperor** has broken
and does not fit well on the neck of the stat
disproportion in the relation of the head and
statue. Moreover, there is a stylistic differen
minute refinement of the face on the one ha

Figure 13 Arrigo Reloaded project example

If we focus on the geographic domain and more precisely the urban domain, we can find many projects implementing annotation and data integration in 3D urban models. The Topos 3D spatial hypermedia system [30], provides a geospatial interface, as shown in Figure 14, for the exchange and organisation of information and for facilitating collaboration among users. The example scenario given is the collaboration of parties during a building construction on site. Topos also integrates a GPS to enable the matching of the real site and its 3D representation. Their work combines spatial hypermedia, GIS and a collaborative virtual environment.

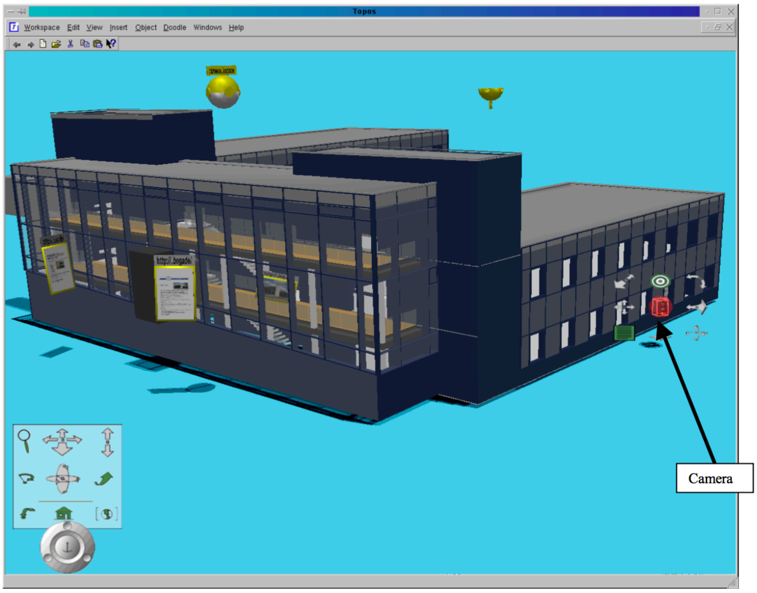


Figure 3: An example of the geo-spatial Topos interface. A layer, which superimposes homepages of institutions and people in our lab building, has been opened on top of the 3D model of our lab. The selected camera to the right represents a mobile user's location in the real world.

Figure 14 Topos geospatial interface

The use of urban 3D city models to facilitate the interpretation of data leads to a visualisation tool to display air quality data in a collaborative environment [31]. This system allows the comparison of different scenarios to facilitate the collaboration. An example of the data visualisation is depicted in Figure 15. A similar system to generate 3D noise calculation and simulation within a 3D city model can be seen in [31]. This system allows visualising the noise impact in the city from the street level up to the building roofs. Such system allows the simulation of urban modification to impact the noise level. Metral *et al.* [33] developed an ontology of 3D visualisation techniques in 3D city models. Depending on the dataset format and with the use of the ontology, systems will be able to select the most appropriate visualisation technique automatically.

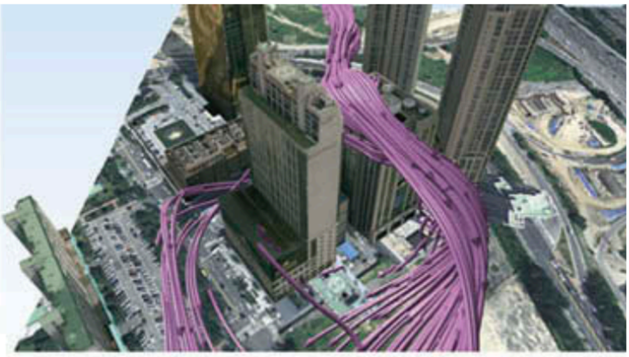


Figure 15 Example of air quality model visualisation

In the catalogue of existing 3D modelling languages, 3D semantic formats have been created to semantically enhance 3D representations. They integrate, within the language, descriptive semantics of the object or the domain. In the context of urbanism, 3D languages are used to model cities or neighbourhoods. In the context of GIS or urban GIS, user environment and/or repository's browsing interface, can be built using a 3D modelling language to create a 3D view of the repository scope.

In this context, the most pertinent language is CityGML, it is used to describe 3D city models. Kolbe describes the CityGML data model adopted by the Open Geospatial Consortium (OGC) [34]. This XML-based language brings semantic to the 3D shape and texture and so represents the following aspects of a city model: semantic, geometry, topology and appearance. The language can represent five levels of detail (LOD). The highest level allows for each building parts to be represented such as windows and doors and indoor details. It also enables the definition of groups and parts of buildings. The semantics describes the following most important geographic features: buildings, water bodies, vegetation, city furniture and land use as described in Figure 16. For example, the building semantics holds information on its function and its usage. Finally, this language allows the description of the objects' 3D spatial properties and interrelationships. It is a complete language and is now the international standard for representing, storing and exchanging 3D urban objects.

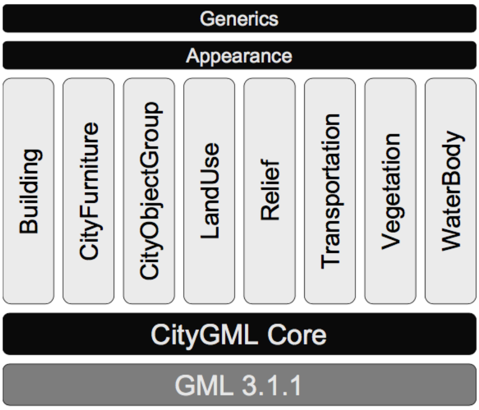


Figure 16 CityGML composition

Many models are already available in CityGML format. Goetz presents a method to automatically generate high-level CityGML, the level of detail (LOD) 3 and 4, level 4 being the highest available LOD including the interior definition [35]. The automated generation is done using crowd-sourced data from OpenStreetMap[17] (OSM) and the proposed feature service *IndoorOSM* [36], [37]. Goetz *et al.* present in [38] the automated generation of CityGML models from OSM for LOD 1 and 2.

As seen here, most of the GISs and GIR systems use 2D cartography as browsing and querying interface. However, a 3D environment is more natural and precise interaction among users, as a certain level of details can be achieved. For example using a CityGML model, each part of the building can be accessed as an object: a particular storey, door or window...

We aim to implement and use the advances in 3D annotation as seen in this section, by using 3D city models as browsing and querying interface. The corpus of documents will be attached correspondingly to our geographic scope definition to the objects or part of objects in the city model. Users will then be able to query or browse documents by navigating in the city model and clearly visualise documents' geographic scopes.

[17] https://www.openstreetmap.org

Chapter 3

Model

Many framework and models for digital libraries exist, such as the DELOS [39] and DL.org [40], the 5S framework [3], [41], and domain oriented models like Inspire [4] for high-energy physics and BRICKS [2] for Cultural Heritage.

With DELOS, and later DL.org as an enhancement of DELOS model, the authors propose a reference model of digital library as a result of the research on digital libraries from European researchers. DELOS defines digital libraries as a three-tier framework composed of a digital library management systems (DLMS) which provides the generic infrastructure, produces and administers the digital library system (DLS); a DLS which is the system that provides the functionality needed by the digital library; and the digital library which manages the content and provides the functionalities to the users. DELOS has defined the domains that compose the digital library as shown in Figure 17, and more precisely, the resource model in Figure 18. A DL domain is composed of resources. Resources are among others, annotated, described, and expressed using information objects in all their forms, such as documents, metadata, images, annotations, queries, results sets... A resource can be part of another resource and can be grouped in a set of resources.

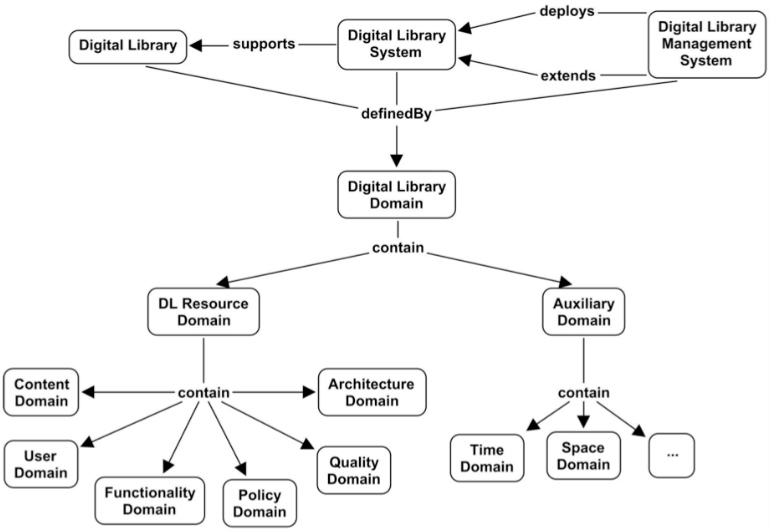


Figure II.2-1. DL Domains Hierarchy Concept Map

Figure 17 DELOS and DL.org domain concept map

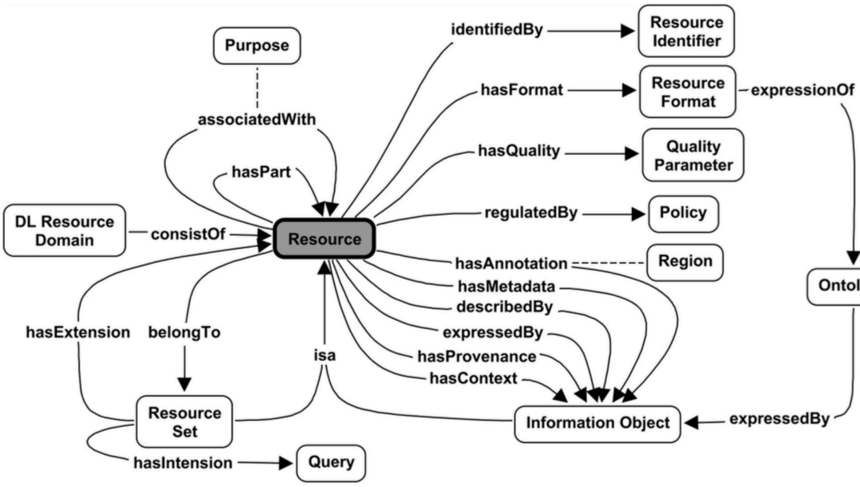


Figure II.2.4. DL Resource Domain Concept Map[13]

Figure 18 DELOS and DL.org resource domain concept map

The 5S framework divides DL in five layers: Streams, Structures, Spaces, Scenarios, and Societies as depicted in Figure 19 [42]. Streams are static or dynamic sequences of elements of any type, for example video delivered to users or document viewing. Structures specify how parts are organised as for example user relationships, taxonomies. Spaces represent set of objects along with the related operations like document space. Scenarios are sequences of events, which involve actions that alter a computation and influence future events, as workflow or dataflow.

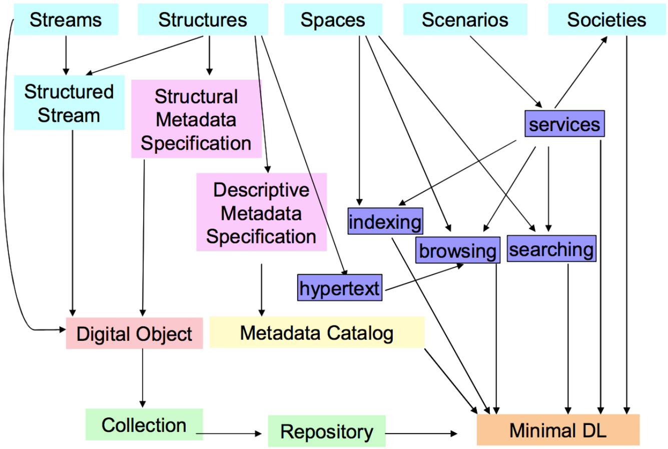


Fig. 1.3.1 5S definitional structure ([50])

Figure 19 5S simple library structure

Our research is based on the concepts and domains highlighted by both DELOS and 5S models. However, none of those generic models provide a geographic axis to the digital library. To fill this gap, we introduce in this chapter a model for the high-level repository model that includes a transversal geographic axis.

This repository is based on four elements as depicted in Figure 20. Those elements are the three repository resources and a transversal notion of spatiotemporal coverage. Our notion of coverage is defined as follow: *the geographic and temporal context of a resource refers to the spatial and temporal regions in which the resource is true or must be true in the real world or in a fiction.* We use here both meanings of “true”: something that is correct, accurate, or something that is real, genuine. The coverage model is detailed in section 3.2. The three repository’s resources are defined below.

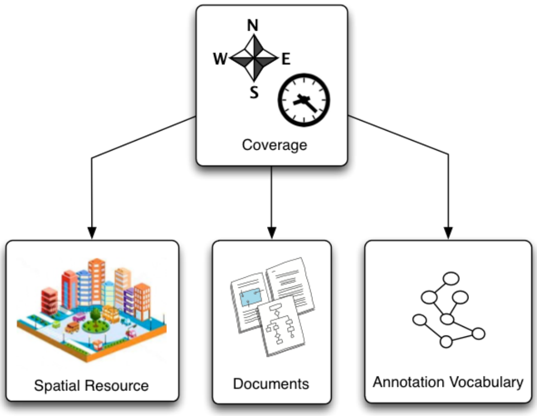


Figure 20 Repository basis bricks

3.1 The Repository Model

The repository is composed of three repository resources: a spatial resource, documents and the annotation vocabulary. As pictured in Figure 21, they are linked as follows: the annotation vocabulary annotates the documents; it also identifies the city objects from the spatial resource; the documents are located in the spatial resource. The following examples describe the links between the repository resources.

- `annotates(d`$_1$`, [‘‘health’’, ‘‘hospital’’, ‘‘medecine’’, …]);` where `d`$_1$ is a document and the concepts are issued from the annotation vocabulary.
- `identifies(obj`$_1$`, HUG);` where `obj`$_1$ is an object within the spatial resource and `HUG` is an instance that represents the public hospital in Geneva in the geographic ontology that composes the spatial vocabulary.

- `coverage(d₁, obj₁)`; the document d_1 is associated with obj_1, the object is defined as its spatial coverage.

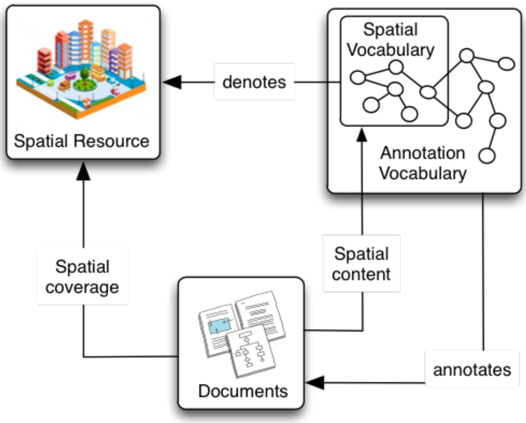


Figure 21 Repository resources connections

We first present the document model and annotation model.

3.1.1 Document Model

The document model represents the non-spatial resources stored in the repository. The document model and the document annotation model are inspired by the DELOS model as depicted in [39] and 5S framework [3]. Our major contribution is on the annotation model and particularly on the coverage annotation.

In DELOS, the space and time coverage in documents are stored in metadata using standards such as the Dublin Core (DC) [43]. Dublin Core proposes a "coverage" property and is composed of a space and a time attribute. It is defined as follows in the DC metadata registry[18]: "The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant". The spatial attribute is of range *dc:location*[19] which is defined as a spatial region or named place, and the time attribute is of range *dcterms:PeriodOfTime*[20] which is defined as an interval of time that is named or defined by its start and end dates. In other cases, such as in the 5S framework [3], system developed a facet called "Spaces" that contains metric or vector spaces which allow the location of a resource in 4D, the fourth dimension

[18] http://purl.org/dc/elements/1.1/coverage
[19] http://purl.org/dc/terms/Location
[20] http://purl.org/dc/terms/PeriodOfTime

being the temporal one. This facet is used in searching, browsing and indexing process by the digital library management system.

We propose in those models to integrate more finely the coverage aspect of a resource by using a semantic annotation model. The spatial coverage is a combination of both semantic instances, geographic features, and classes, geographic feature classes.

The documents are accessed by the users through the query engine and are positioned within the spatial resource in 3D. The term document refers here to all kind of texts, images, datasets, 3D models, *etc.* A document is composed of two parts: its content and its metadata, as shown in Figure 22. The metadata holds information such as title, language, date, creator, type, *etc.* Metadata are expressed in XML format, and gives information about the document as a whole. Many metadata schemas exist; we propose to use a generic one such as the one provided by Dublin Core that allows the description of a broad range of type of resources. The content is expressed in the language corresponding to its format (*e.g.*: pdf, csv, jpeg....).

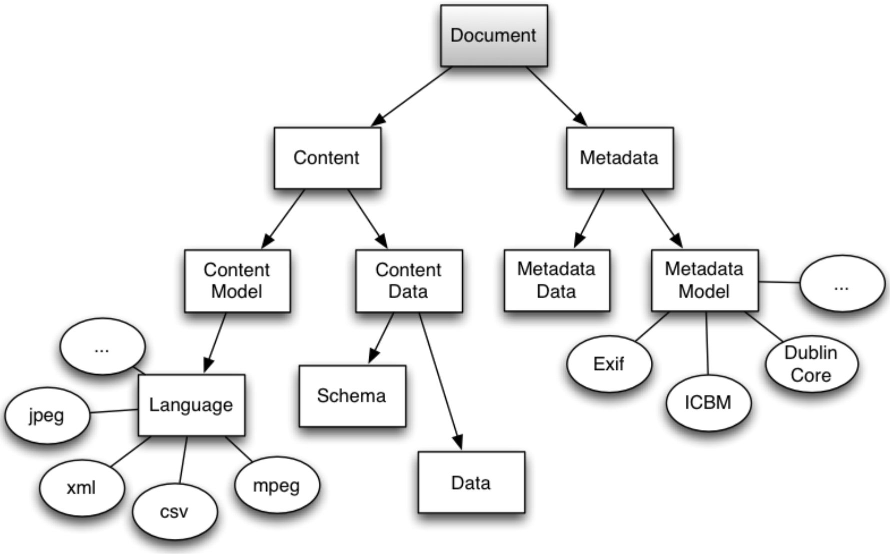


Figure 22 Document model

The repository is axed around a transversal annotation system. So, the documents are annotated. As represented in Figure 23, the document has two sets of annotations: the coverage and the content. Those annotations allow the matching and query system to index and retrieve the documents, for the users.

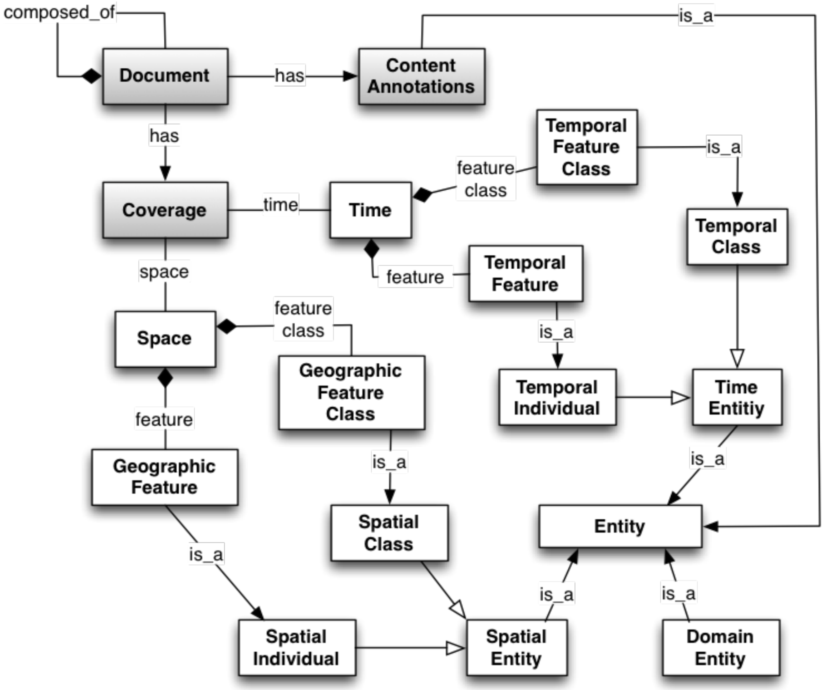


Figure 23 Document annotation model

First, the content annotation is a combination of entities from the annotation vocabulary described in the next section, 3.1.2. It can also be composed of spatial and temporal entities from the content of the object. But it is mainly composed of entities from the different domain ontologies. The coverage annotation is composed of temporal and spatial entities from the specific facets of the annotation vocabulary as described in 3.2.

Finally, in our model a document can be divided in sub-documents that have their own coverage, for example a chapter or a paragraph.

3.1.2 Annotation Vocabulary Model

The annotation vocabulary is imported and constructed by domain experts, and is used to annotate the repository resources. This annotation is used to index temporally, spatially and textually the documents. It also helps users to formulate the queries but most of all, helps in creating the links between the documents and the objects of the city model. As shown in Figure 24, the vocabulary model is composed of three facets: spatial, temporal and domain. Each of those facets are composed of different vocabulary aligned together such as ontologies, thesauri... An alignment is the process of determining semantic equivalence between concepts, properties or instances in vocabularies.

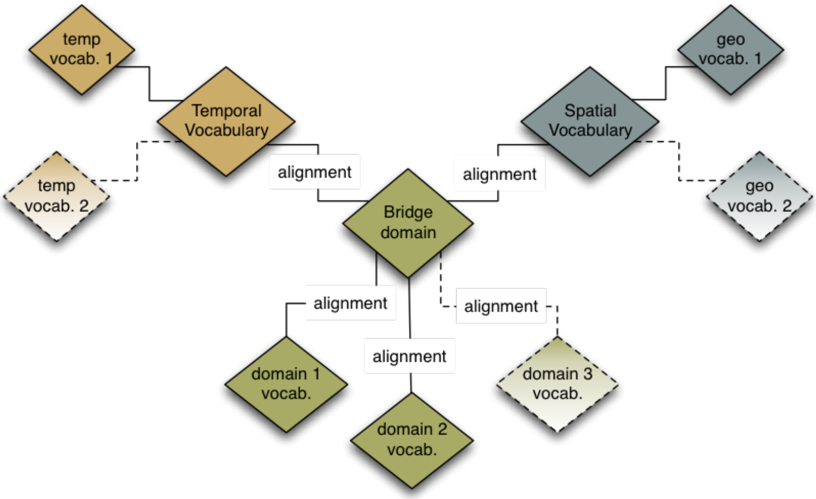


Figure 24 Annotation vocabulary facets

This creation of the annotation vocabulary is done through the following steps. First, the terminological resource to import, *e.g.* ontology or thesaurus, is selected and transformed in OWL format. The Web Ontology Language (OWL) [44], by facilitating greater machine interpretability as well as human readability, is a commonly used format to represent ontologies, to formalise thesauri, *etc.* The new OWL resource is transformed following the model represented in the upper part of Figure 25. Then, the resource and its entities are contextualised with a coverage. Finally, the entities of the resource are aligned with the ones already imported.

Each terminological resource is a domain ontology for a specific discipline and represents a specific point of view on this discipline. It provides the annotation vocabulary, with the terminological entities for this domain. So, the annotation vocabulary is created with the alignment of the different domain ontologies. This alignment consists in adding equivalence and subclass axioms that link classes in different ontologies.

This domain specific knowledge is also linked to the core of the annotation vocabulary, which is composed of a spatial and a temporal facet. These spatio-temporal entities are mainly used to define the coverage within the repository. The link between the domains and the spatiotemporal facets is done through a defined “bridge” ontology. This “bridge” fills the semantic gap between a concept and a city object or geographical name. It consists on a categorisation of the domains. It can typically be composed of the first levels of a generic thesaurus.

A terminological resource is composed of entities linked with each other that can be axioms, concepts, individuals or properties. The entities and the resource itself are tagged with a coverage, which forms the contextualised terminological resource and the contextualised entities.

The coverage, as described in 3.2.1, is composed of time and space elements. They describe the spatiotemporal context that represents the validity of the entity, or the resource. For example, the individual "Département" (defined as a French administrative division) will have as spatial coverage the entity "France" that represents the country it is used in, and as temporal coverage "1789-Nowadays", as it was created in 1789 and is still in use. A group of entities, that share the same coverage, is referenced as a new contextualised terminological resource that composes the original one.

The resulting annotation vocabulary is represented in the lower part of Figure 25.

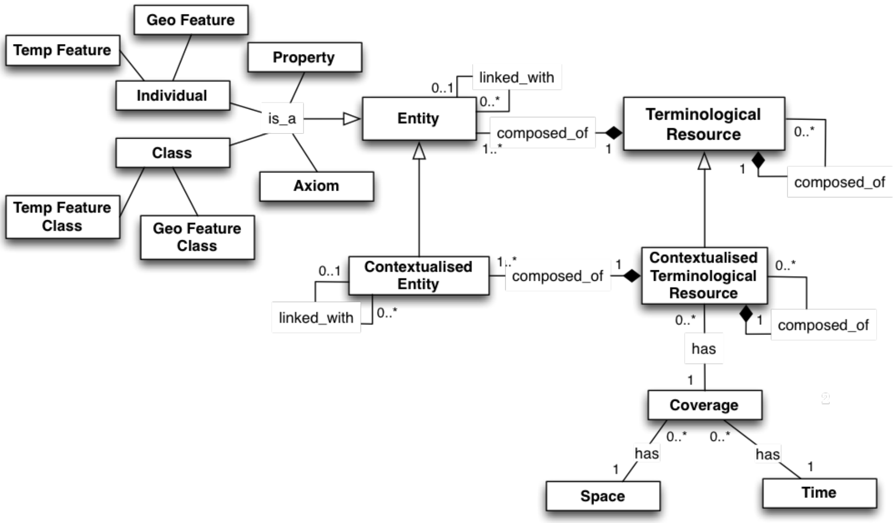


Figure 25 Annotation vocabulary model

3.1.3 Spatial Resource Model

The spatial resource is a 3D city model (3DCM), which represents the spatial boundaries of the repository.

Every document imported in the repository is tested, mainly its spatial coverage, with each city object contained in the 3DCM. The documents are then linked, when possible, to the set of city objects or geographic zones that represent their spatial coverage. The process is detailed in Chapter 4, section 4.3.

This link is also the key to the presentation of those documents. With the use of visualisation techniques, this 3DCM offers the visualisation of the spatial coverage and doing so increases the readability of the documents.

This spatial resource is also where the user navigates in order to query and browse the repository.

To integrate the 3DCM within the repository, we need to annotate it. The linking, querying and indexing are based on those annotations. The annotation process consists here in identifying the 3D city objects with the corresponding place names or concepts from a given vocabulary. This vocabulary is part of the spatial facet of the annotation vocabulary defined in 3.1.2. We propose an identification technique in Chapter 4, section 4.2. Thanks to this annotation, we are able to match the documents annotated with similar concepts and the 3D objects from the 3DCM.

3.2 The Coverage Model

3.2.1 Coverage Definition

We propose to use a definition of coverage inspired by the Dublin Core one, which states the coverage to be "the spatial applicability of the resource or the jurisdiction under which the resource is relevant". Accordingly, we have defined our notion of coverage as follows: *the geographic and temporal context of a resource refers to the spatial and temporal regions in which the resource is true or must be true in the real world or in a fiction.* We use here both meanings of "true": something that is correct, accurate, or something that is real, genuine. The coverage does not contain geographic and temporal entities that could appear in the content of the resource but doe not match the definition of coverage mentioned before. In this section, the resources represent both the documents and spatial resources held in the repository.

Contrary to the Dublin Core coverage, which defines the spatial coverage as a list of place names or coordinates, we propose to define it as a composition of entities such as place names together with functions or properties. To do this, we use semantic vocabularies and so allow axioms (logical expressions) as coverage. For example: `'feature code' some P.PPL and 'population' value 1000 and 'level 1 Administrative parent' value Rhône-Alpes` (all the cities (P.PPL) with a population of 1000 and within the Rhône-Alpes region of France). This use of semantics within the definition of coverage allows us to be more precise than a simple list of coordinates or place names. It also allows us to describe non-connected region as coverage. The spatial coverage is a composition of semantic classes or entities from a specific geo-spatial ontology and coordinates or geographic region as shown in Figure 29. The formal model of the coverage is given in section 3.2.3. For example: `'feature code' some S.SCH and 'level 1 Administrative parent' Geneva` (all the buildings with a function of school (S.SCH) within the Canton of Geneva in Switzerland (Admin

1)). The time coverage is composed of classes or entities from a specific time ontology. They can be of the form of a range of time and dates or a specific point in time. For example: “the Tuesday’s evenings”, “the 19th century”, “the 2$^{\text{nd}}$ of April 2010”... Moreover, we can apply a time property to each spatial concept that composes the spatial coverage. This allows defining the exact temporal range, when the document is associated to a particular spatial region. Both ontologies models are described in the annotation vocabulary model in 3.1.2.

In our definition, for the particular case of a resource that has sub-resources, the coverage of the parent resource must be greater or at least equal to the union of all the sub-resources coverage. No coverage of sub-resource can be greater than the one of the whole resource. For example a resource, which is composed of three sub-resources that have respectively for coverage `Paris`, `Bruges` and `Geneva`. The main resource coverage could be `Europe` (as it comprised the three cities) or `Geneva U Bruges U Paris`, whichever level of precision suits better the resource.

The coverage we present is a transverse concept that is applied to all of the repository resources. Each resource has a spatial and temporal region where its content is true or must be true. It is a key aspect to enable the coordination between the different repository resources. As a consequence, the spatial facet of the coverage allows us to retrieve complementary documents. Thanks to their spatial closeness, the query engine returns objects to the user outside of the requested domain, but spatially pertinent. Resources or documents, annotated with a particular spatial feature might be complementary. For example if a user requires documents on the norms applying to a particular building, it will not only cover the construction domain but also the security and the sanitary domain at least. One of the links between those documents is their coverage.

Examples of coverage identification and application for different type of repository resources are detailed in 3.2.2. The formal model of the coverage is given in 3.2.3.

We did not develop the time facet of the coverage further than its generic integration within our models. This choice was made as the time is out of scope of our research, which is centred on the introduction of the spatial coverage to a repository model. However, we researched its implementation and have developed some tracks as detailed in Chapter 4.

3.2.2 Spatial Coverage and Spatial Content

In the models and systems presented in the state of the art, the spatial context is generally obtained by extracting place names or coordinates from the document content. In the approach we study here, the coverage (*i.e.* context) is not usually simply inferred from the list of geographic features found in the resource content. As explained in 3.2.1, our spatial coverage is a combination of semantic and geographic notions that describes precisely the area where the resource is true. We propose to make a distinction between the geographic and spatial content of a resource and its actual spatial coverage. The goal is to differentiate the spatial content: a reference to something in a given place for example and the actual context of the resource.

This distinction allows us to be more precise. Comparing coverages and contents separately is more precise than comparing both annotations together, *i.e.* checking the entities indiscriminately. For example the following query “content: Italy pictures exhibition; coverage: Geneva” is more precise than “Italy pictures exhibition Geneva”, where there is no distinction of the coverage for the geographic terms in the query. In the first example it is clearer that the person searches for an exhibition in Geneva about Italian pictures. Mixing all the entities together looses the refining possibility. We can order more finely the results by focusing on the coverage for example, or on content if the distinction exists. It also enables us to be more accurate in the process of presenting related resources to the user based on their spatial closeness.

Typology

Each resource stored in the repository has a coverage. We present here different ways of refining this coverage for each type of resource. A given resource can have multiple interpretations and can be considered of multiple natures. The annotator will then choose to annotate the resource with a coverage corresponding to a combination of all the possible values.

Coverage in the annotation vocabulary

The vocabulary is an important part of the repository as it is key to the indexation of the resources held within the repository. Concepts in languages evolve over time and space. The contextualisation of each of their definitions allows precision and disambiguation.

The contextualisation of a semantic vocabulary or ontology is done through the annotation of its entities. The coverage is represented using annotation properties. Instead of tagging every entity contained in the ontology we propose to contextualise, with the coverage, the full resource and the major entities. The annotation is spread to the child entities of the ones tagged. The resource can then be decomposed in sub-resources according to the coverage of the entities. A group of entities with the same coverage is considered as a sub-resource.

The vocabulary is mainly used within the repository to annotate and query the stored resources. The examples below describe the use and necessity of the vocabulary's spatial and temporal coverage in such context.

The time coverage of a concept describes the time period when a given definition is true. This annotation is given to each definition if many exist, accentuating the concept evolution over time. For example, the geographic term "Canton de Genève", which represents one of the Swiss cantons, is true since December 31th 1815. Before that, it was known as the "Département du Léman" and was part of France between 1798 and 1813.

The spatial contextualisation of a concept can help solve ambiguity if its definition varies in different countries. For example the concept "*Canton*" represents different kinds of administrative division in France and in Switzerland. If we look at two queries such as Q$_1$(Content: ‘‘Canton’’, Coverage: ‘‘Europe (continent)’’) and Q$_2$(Content: ‘‘Canton’’, Coverage: ‘‘Geneva’’), we can see that the coverage of Q$_1$ is not precise enough to determine which definition of "*Canton*" the user uses. The entity "Europe" is parent to both "France" and "Switzerland". So it is impossible to guess which definition is to be used, as both are true in Europe. This leads to an ambiguity only solvable by more precision given by a user. Conversely, Q$_2$ is clearly more precise than Q$_1$, as it uses a lower level entity. "Geneva" is contained in "Switzerland" and Switzerland possesses a unique definition of "*Canton*". So we can easily deduce, that Q$_2$ is using the Swiss definition of "*Canton*". In this case, there is no ambiguity possible.

Finally, by specifying the coverage for some or all the terms inside a query's content, the user can express more accurate queries to the system. We can illustrate this by presenting an example using different coverage within the same query. The concept "*Collège*" in Geneva canton has the same meaning as "*Lycée*" in France. It is a school for students from 16 to 18 years old, just before University. But in France, "*Collège*" is the concept used for the school before the "*Lycée*". As an example, we have a student from Geneva who wishes to attend a "*Lycée*" in France and searches for resources concerning his transfer and the diploma (*Baccalauréat*) given by those schools. But the student only knows his own vocabulary, *e.g.* "*Collège*" and not the local term. The resulting query would be: `Q(Coverage:‘‘France’’, Content:‘‘Collège’’(Geneva), ‘‘Baccalauréat’’(France))`. The coverage of the query is "France", and each entity used in the content part of the query (*i.e.* keywords) has its own coverage to avoid ambiguity.

Coverage in the documents

Based on several thesauruses and indexation system, such as Wordnet [45], Schema.org[21] and the Dewey System [46], we have referenced and defined a number of formats and subjects that represent most of the documents stored in the repository. The resulting table illustrates this combination, using the union operator, of coverage definitions for each type of documents. The generic definition given in 3.2.1 is applicable to all types of objects. But certain cases need disambiguating and we propose here examples and precision on the coverage selection, for those.

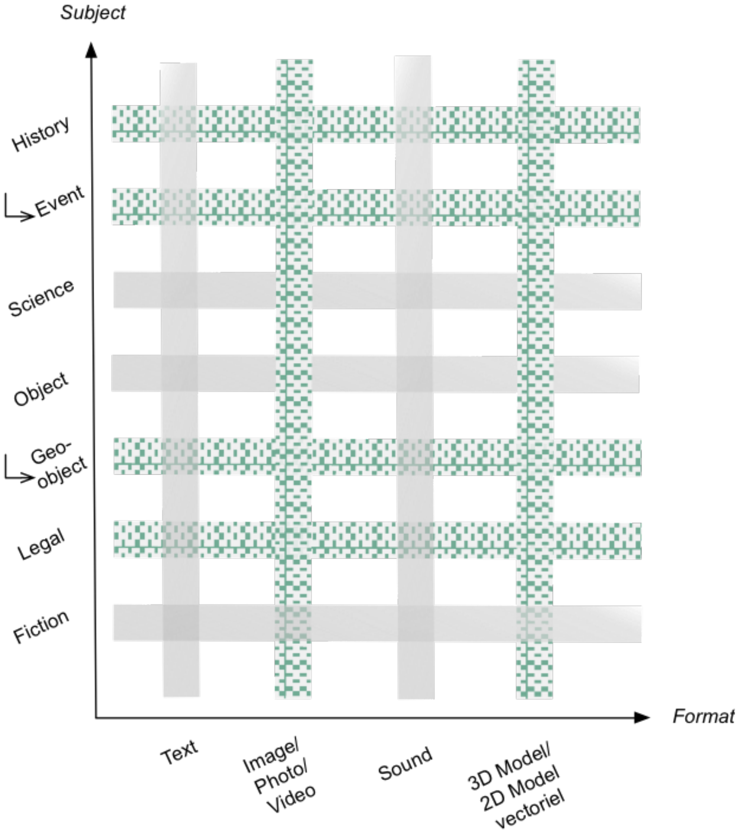


Figure 26 Resource's coverage table

The table, depicted in Figure 26, represents a selection of elements, whose characteristics match the most the implementation context presented in Chapter 4. It is composed of two main types of coverage: the green dotted lines, which

[21] https://schema.org

represent a specific coverage, and the grey lines that represent the “universe” coverage. It became obvious while building this table that some documents might have a coverage that can be too large (e.g. Universal) for it to be relevant to define it precisely. So, for those cases, we use the “universe” coverage, where the object is true everywhere at any time.

The selected coverage’s subjects (themes) and formats specifications are described below. We begin with the list of subjects.

History. We mean here the aggregate of past events, for example a document, which describes the life in the 14th century. The coverage of such resource is the geographic place it describes and the time of validity of its content.

Event. We mean here any action, process, or thing that happens or takes place. It is considered here as a sub-subject of History. For example, a newspaper article that comments on the meeting between the French president and the American president in Switzerland. In this case the coverage of the article is Switzerland as it is the place where the meeting was held. America and France are considered as part of spatial content. The coverage is the one of the event itself, where and when it took place; any other temporal, geographic or spatial references are associated to the content.

Science. We mean here the factual science such as facts or scientific generality. For example, a document presenting Pythagoras’s Theorem is true everywhere and not only inside a given space and time. Science is considered having “universe” coverage.

Object. We mean here the description of an object. For example, the user manual to change a car tire is basically true everywhere. Object is considered having “universe” coverage.

Geo-object. We mean here the description of a geographic object such as a particular building in a city. The spatial coverage of such document is the one of the object: its exact footprint and its time validity.

Legal. We mean here any document with a legal aspect. The spatial coverage of such resource is the jurisdiction of the resource.

Fiction. We mean here any fiction document that takes place outside the earth, in a fictional world or that is not linked to a special geographic place or time. Fiction is then considered having “universe” coverage.

The types of text used in the document can impact the coverage derived from the type of subject, described below. There exist five types of text as stated by [47]: narrative, descriptive, expository, argumentative, and prescriptive.

The **narrative** type generates a temporal aspect, as it is composed of a series of chronological events. So it implies the coverage of the document to at least be the union of all the coverage of the events it is composed of.

The **descriptive** and **expository** types, which aim at describing and explaining, both inherit the coverage of what they describe. They do not imply more precision for the coverage definition.

The **argumentative** type is the representation of a subjective judgment, of person thoughts. These judgment or thoughts can evolve in time and space. The location or the environment of a person influences her ideas. This type of text infers the spatial and temporal coverage of the argument that the document contains, *i.e.,* when and where the argument is a true argument.

The **prescriptive** (or **directive**) type represents rules, law, orders, or instructions, *etc.* It infers as coverage to the document, the jurisdiction of the content.

Finally, the precisions for the formats are described in the list below.

Text. The text format does not induce any time or geographic properties to a resource. So the coverage implied by the text format is “universe”.

Images, Photos and Videos. Those visual formats imply as coverage the subject in the frame, and not the position of the photographer. More precisely, the coverage is what spatial region we see in the resource. But an image picturing something that does not exist, like a drawing of a triangle, does not imply a spatial coverage. The time is the moment the film or photo was taken. Persons and objects in the frame are considered as content of the resource.

Sound. The sound format does not induce any time or geographic properties to a resource. So the coverage implied by the sound format is “universe”.

2D and 3D Vector Model. We mean here any resource 2D or 3D that depicts a vector model of data. For example, a wind simulation, the postman path within the city... The spatial coverage is the footprint of the data held and express by the model. The time coverage is the time where the data is true (valid).

To illustrate the use of the matrix, we choose to present two examples of complex documents.

First, the photo of Jeff Widener, see Figure 27, at Chang'an Avenue, on June 5th, 1989. The day after China's government began violently repressing the protestors in Beijing's Tiananmen Square, a lone man decided to take a non-violent action against an approaching line of military tanks.

The event illustrated by this photo is the Tiananmen protest of 1989. So if we follow the rules defined before, the event implies the spatial coverage to be the Tiananmen Square. The fact that the document is a photo implies that the spatial coverage is the spatial region we see in the document, here Chang'an Avenue. We must combine the two precisions in order to generate the spatial coverage.

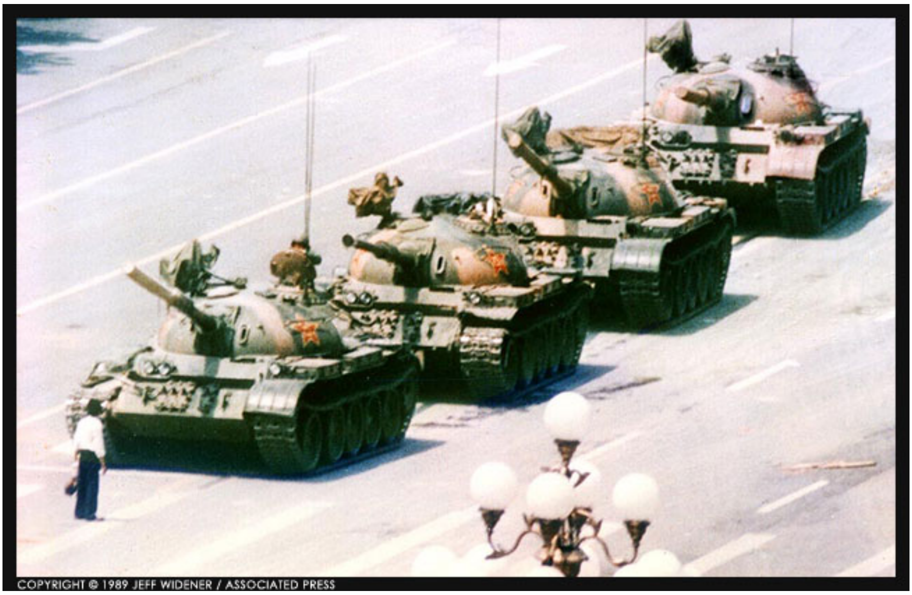


Figure 27 "Tank Man" near Tiananmen Square by, Jeff Widener 1989

The picture must appear in the results for both Chang'an Avenue and the Tiananmenn Square. To do so, the coverage of the document is the following axiom, based on the union operator: `Chang'an Avenue U Tiannanmen Square (June 5th, 1989 - Today)` and the content of the document: `Tiannanmen Square (April 15th, 1989 - June 5th, 1989).`

The second example is the picture, Figure 28, of the signature of the Evian Agreements in March 18th, 1962. It shows the Algerian delegation in front of the hotel before the signature. The agreements, led to an instant cease fire in the war between France and Algeria (1954-1962) and declare the independence of the Algerian country.

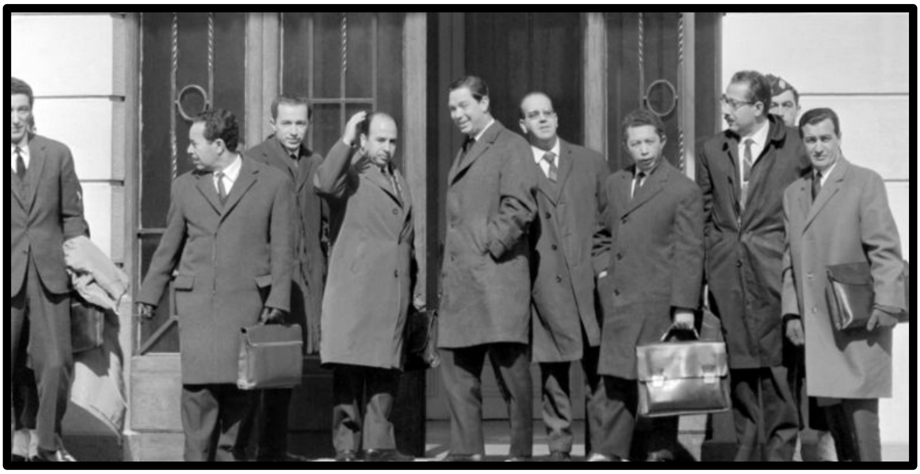


Figure 28 Algerian delegation arrivals to the hotel, for the signature of the "Accords d'Evian", © AFP/STF 1962

The meeting for the signature took place in the "Hotel du Parc" in Evian-les-Bains, France. Accordingly, this picture should appear in results for both this particular hotel in Evian, as it is what is shown in the picture, and Algeria, as the war took place there. We can define the content of the document as the time and place where the signature took place combined with the time and space of the war.

To do so, the coverage of the document is the following axiom: `Hotel du parc (March 18th, 1962 - Today)` and the content of the document: `France U Algeria (1954- 1962).`

Coverage in the spatial resource

The spatial coverage for a spatial resource is simply what is represented by the resource itself. Its temporal coverage is the date or range of time the resource represents. The coverage of the 3D city model of Geneva city in the 1950's is simply the "Geneva city" concept with the temporal coverage of "1950's".

3.2.3 Coverage Model

As described in 3.2.1, the coverage is composed of a temporal and a spatial facet. The coverage references spatial and temporal entities that represent where and when the resource is true. It differentiates them from the geographic and temporal mentions that appear in the content of the resource but do not reflect the coverage. The resource represents here any document or spatial resource held in the repository.

Our model for the spatial coverage is composed either as a set of (named) spatial features, such as canton names or cities names, or as (a set of) class of geographic entities (*e.g.* the buildings, the airports, *etc.*), or both. After adding the temporal dimension we obtain a coverage model made of three components as shown in Figure 29:

1. Coverage Features (`CF`): a set of geographic features (place names, polygon and coordinates...)
2. Coverage Classes (`CC`): a set of geographic feature class definitions (in the current implementation class definitions are expressed in description logics, *e.g.* `‘feature class’ some S and (‘feature code’ value SCH)`)
3. Coverage Time (`CT`): a set of time intervals or time entities

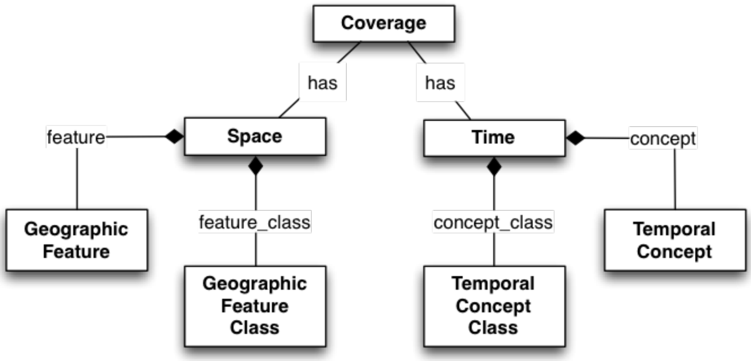


Figure 29 Coverage model

The entities that compose the different parts of the coverage are issued from the geographic and spatial facet of the annotation ontology described in section 3.1.2.

3.3 Method

In this section we describe the methodology needed to implement our model described in the previous sections.

In order to build this model, we first need to build the annotation vocabulary. As detailed in section 3.1.2, the base of the annotation vocabulary is composed of three facet vocabularies. The time vocabulary must be able to represent all the time elements used to describe the documents stored in the repository. The spatial vocabulary must also be able to describe all the geographic and spatial elements used to describe the documents stored in the repository, but also be able to annotate the spatial resources that will be integrated in the repository. We propose to build the spatial vocabulary as a composition of a geographic vocabulary and a spatial vocabulary. The spatial vocabulary can also be enhanced with the implementation of gazetteers for the completion of the dataset spatial. Finally, the bridge domain vocabulary must be generic enough to be able to be aligned with any domain vocabulary. A high level thesaurus or ontology is a good candidate for this facet.

Once the annotation vocabulary has been assembled, we need to annotate the documents and define their coverage using entities and classes from the vocabulary. The document's annotations describe the content of the document and can be issued from any of the facet. The coverage of the documents must be a composition of classes and/or instances form the time and spatial facets.

Finally, we need to annotate the spatial resources where the documents will be linked. Those annotations are identifications of the geographic and spatial elements of the spatial resources using exclusively the spatial facet of the annotation vocabulary. The identification process associates with each spatial object from the spatial resources the corresponding geographic instance from the annotation vocabulary. The identification is based on the characteristics of the object as given by the spatial resource.

3.4 Query and Matching

This last section describes the query and matching models that allow the user to access the resource held by the repository and interact with it.

In the remainder of the thesis, starting from this section, we only consider the spatial facet of the coverage. The time facet can be implemented following similar rules.

3.4.1 Query Model

The repository users use the query system to retrieve documents in order to achieve a given task.

As presented in Figure 30, a query is composed of a coverage and a content. The coverage of a query represents the location the user is focusing on for his/her research. It contains a combination of spatial and temporal entities, as described in section 3.2.1. For example the query coverage could be all the building with the function "*hospital*" (`CC`) in the *state of Geneva* (`CF`). The content of the query is a list of terms or entities that describes the interest of the research, from different ontologies of domain. Those entities can be annotated with a coverage to precise the meaning used for the query. For example, for the disambiguation of a term that has different meanings in different countries or times.

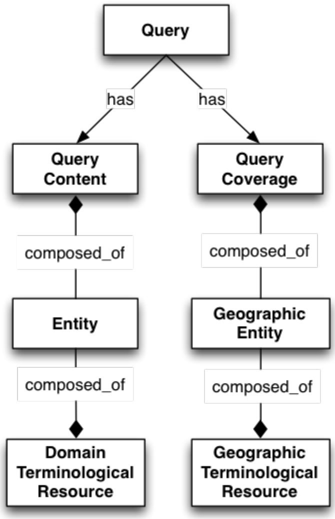


Figure 30 Query model

A query example for a query on brain surgery with the coverage being all the hospitals in Geneva, would be as follows:

`Q₁(Content: ‘‘brain surgery’’; Coverage: {‘feature code’ some S.HSP and ‘level 1 Administrative parent’ value Geneva})`

3.4.2 Matching Algorithm

Now that we have modelled the repository entities and the queries, we can focus on the matching as part of the information retrieval process. The first step is to filter the documents that have a spatial coverage that matches the query's documents. There is no need to return documents that matches the query terms if their geographic footprints do not match.

After the filter process we rank the selected documents according to their matching result with the query's coverage and keywords. We primarily focus on the spatial facet. The most pertinent documents are the ones were both coverage are identical. Depending on the context of the query, we might prefer documents with a more generic coverage (inclusion case). For example, if the query is focused on finding information located in the city of Geneva, a document with a coverage set to Budapest (disjoint case) might be less significant than a document set in Switzerland or Europe (inclusion case). The coverage being composed of entities from ontologies, as described in section 3.2; we use semantic similarity measures between, for example, two features or classes in the ontology, as presented for example in [48]-[51]. Doing so allows us to rank the resources according to their geographical closeness with the query's coverage. The closest the entities, the more relevant to the researched location the document would be.

As for the content or keywords matching, we can use the techniques on term indexing and ranking. We first need to match and rank the annotations of the document content, composed of entities from the annotation ontology, with the query's content. We apply, for this step, semantic similarity methods as described for example in [50], [51]. A second step, in order to deal with possible ambiguities and to refine the ranking, is to use traditional text indexing methods like TF*IDF [52], that parses the document content to retrieve keywords.

Both matching processes should return a score that represents the pertinence of the document with the query for each aspect, *i.e.* content and coverage. To compute those scores we can apply methods from the literature. In the domain of geographic information retrieval many systems have been prototyped as described for example in [53], [54] and in the SPIRIT project [12]. They have applied indexing and matching methods such as R-trees, grid based indexing, inverted indexes and probabilistic IR. As for the semantic aspect, there exists different methods to compute semantic similarity [50], [55] and more specifically semantic similarity applied to geographic information retrieval as listed in [48], [56].

In order to identify the pertinence of each document according to a given query, we represent the matching as shown below. The matching focuses on two aspects: the coverage that is composed of coverage feature (CF), coverage classes (CC) and the content (Cont).

For a query $Q=(Q_{CF},Q_{CC},Q_{Cont})$ and a document $D=(D_{CF},D_{CC},D_{Cont})$ the matching score is defined as:

$$M(Q,D)=a_1M_{CF}(Q_{CF},D_{CF}) + a_2M_{CC}(Q_{CC},D_{CC}) + a_3M_{Cont}(Q_{Cont},D_{Cont})$$

Where the a_i are the relative weights of the matching scores and M_{CF} is a matching function based on the geometry of the feature lists to match [12]. By tuning the weights, we can give more or less impact to each part that composes the matching algorithm: the content, the coverage classes or the coverage feature. In our work, the geographic coverage should have a greater impact for the matching than the content. a_1 and a_2 should be greater than a_3. M_{CC} is a matching function for lists of classes. It is based on a notion of semantic similarity between classes. M_{Cont} is a classical matching function for indexed documents, such as TF*IDF.

The closeness of both coverage features (CFs) needs to be verified before calculating the pertinence of the coverage classes (CCs), as they describe types of geographic elements and so help precise the spatial coverage.

In addition, if $a_1M_{CF}(Q_{CF},D_{CF}) + a_2M_{CC}(Q_{CC},D_{CC})=0$, then $M(Q,D)$ is set to 0 to represent the fact that documents that have no coverage intersection with the query must not appear in the result.

The final relevance value is a composition of the two matching scores (content and coverage). This value allows us to filter and rank all the possible documents, and to return them ordered according to their pertinence.

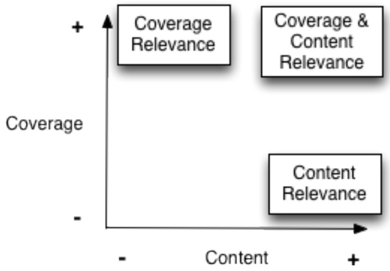


Figure 31 Document relevance scale

3.5 Summary

In this chapter, we have presented the full repository model. Each part composing this repository has been detailed and modelled. First we have defined and modelled the three repository's resources: The annotation vocabulary creation and composition; the documents model and annotation model; finally, the spatial resource usage and annotation. Next we have presented our notion of coverage. Its application to the rest of the repository was also detailed and illustrated with examples.

In the last part of this chapter we presented the model of the query and matching system used with the previously described resources. We also proposed famous techniques from the literature to use for the matching algorithm.

The following chapters focus on use cases and the implementation of this proposed model.

Chapter 4

Use Cases for Associating Documents and City Objects

In this chapter, we will detail applications of our model described in Chapter 3. The first applications are the implementation of models, with the creation of the annotation model, and more practically the annotation of spatial resources and the spatial annotation of documents. We also present the existing sources to retrieve and process geocoverage for different document types. Finally, we will highlight in this section 4.5 the relevancy of the notion of coverage with our use case.

4.1 Building the Annotation Model

4.1.1 Method implementation

In this section we describe the implementation of our model methodology described in 3.3.

The first step of the implementation is to build the annotation vocabulary. This vocabulary is composed of three facets: time, spatial and domain. The time facet can be built using the Owl-Time ontology [63] from W3C. The spatial facet is built using a combination of geographic and spatial ontologies and thesaurus. We have combined the Geonames [60] and CityGML [61] ontologies to create the spatial facet. Finally, the domain facet is built using the Urbamet Thesaurus[22]. All the previously mentioned vocabularies are described in the section 4.1.2.

The documents are then annotated using the annotation vocabulary. The documents annotations are composed of entities and classes from the annotation vocabulary. The definition of a document coverage is composed of classes and entities from the time and spatial facet of the vocabulary.

The final step is the annotation of the spatial resources. Those identify the geographic and spatial element that composes the spatial resource. In this identification task, only instances and classes from the spatial facet of the annotation vocabulary are used. The identification is based on the characteristics of the object as given by the spatial resource. The process is described in section 4.2.

[22] http://www.urbamet.com/thesaurus-urbamet-r13.html

4.1.2 Building the Core: Implementation Choices

As explained before, the core of the annotation ontology is composed of three aligned ontologies: a geo-spatial ontology for the space facet, a time ontology and a domain bridge ontology. Our choices are detailed below.

Space. The spatial facet of the annotation ontology is composed of a geographic ontology that describes and holds all the geographic named entities and a spatial ontology to describe the 3D support. The alignment is constructed around the function and the address (coordinates, footprint) of the objects and features they describe.

For the geographic aspect, we have chosen the Geonames ontology[23] [60]. It provides instances called `feature codes` that differentiate the geographic features such as the villages, the cities, the parks, the mountains, *etc.* The individuals of this ontology represent place names and are annotated with additional information as for example the population number, postal codes, links to Wikipedia articles related to the place, feature codes and more. Feature codes describe the function of the entity (*e.g.* School, River, Park). We have transformed this ontology so that each feature code yields a class (*e.g.* Monument) whose instances are the geonames places with that code (*e.g.*, the Eiffel tower, Puerta del Sol). Each geoname has a RDF file attached, which describes its parents and its geographical hierarchy (*e.g.* parent country, parent administrative division).

That information can also be used in a logic expression to describe the spatial coverage of an entity. An example of such spatial coverage could be a logic expression that returns all the places categorised under a certain level of administrative division, or all the cities that have more than a certain number of citizens. This example could be described as follow with the Geonames ontology: `('feature code' value P.PPL) and population min 10000`, which describes every city (`P.PPL` feature code) with a population of minimum 10000.

For the spatial aspect we have chosen to use the CityGML ontology[24] [61]. It describes 3D city models and landscapes, composed of city objects, with respect to their geometry, topology, semantics and appearance. We have created this ontology using the XML schemas provided in the specification [62]. The top classes represent the different types of city objects (*e.g.* Land use, Building, City furniture). Each type has functions, which are described as subclasses. For example, the class `LandUseFunctionType` contains subclasses such as *Forest*, *Sea*, and *Airfield*. A city object has also a footprint that describes the polygon that marks the object sitting on the ground. For a building it is the

[23] www.geonames.org/ontology/

[24] http://www.citygml.org

`GroundSurface` class that contains those data. In the specifications, the address of an element can be composed of the postal address of the object, described in xAL format and the exact point of entry to the object.

Time. The Owl-Time ontology [63], as described by the W3C, is our choice as ontology for the time facet. It allows us to create all needed individuals and is compliant with the standards. It models temporal properties and content, date-time information, time zones and also scheduling of events. The class `Interval` in the ontology is used to describe the time validity of an entity. As an example, we could look at the axiom that states that mercury is toxic. This axiom has a time validity starting on the date when the sanitary administration stated it and is still valid today.

Domain. Any general thesaurus can be used to describe the list of domains, as for example, WordNet Domains [45], Gemet[25], *etc.* It needs to include all the disciplines of the multidisciplinary domains the library is centred on. As our use case is on urbanism, we focused on this multidisciplinary domain and so choose to use the two first levels of the Urbamet thesaurus[26]. In fact, although this resource is covering a wide range of domains it is centred on the urban aspect. If we look at the entity "*vehicles*", for example, in a resource we are importing, we could annotate it using the "*transportation*" term of the Urbamet thesaurus to describe its domain of application. In the same ontology, we could also have entities that apply to the "economy" term as their application domain.

4.1.3 Aligning the Ontologies

The alignment of Geonames and CityGML creates an ontology of geographic urban entities. The first one provides identification, information on the hierarchy between geographic features and their affiliation to the hierarchy of administrative divisions in a country. The later describes their geometry, their localisation, and defines city furniture such as benches or public lights. Finally, both give details on their function and this common attribute is the articulation of the alignment. For example, as described in Figure 32, the Geonames feature code `University (S.UNIV)` is aligned with the CityGML type `College_or_University`, which is a subclass of `BuildingFunctionType`.

[25] http://www.eionet.europa.eu/gemet

[26] http://www.urbamet.com/thesaurus-urbamet-r13.html

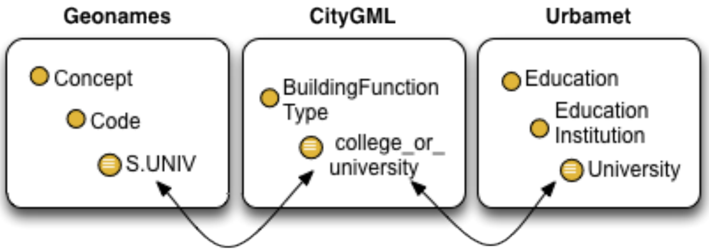


Figure 32 Ontologies alignment example

We had to modify the official Geonames ontology for the need of the alignment. The features class and features code, that represent the semantic of places, in the original Geonames ontology are instances. However in CityGML those codes are defined as classes. So we adapted the Geonames ontology by transforming the instances in subclasses, of the corresponding parent class *Code* or *Class*, and defining the place names as instances of those classes. So each geoname entity that has a geonameID is an instance of our new subclasses in our edited geonames ontology.

To complete the annotation vocabulary, as described before, the Urbamet thesaurus has been used as the "bridge" between domain ontologies and the spatial vocabulary. For example, the concept "*surgery"* will be linked to the city objects that have for function `Hospital`. In Figure 32, the Urbamet concept "University" is linked with the Geonames feature code `UNIV` and the CityGML type class `College_or_University`. So it is aligned with the geographic and geometric ontologies previously described. Then all the other domain's ontologies that describe the different concepts, used in a particular expertise, are linked with this "bridge" ontology, and through it, to the geographic ontology.

The aligned ontology is based on a main owl file called "GeoAnnotation" that contains all the alignments between the three ontologies we have chosen to form our annotation vocabulary: CityGML, Geonames and Urbamet.

- The CityGML owl file contains all the properties and main classes of CityGML, such as `FloorSurface`, `BuildingPart` and imports the CityGMLcodes[27] owl files that contain the attributes which are used to classify object such as roof type, city furniture's *etc.*

[27] http://www.citygmlwiki.org/index.php?title=CityGML_Code_Lists

- The Geoname owl file as described before has been tailored to fit the alignment with the other ontologies. The instances from the original ontology have been transformed into subclasses. The geoname owl file imports the instances files that contain all the named individuals such as countries, cities, point of interest, forests, water bodies *etc*. We created the instances owl files from the database dumps[28] available from the Geonames website. We transformed the exported file using Open Refine[29] tool in order to create the instances owl file.

The final ontology contains 6'239 classes, 251 object properties and 39 data properties.

The ontologies and their alignment are available in the following git repository: https://gitlab.com/CamilleTrd/GeoAnnotation-ontology.

4.2 Annotating the Spatial Resource

4.2.1 Identification of Spatial Objects

To increase the information on our 3D objects and to identify the place names, we have aligned the Geonames and CityGML ontologies a described in 4.1.3. Thanks to this work we can then identify a 3D object with a geoname individual.

The alignment process is based on the following heuristics:

1. The place position (longitude, latitude) must be contained inside the 3D object footprint;
2. The place function must match the object's function. This can be tested because Geonames feature codes and CityGML `functionType` classes have been aligned;
3. Rules to solve ambiguities, when two or more 3DCM objects satisfy

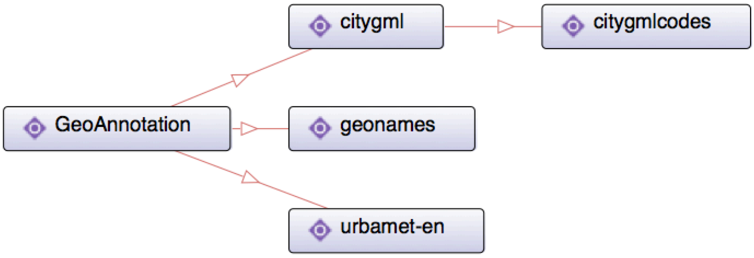


Figure 33 Onotlogies alignement schema

[28] http://download.geonames.org/export/dump/

[29] http://openrefine.org

conditions 1 and 2.

In Figure 34, we see how we identify two CityGML entities (`City1` and `City2`) with two geonames individuals (`G1` and `G2`).

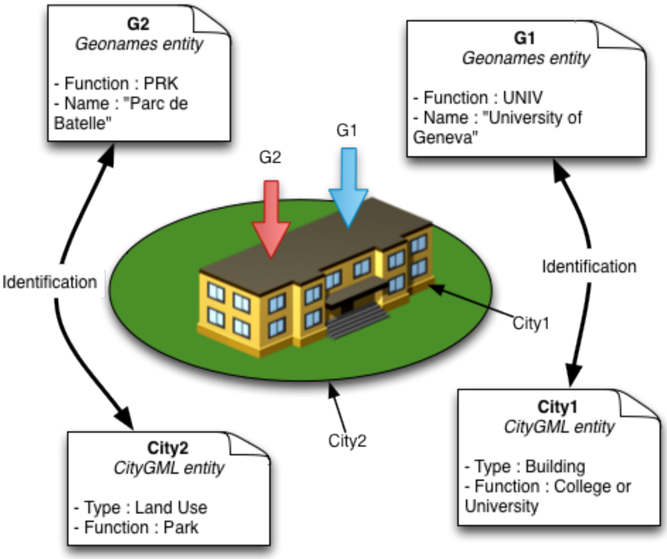


Figure 34 Example of identification

But not all of the urban entities are linked to a distinct building or area. Some can share the same building, in different storeys or different parts. Those entities will most probably not be related to the same resources, as their domain of application, or their function will differ.

As an example to motivate the need to identify sub-objects within the 3D models we present the case of the airport of Geneva. This airport has the singularity to hold within the same building, two airports, from two countries: France and Switzerland. This means that one half is under the French legislation and will be associated with resources regarding French airports and shops. As for the other one, it is under the Swiss law and will be attached to documents regarding Swiss restaurant shops, airline traffic, *etc.*

We describe below the techniques for the identification of sub-objects: storeys and groups.

4.2.2 Sub-Object Identification

In order to identify parts of a building with geonames, as for the identification of building, the geonames must be contained in the footprint of the objects. In the case of storeys, their elevation must match the one of the storey the sub-object is representing. A building with three storeys might have four geonames inside its footprint, three of them describing each floor and the last one describing the whole building. The storeys and sub-objects in geonames are described as contained in the building geoname using the `part_of` or `contains` attributes of the geonames. In CityGML, the storeys and groups are represented using the `CityObjectGroup` class with the following attributes: class having for value "`building separation`". In the case of storeys, function has for value "`lodXStorey`" where X is the level of detail number and `gml:name` is the name or the number of the floor. The group is then composed of rooms, doors, *etc.*, and is associated with a `BuildingPart` or a `Building`.

If the 3D object is known to be a group, if the coverage of the object contains the geonames `e` and `f`, and if `f` is known to be part of `e` (*e.g.* if `e` is a building, `f` is a building-part of `e`), then there are two possibilities:

- The group is defined in the model and then we identify the different objects individually.
- The sub or top-objects are not defined and we have to assign to the same object different geonames.

We have created a class `Storey` in the CityGML ontology to represent those rules.

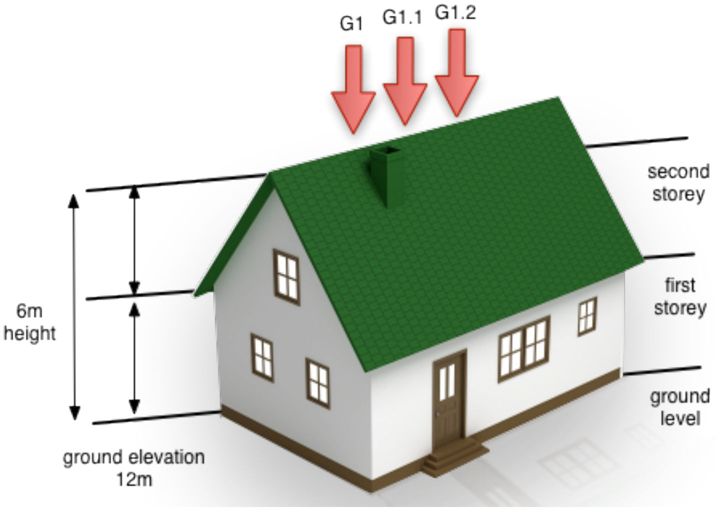


Figure 35 Example of storey identification

The next step is to identify each storey with the proper geoname. Each building has four attributes: `storeyAboveGround` and `storeyBelowGround` that contain the number of floors below or above the ground, and `storeyHeightsAboveGround` and `storeyHeightsBelowGround` that contain the total height of all the floors below or above the ground. In order to know the average height of each storey we need to divide the total height of the storey or above the ground with the corresponding number of floors. The ground elevation (`g`) added to the storey average height multiplied by its number (`n`), will give the appropriate elevation (e) of the geoname corresponding to the storey in question, above the ground. To calculate the underground elevation, g needs to be subtracted to the storey average height multiplied by -n, n being a negative number. We imply here that the ground floor is the storey zero; that the first floor is the first storey, *etc.* For the storeys above the ground, n has to be greater or equal to 0, and for the ones below, n is strictly less than 0.

$$g + \frac{storeyHeightsAboveGround}{storeyAboveGround} \times n = e$$

$$g - \frac{storeyHeightsBelowGround}{storeyBelowGround} \times (-n) = e$$

When the CityGML object does not contain the different storeys in its description, we link the geonames that describe the floors with the building itself.

4.3 Resource Annotation Techniques

There exist metadata models for different types of documents. If the metadata model contains a geographic property, it can be used to help define the geographic coverage annotation for a given resource. We present here some models that can be used to define a resource's coverage.

The most versatile and popular one is Dublin Core (DC) metadata model. It defines a "coverage" property[30] that contains the spatial or temporal topic of the resource. As described in Chapter 3, section 3.1.1, the "coverage" property is composed of a space and a time attribute. It is defined as: "The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant". The spatial attribute is of range *dc:location*[31] which is defined as a spatial region or named place. DC can be used for text and web resources, SVG (Scalable Vector Graphics)[32]... The Adobe XMP metadata model is based on DC and is used for videos, images, pdfs...

The EXIF metadata [64] is used for pictures and videos. It is composed of a GPS attribute and more precisely "exif:GPSLatitude" and "exif:GPSLongitude" tags. Both tags represent the coordinates where the camera was when taking the picture. It is expressed in degrees, minutes and seconds.

[30] http://purl.org/dc/terms/coverage

[31] http://purl.org/dc/terms/Location

[32] http://www.w3.org/TR/SVG/metadata.html

The IPTC photo metadata standard [65] , replaces the Information Interchange Model (IIM) from the International Press Telecommunications Council (IPTC) since 2007. It is composed of a core and an extension model. The extension introduces a clear distinction between where an image has been taken "Location Created" and a location shown in the image "Location Shown in the Image" (or Location shown). The metadata is of type "Location Details". It includes a Sub-location, a City, a Province or State, a Country (Name and ISO-Code) and a World Region.

IPTC also developed a series of news related metadata models like the rNews, NewsML-G2. They propose the "located" attribute that represents where the news item originates from and not especially where the news is about.

For HTML pages, there exist the "ICBM" meta and the Geo Tag format, with the property "geo.position", that take coordinates as value. Geo Tag also proposes a "geo.placename" and "geo.region" tags that receive text inputs. The geo tag format is described in [66].

Finally, ISO defined a standard [33] (ISO 19115: 2003(E)) for geographic information that can be used to describe digital or physical objects or dataset that have a spatial dimension.

4.4 Locate Resources

4.4.1 Align Documents and Spatial Objects

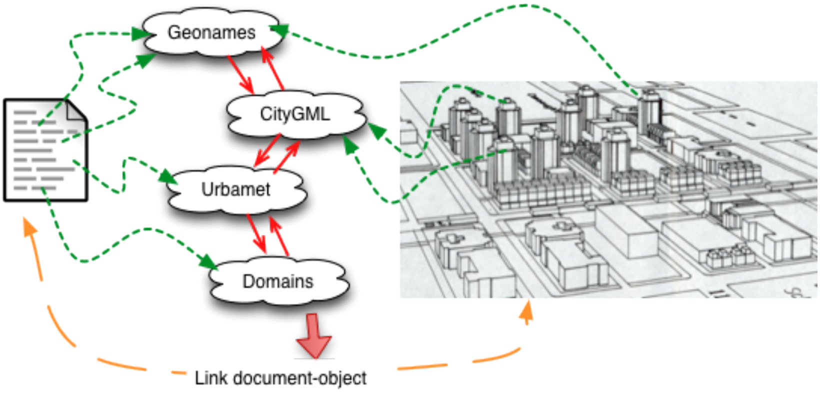


Figure 36 Link document-3D city model

The following algorithm, as an implementation example of our model, creates links between 3D objects and documents. The links can be of two types: *explicit* or *suggestion*. The first one describes a direct or indirect connection. If a document and a 3D object are indexed with the same entity (or with aligned

[33] http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

entities), they are directly connected, but if they are indexed with different ones, there is an indirect link if there is an ontological connection between those entities. The second one can be used in two cases. First, it can represent a connection made using a non-geometric or non-geographic vocabulary. Finally, it can represent a link with an object that is not contained in the current 3D city models (3DCM). However, the document is considered as relevant for the user and needs to be cited in the results. So a link will be created with an object that is contained inside the 3DCM and that is close to the location of the appropriate one.

So the algorithm takes as input a document `doc`, a 3D city object `obj` from a 3DCM, both indexed with the indexing vocabulary. It will return as an output, a Boolean value for the `explicit_link` and a number value for the `suggested_link` that represent the weight of the suggestion link. This algorithm is using classes and individuals from ontologies. `Cov(doc|obj)` is the coverage of the document or the object. For the clarity of this explanation we are only treating the spatial aspect of the coverage in this algorithm, but a similar process has to be apply to the time facet. `Cov()` can only contain individuals or classes from the Geonames or the CityGML ontology. `Idx(doc)` is the list of concepts attached to a document to define its content. It can contain individuals or classes from any of the vocabulary's ontologies. `Idx(obj)` is the list of instances of CityGML classes that represent the object.

First of all the algorithm needs to check that the coverage of the document matches the coverage of the object, otherwise no link can be created. So `cov(obj)` needs to be included in or equal to `cov(doc)`, or vice versa. If this assertion is true, we can search for a link via the `searchLink(doc,obj)` function described below.

There are four possible cases where a document can be linked with an object. Cases 1 and 2 (1^{st} and 2^{nd} point below), represent the cases where the entities contained in both indexing list, are identical or hierarchically linked. If the indexing terms for the document are linked to the function of an object, as explained in the case 3 (3^{rd} point below). The 4^{th} point describes the three first cases where the object is a group. Finally, as described in the 5^{th} point below, the 4^{th} case represents the case where the object is not present in the model but the document is relevant and needs to be returned to the user.

1. Direct link

If `cov(doc)` and `cov(obj)` contain the same entity, then a link can be created between `doc` and `obj`, so `explicit_link=true`.

Example. If `cov(doc)` and `cov(obj)` both contain the "Geneva airport" geoname, so we can create an explicit link between the two entities.

2. Direct link by class

If `cov(doc)` contains a class `c`, and `cov(obj)` an individual `e`. If `e` is an individual of the class `c`, then a link can be created between `doc` and `obj`, so `explicit_link=true`.

Example. If `cov(doc)` contains the "*airport*" class from CityGML and `cov(obj)` the "*Geneva airport*" geoname. This individual is linked with the "*airport*" class, so we can create an explicit link between the two entities.

3. Link through non-geographic entities

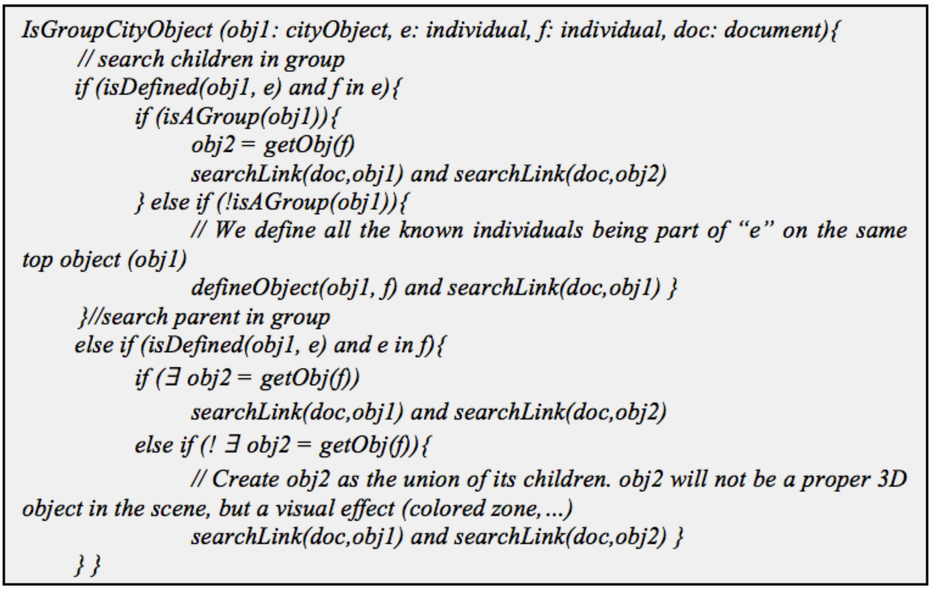
`g` is an individual or a class, from one of the domain ontologies (*i.e.*, not from Geonames or CityGML). If `idx(doc)` contains `g`, `cov(obj)` contains `f`, and there exist an alignment in the ontologies between `f` and `g`, then `suggested_link = 1/semantic_distance(g,f)`. The semantic distance can be computed according to one of the well-known concept distance (or similarity) measures that have been defined for ontologies.

Example. If `cov(doc)` contains the "*surgery*" concept and `cov(obj)` the "*Geneva university hospital (HUG)*" geoname. This concept is linked with the *hospital* class, and "*HUG*" is a hospital. We can then create a suggestion link with the value of the semantic distance between the concept "*surgery*" and the geoname "*HUG*".

4. The object is a group

The three previous cases can be in a particular situation if the object is known to be a group.

In this case for the identification of group objects, the algorithm has to test if the individual `e` is defined as a group in the 3DCM. The test is done by calling `isGroupCityObject(obj,e,f,doc)` as described in Figure 37. This function will then call `searchLink(doc,obj)` with the new identification and fall back on the cases 1, 2 and 3. In order to know if an object is a group of city objects, it calls the `isAGroup(obj)` Boolean function. To verify if an object has been defined with a geoname entity, and to retrieve an object from a geoname that identifies it, it respectively uses `isDefined(obj,e)` and `getObj(e)`. To assign to an object a geoname entity it calls `defineObject(obj,e)`. This function will associate a specific entity to a city object.



```
IsGroupCityObject (obj1: cityObject, e: individual, f: individual, doc: document){
        // search children in group
        if (isDefined(obj1, e) and f in e){
                if (isAGroup(obj1)){
                        obj2 = getObj(f)
                        searchLink(doc,obj1) and searchLink(doc,obj2)
                } else if (!isAGroup(obj1)){
                        // We define all the known individuals being part of "e" on the same
top object (obj1)
                        defineObject(obj1, f) and searchLink(doc,obj1) }
        }//search parent in group
        else if (isDefined(obj1, e) and e in f){
                if (∃ obj2 = getObj(f))
                        searchLink(doc,obj1) and searchLink(doc,obj2)
                else if (! ∃ obj2 = getObj(f)){
                        // Create obj2 as the union of its children. obj2 will not be a proper 3D
object in the scene, but a visual effect (colored zone, ...)
                        searchLink(doc,obj1) and searchLink(doc,obj2) }
        } }
```

Figure 37 IsGroupObject algorithm

Example. If `cov(doc)` and `cov(obj)` both contain the geonames $\mathtt{g}_1$ that represents a building that hosts many different companies, and $\mathtt{g}_2$ that represents a part of this building owned by one of the enterprises. `obj` is defined by $\mathtt{g}_1$. The algorithm calls the function `isGroupCityObject`. We know that $\mathtt{g}_1$ `contains` $\mathtt{g}_2$ in the Geonames ontology, but `obj` is represented as a simple building with no building part in CityGML. The function `isAGroup(obj)` will then return false. So it defines $\mathtt{g}_2$ to be attached to `obj` as $\mathtt{g}_1$. Then it calls `searchLink` again to check if there is any documents related to $\mathtt{g}_2$, and if there are, it will link them with `obj`. In the case where `obj` would have been defined with building parts, the function `isAGroup(obj)` would have returned true. Then the following steps would have been to retrieve $\mathtt{obj}_1$ as the proper building part matching $\mathtt{g}_2$, to associate them, and to run the `searchLink(doc,`$\mathtt{obj}_2$`)`.

The last case is particular; in order not to miss relevant resources to be displayed to the user because the 3D object they should be linked to, is outside the current 3DCM. We introduce here the coverage of the current 3D model as `cov(model)`. Its behaviour is identical to `cov(obj)`.

5. Link by spatial proximity

If `cov(doc)` contains an individual e, and `cov(model)` an individuals f . If e is geographically close to f, then a suggestion link can be created between `doc` and `obj`, where `obj` is an object from the model and is spatially positioned next to e. The `suggested_link` is equal to the addition of the `suggested_link` value from the case 3, and `1/euclidean_distance(e,f)`.

Example. If `cov(model)` contains the geoname (*f*) “Lyon 3rd” which is the 3rd district of the city of Lyon in France. `Cov(doc)` contains the geoname (*e*) “Lyon 6th” which is the 6th district of Lyon. *e* and *f* are spatially close, as they have a border in common. The algorithm will create a suggestion link between `doc` and an object `obj` contained in *f* that is close to the border between *f* and *e*.

This algorithm generates a matrix `M`, where each column corresponds to an object in a 3DCM and each line represents a document stored in the library. If there is a link between a document and a city object, then `M(doc,obj)` will contain the link value : `explicit_link` or `suggested_link`.

4.4.2 Example

To illustrate the principle of the algorithm, we detail below a first example of identification and link through non-geometric entities.

Example 1 - Airport

`doc` is an article from the “Journal de l’aviation”[34] explaining that Swiss airline is now equipping its head crew members with digital tablets. The document is annotated with different concepts, such as “`Swiss`”, “`airline`”, “`tablets`” and “`crew`”. `Cov(doc)` is the geoname *s* that represents the whole Switzerland country (geonameId: 2658434).

We use in this example the 3DCM of Geneva Canton. It contains `obj`, which is a 3D building representing the Geneva airport and described with a `BuildingFunctionType` with the value `Airport_Building`. *e* is an individual of the Geonames ontology:

{ label: “*Aéroport Genève Cointrin*”,
geonameId: 2660644,
feature code: `Airport (AIRP)` }.

In the annotation ontology, `AIRP` is aligned with `Airport_Building`, and *e* is positioned inside the object footprint. So `obj` is identified with *e*.

Next step is to find if there is a link between the document and an object in the model. `Cov(doc)=s` and this geoname contains the 3DCM coverage (Geneva Canton). This implies that a link is possible with an object in the model. `Idx(doc)` contains “`airline`” which is linked with the concept “`airport`” in Urbamet. In the ontology, this Urbamet concept is linked with the Geonames feature code `AIRP`. So there is a link between `doc` and all the objects in the model

[34] Can be accessed at: http://www.journal-aviation.com/actualites/17176-le-metier-de-chef-de-cabine-se-numerise-chez-swiss

that are defined as airports, here `obj`. Finally, `M(doc,obj)` contains `explicit_link=true`.

In Figure 38, there are six geonames represented by the different icons. `doc` is associated with the object that is defined by the geoname with the plane icon.

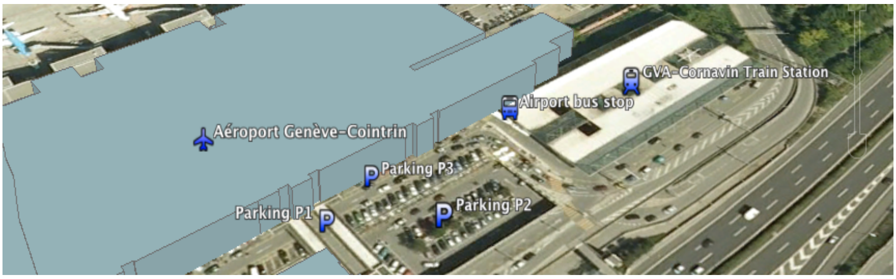


Figure 38 Geneva international airport

Example 2 - University

This next example illustrates the cases where the algorithm needs to use objects that are not identified with geonames, but only with CityGML classes, in the model.

doc is the contact webpage of the Human Resources of the University of Geneva[35]. They specify that their offices are located in the first floor of the "Uni Dufour" Building in Geneva. So `cov(doc)` is composed first of a geoname g, that represents the "Uni Dufour" building of the University of Geneva (geonameId: 8285539), and of a class Storey, with the specified value for the name attribute "1". The coverage is the element of the building "Uni Dufour" that is of class Storey and is located in the first floor.

`obj` is the 3D representation of the building "Uni Dufour" in Geneva, and `cov(obj)=g`. This object is a group and is composed of four storeys above the ground. The 3D objects o_1, o_2, o_3 and o_4 represent those four different building parts. They have their name attribute set to the floor level they represent.

`Cov(doc)` and `cov(obj)` both contain the same geoname `g`. So they can be associated, but `cov(doc)` contains also a specification, which is the class `Storey` with the value of the *name* attribute set to "1". As `obj` is a group we can run the `searchLink` function with each of its sub-object and try to find a match with `doc`. Cov(o_2) matches this requirements, so we can complete the matrix with M(doc,o_2) containing `explicit_link=true`.

[35] Can be accessed at: http://www.unige.ch/adm/dirh/contact.html

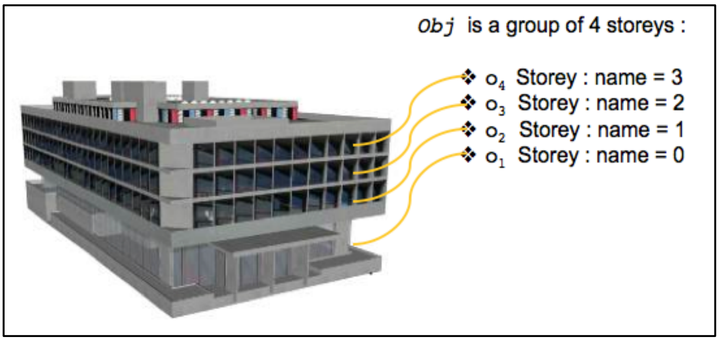


Figure 39 University of Geneva "Uni Dufour" building

4.5 Use Case: Geographic Coverage in Web Search Queries

This use case aims at demonstrating the effectiveness of the geographic coverage in the task of ranking search results. We use the DuckDuckGo[36] search engine with the query "legislation eau potable" *(drinking water laws)* to retrieve our result sets. Our aimed query coverage is Switzerland.

The first set of results is retrieved using a given country as region setting (here Switzerland). Then applying our ranking system described below we get the second and final ranking set. For this test our ranking algorithm is defined as follows: if the geographic coverage of the document is equal to the geographic coverage of the query, then the ranking weight is 1. This weight decreases along the parents and siblings ancestor tree of our desired coverage, as shown in Figure 40.

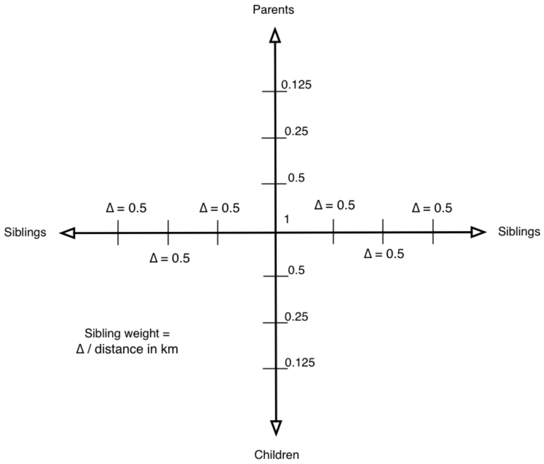


Figure 40 Ranking weight scale

[36] http://duckduckgo.com/

If the geographic coverage of the result is a child or a parent of the query coverage, then the ranking weight is 0.5. For example, if the query coverage is “Switzerland” and the document coverage is “Europe”, the ranking weight will be 0.5. The next rank in the ancestor level is then weighted at 0.25, *etc.* Each step further from the requested coverage divides the previous weight by 2. Siblings are less valuable and so decrease significantly. The sibling’s coverage has a ranking weight value of 0.5 divided by the kilometric distance with the desired coverage. For the test below, the kilometric values were taken from Google maps with a straight path between the Switzerland and the document coverage. For example, if the query coverage is “Switzerland” and the document coverage is “Quebec”, the ranking weight will be 0.5/5580 = 0.00009.

Figure 47 and Figure 48 show three different ranking results. Highlighted in green are the results that match the desired geographic coverage: either France or Switzerland. The second column is the result of the query using DuckDuckGo with the region settings set to France Figure 47, or Switzerland Figure 48, and their respective geographic coverage in the third column. The fourth column is the reorganisation of the results according to the ranking process.

For the Switzerland example, we can see that the first and fifth results have a coverage value equal to France. The third, fourth and sixth results are closer to the requested coverage, as their respective coverages are states of Switzerland (Fribourg and Geneva). Only the second result has the required geographic coverage.

These results demonstrate clearly the effectiveness of the geographic coverage in ranking task. When we apply our ranking scale, the local results are placed first on the list. Moreover, the rest of the results are organised following their closeness with the query coverage.

To evaluate the results ranking, we calculated the Cumulated Gain (CG) and Discounted Cumulated Gain (DCG) as detailed in [57] and based on [58], [59] for the Switzerland example. We graded the 30 first results on a scale from 0 to 3 according to their relevance on both the geographic and the content aspect. We did the calculations for both the query results according to DuckDuckGo and the same list reordered according to our ranking method based on our notion of coverage.

$$\mathrm{CG}[i] = \begin{cases} \mathrm{G}[1] & \textit{if } i = 1; \\ \mathrm{G}[i] + \mathrm{CG}[i-1] & i > 1; \end{cases}$$

Where $\mathrm{CG}[i]$ refers to the cumulated gain at the i-th position of the ranking of the query. G is the gain vector for a query and reflects the relevance score for each result.

$$\mathrm{DCG}[i] = \begin{cases} \mathrm{G}[1] & if\ i = 1; \\ \dfrac{\mathrm{G}[i]}{\log_2 i} + \mathrm{DCG}[i-1] & i > 1; \end{cases}$$

Where $\mathrm{DCG}[i]$ refers to the discounted cumulated gain at the i-th position of the ranking of the query. Similarly to $\mathrm{CG}[i]$, G is the gain vector.

As shown in Figure 43 and Figure 44, in both the CGs and DCGs graphs, the graph with our ranking algorithm, represented in blue, values is on top of the ones from the DuckDuckGo values. We can deduce that our ranking algorithm is slightly more efficient than DuckDuckGo's if we take into account the geographic coverage of the query.

We also run this test for the search engines Qwant[37], SwissCows[38], both with the region setting set to Switzerland, Google.ch[39] and DuckDuckGo with no region settings. For each search engine we found similar results as in our first experiment as shown in Figure 45. Each time our ranking DCG graph is on top of the one for the search engine.

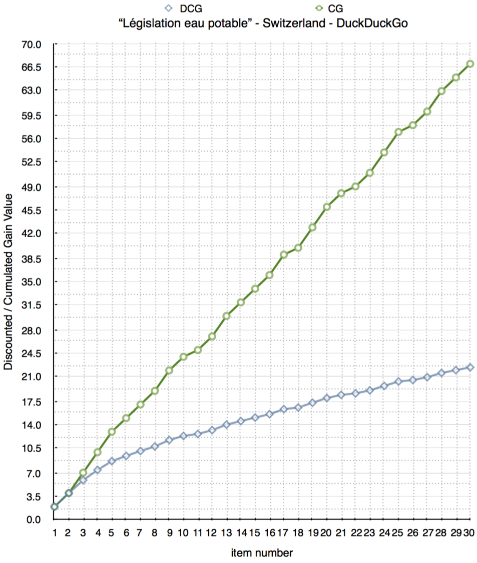


Figure 41 CG and DCG for DuckDuckGo

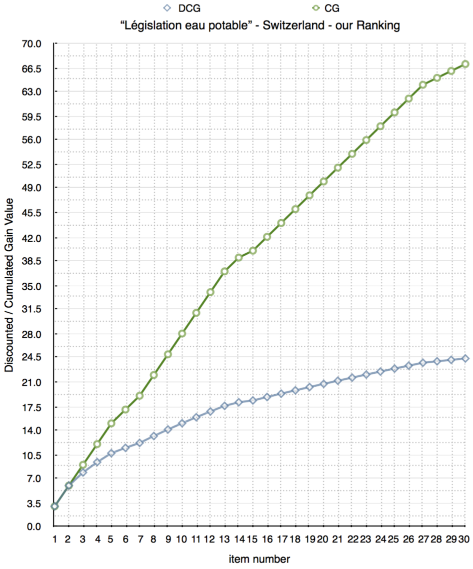


Figure 42 CG and DCG for our ranking algorithm

[37] https://www.qwant.com

[38] https://swisscows.ch

[39] http://google.ch

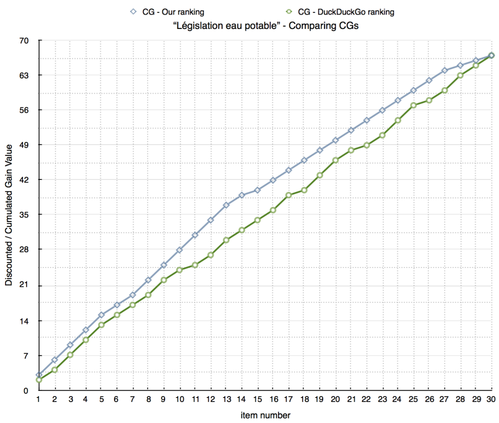


Figure 43 Comparing CGs

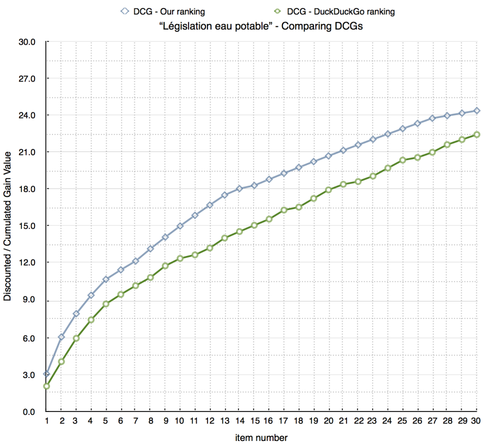


Figure 44 Comparing DCGs

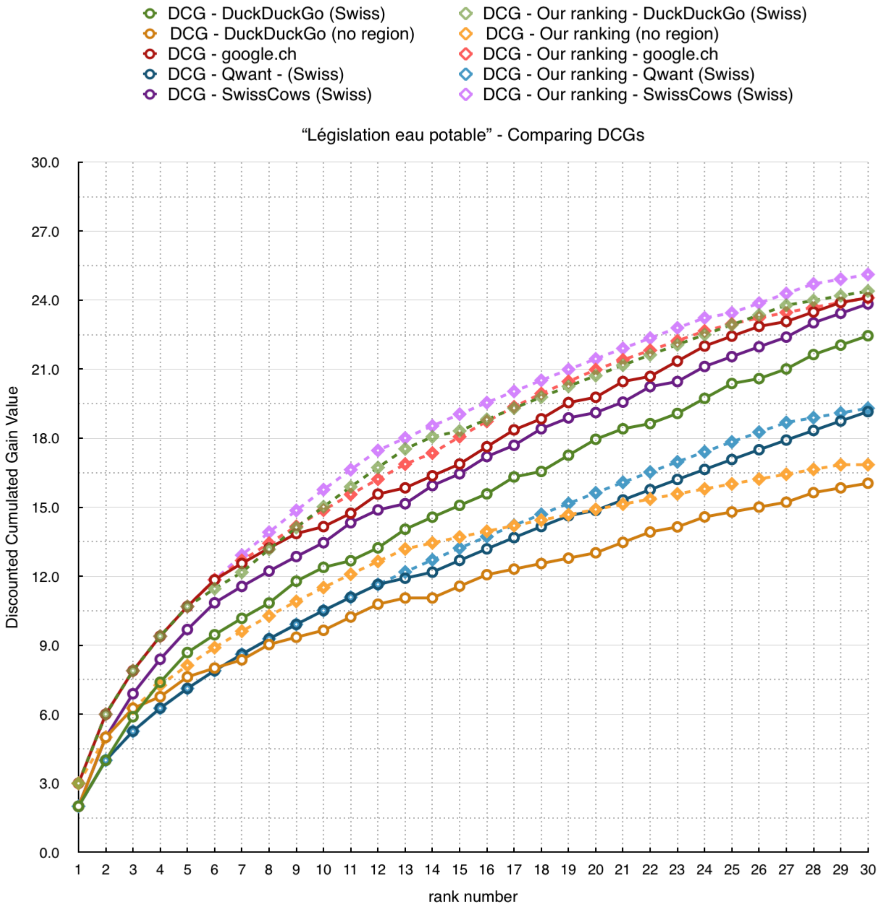


Figure 45 "legislation eau potable" DCGs full comparison

To validate further those results, we also run the test with the query “formation continue informatique” (Advanced studies computing) for the Switzerland coverage on DuckDuckGo with Swiss coverage and with no coverage. The result for the Discounted Cumulated Gain is shown in Figure 46. We can see that we again have a DCG curve on top using our ranking, even though the difference lessens when we add the region settings.

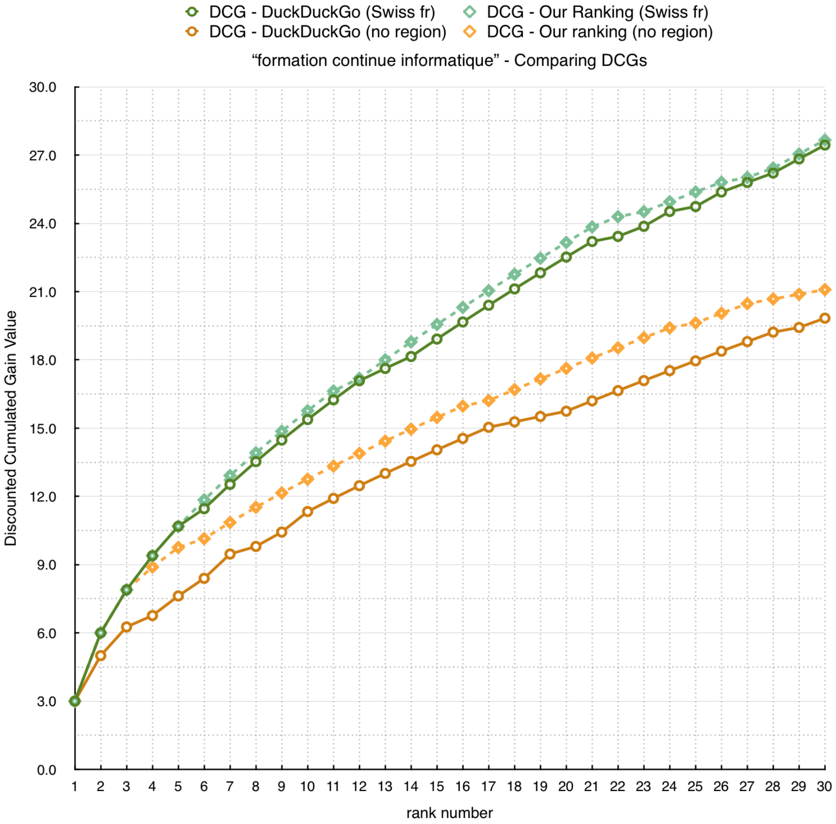
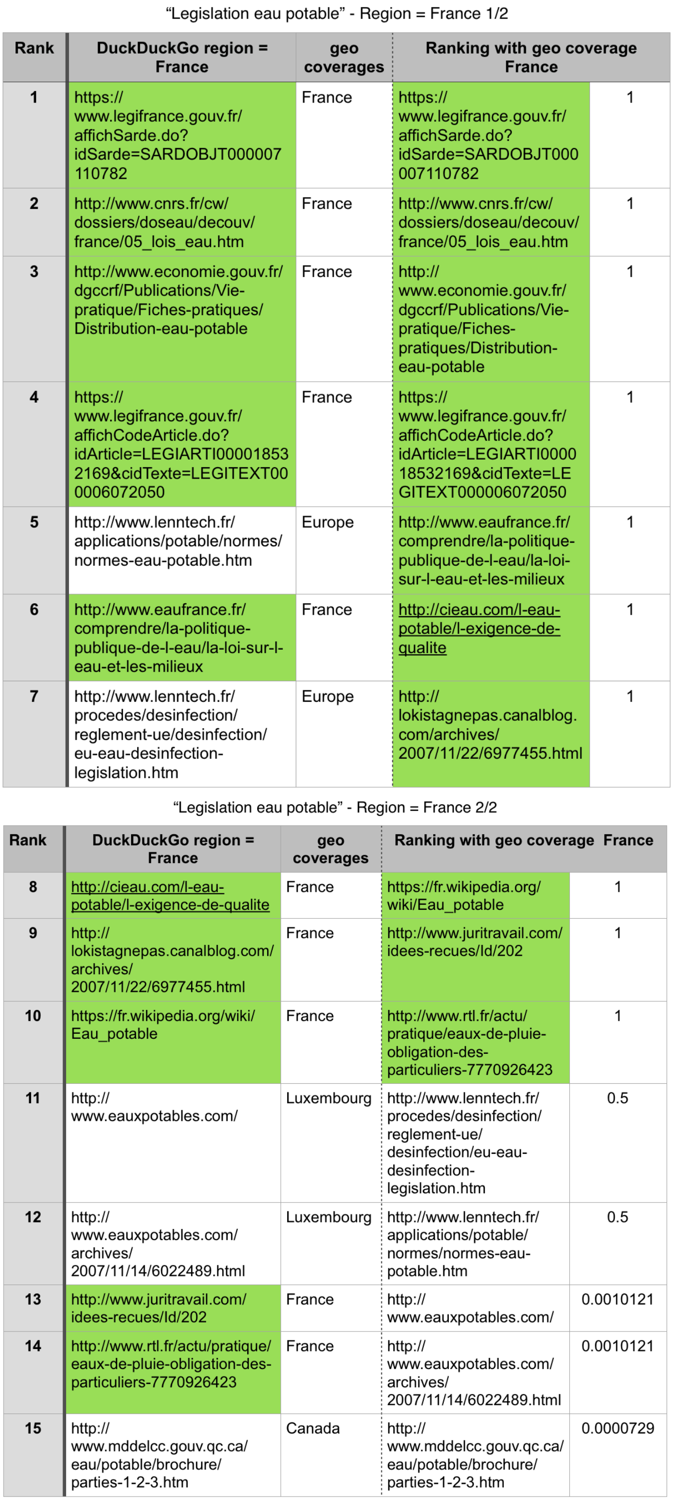


Figure 46 "formation continue informatique" DCGs comparison



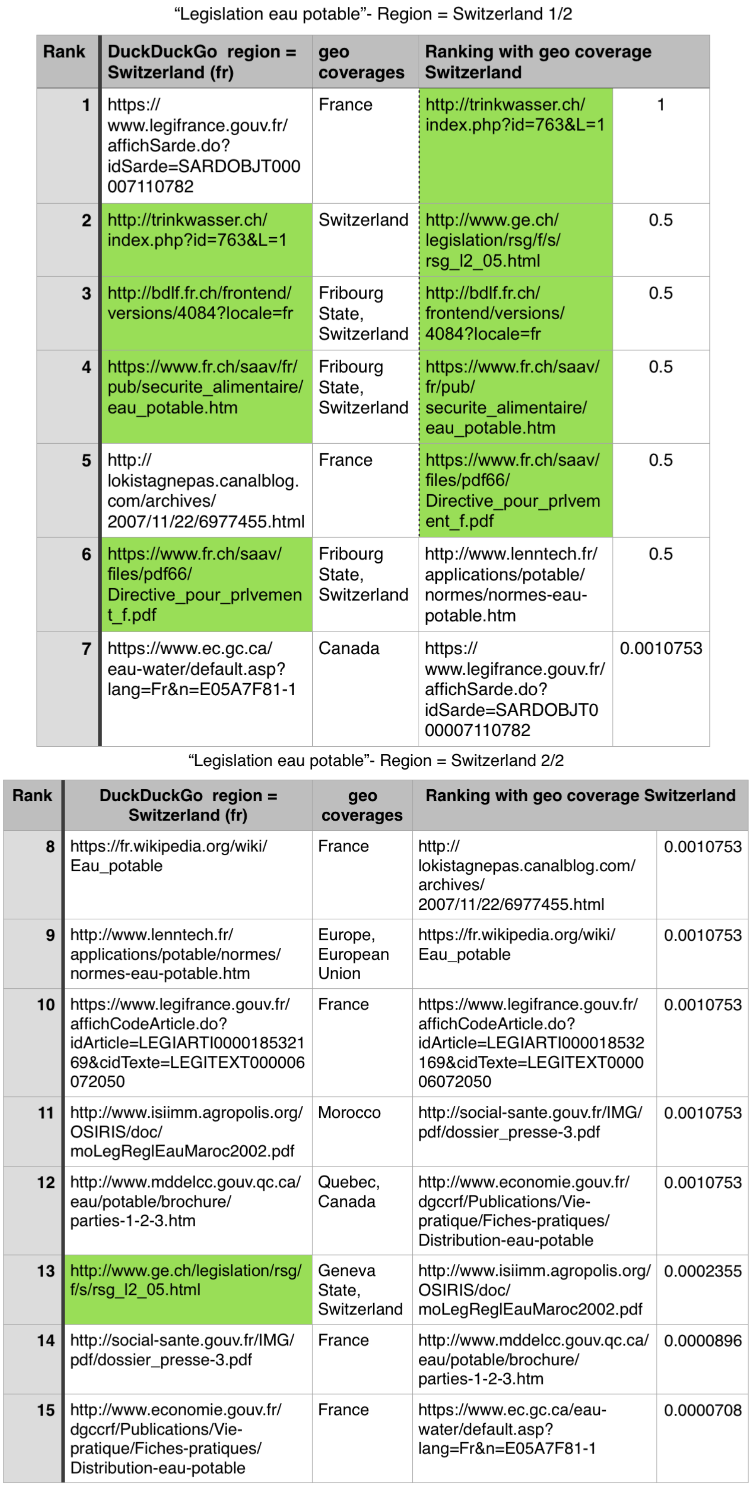
“Legislation eau potable” - Region = France 1/2

Rank	DuckDuckGo region = France	geo coverages	Ranking with geo coverage France	
1	https://www.legifrance.gouv.fr/affichSarde.do?idSarde=SARDOBJT000007110782	France	https://www.legifrance.gouv.fr/affichSarde.do?idSarde=SARDOBJT000007110782	1
2	http://www.cnrs.fr/cw/dossiers/doseau/decouv/france/05_lois_eau.htm	France	http://www.cnrs.fr/cw/dossiers/doseau/decouv/france/05_lois_eau.htm	1
3	http://www.economie.gouv.fr/dgccrf/Publications/Vie-pratique/Fiches-pratiques/Distribution-eau-potable	France	http://www.economie.gouv.fr/dgccrf/Publications/Vie-pratique/Fiches-pratiques/Distribution-eau-potable	1
4	https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000018532169&cidTexte=LEGITEXT000006072050	France	https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000018532169&cidTexte=LEGITEXT000006072050	1
5	http://www.lenntech.fr/applications/potable/normes/normes-eau-potable.htm	Europe	http://www.eaufrance.fr/comprendre/la-politique-publique-de-l-eau/la-loi-sur-l-eau-et-les-milieux	1
6	http://www.eaufrance.fr/comprendre/la-politique-publique-de-l-eau/la-loi-sur-l-eau-et-les-milieux	France	http://cieau.com/l-eau-potable/l-exigence-de-qualite	1
7	http://www.lenntech.fr/procedes/desinfection/reglement-ue/desinfection/eu-eau-desinfection-legislation.htm	Europe	http://lokistagnepas.canalblog.com/archives/2007/11/22/6977455.html	1

“Legislation eau potable” - Region = France 2/2

Rank	DuckDuckGo region = France	geo coverages	Ranking with geo coverage France	
8	http://cieau.com/l-eau-potable/l-exigence-de-qualite	France	https://fr.wikipedia.org/wiki/Eau_potable	1
9	http://lokistagnepas.canalblog.com/archives/2007/11/22/6977455.html	France	http://www.juritravail.com/idees-recues/Id/202	1
10	https://fr.wikipedia.org/wiki/Eau_potable	France	http://www.rtl.fr/actu/pratique/eaux-de-pluie-obligation-des-particuliers-7770926423	1
11	http://www.eauxpotables.com/	Luxembourg	http://www.lenntech.fr/procedes/desinfection/reglement-ue/desinfection/eu-eau-desinfection-legislation.htm	0.5
12	http://www.eauxpotables.com/archives/2007/11/14/6022489.html	Luxembourg	http://www.lenntech.fr/applications/potable/normes/normes-eau-potable.htm	0.5
13	http://www.juritravail.com/idees-recues/Id/202	France	http://www.eauxpotables.com/	0.0010121
14	http://www.rtl.fr/actu/pratique/eaux-de-pluie-obligation-des-particuliers-7770926423	France	http://www.eauxpotables.com/archives/2007/11/14/6022489.html	0.0010121
15	http://www.mddelcc.gouv.qc.ca/eau/potable/brochure/parties-1-2-3.htm	Canada	http://www.mddelcc.gouv.qc.ca/eau/potable/brochure/parties-1-2-3.htm	0.0000729

Figure 47 Geographic coverage - ranking test result for region = France



"Legislation eau potable"- Region = Switzerland 1/2

Rank	DuckDuckGo region = Switzerland (fr)	geo coverages	Ranking with geo coverage Switzerland	
1	https://www.legifrance.gouv.fr/affichSarde.do?idSarde=SARDOBJT000007110782	France	http://trinkwasser.ch/index.php?id=763&L=1	1
2	http://trinkwasser.ch/index.php?id=763&L=1	Switzerland	http://www.ge.ch/legislation/rsg/f/s/rsg_l2_05.html	0.5
3	http://bdlf.fr.ch/frontend/versions/4084?locale=fr	Fribourg State, Switzerland	http://bdlf.fr.ch/frontend/versions/4084?locale=fr	0.5
4	https://www.fr.ch/saav/fr/pub/securite_alimentaire/eau_potable.htm	Fribourg State, Switzerland	https://www.fr.ch/saav/fr/pub/securite_alimentaire/eau_potable.htm	0.5
5	http://lokistagnepas.canalblog.com/archives/2007/11/22/6977455.html	France	https://www.fr.ch/saav/files/pdf66/Directive_pour_prlvement_f.pdf	0.5
6	https://www.fr.ch/saav/files/pdf66/Directive_pour_prlvement_f.pdf	Fribourg State, Switzerland	http://www.lenntech.fr/applications/potable/normes/normes-eau-potable.htm	0.5
7	https://www.ec.gc.ca/eau-water/default.asp?lang=Fr&n=E05A7F81-1	Canada	https://www.legifrance.gouv.fr/affichSarde.do?idSarde=SARDOBJT000007110782	0.0010753

"Legislation eau potable"- Region = Switzerland 2/2

Rank	DuckDuckGo region = Switzerland (fr)	geo coverages	Ranking with geo coverage Switzerland	
8	https://fr.wikipedia.org/wiki/Eau_potable	France	http://lokistagnepas.canalblog.com/archives/2007/11/22/6977455.html	0.0010753
9	http://www.lenntech.fr/applications/potable/normes/normes-eau-potable.htm	Europe, European Union	https://fr.wikipedia.org/wiki/Eau_potable	0.0010753
10	https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000018532169&cidTexte=LEGITEXT000006072050	France	https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000018532169&cidTexte=LEGITEXT000006072050	0.0010753
11	http://www.isiimm.agropolis.org/OSIRIS/doc/moLegReglEauMaroc2002.pdf	Morocco	http://social-sante.gouv.fr/IMG/pdf/dossier_presse-3.pdf	0.0010753
12	http://www.mddelcc.gouv.qc.ca/eau/potable/brochure/parties-1-2-3.htm	Quebec, Canada	http://www.economie.gouv.fr/dgccrf/Publications/Vie-pratique/Fiches-pratiques/Distribution-eau-potable	0.0010753
13	http://www.ge.ch/legislation/rsg/f/s/rsg_l2_05.html	Geneva State, Switzerland	http://www.isiimm.agropolis.org/OSIRIS/doc/moLegReglEauMaroc2002.pdf	0.0002355
14	http://social-sante.gouv.fr/IMG/pdf/dossier_presse-3.pdf	France	http://www.mddelcc.gouv.qc.ca/eau/potable/brochure/parties-1-2-3.htm	0.0000896
15	http://www.economie.gouv.fr/dgccrf/Publications/Vie-pratique/Fiches-pratiques/Distribution-eau-potable	France	https://www.ec.gc.ca/eau-water/default.asp?lang=Fr&n=E05A7F81-1	0.0000708

Figure 48 Geographic coverage - ranking test result for region = Switzerland

4.6 Summary

In this chapter we have presented an implementation solution for the annotation model defined in Chapter 3. The implementation consists in the alignment of ontologies and thesaurus: CityGML, Geonames and Urbamet. The result is published in our repository on GitLab.

In section 4.2, we implemented the annotation algorithm for spatial resources. The annotation is done through the identification of the different objects that compose the 3D scene.

Section 4.3 presents the location of documents, held in the repository, in the 3D scene. The location of the documents is found by matching the documents & the corresponding 3D objects according to their respective coverage. The output of this process is a matrix that references the relations and the weight of the relation between each document and each object from the spatial resource.

Finally, we have presented a use case to validate the effectiveness of geographic coverage in ranking tasks for search results. We computed the CG and DCG formulas to validate our results.

Chapter 5

Use Case: Creating a New Technique for Extracting Geographic Information from Tags

The advent of social multimedia repositories such as Flickr or YouTube has raised a lot of interest for mining these sources in order to extract and discover knowledge for many purposes such as geographic information [67], [68]. These implicit volunteered geographic information (VGI) data sources offer promising opportunities to discover geospatial knowledge. Unfortunately the data cannot be directly exploited and need to be processed [69].

Our objectives here are to propose a simple non-statistical framework, using existing VGI sources, to find the characteristics of geographic places, and improve the formulation and the precision of search queries.

Most of the social repositories rely on folksonomies to annotate the multimedia contents. Folksonomies offer a very simple and attractive framework to annotate multimedia contents. However, they are greatly lacking of semantics in order to exploit them appropriately. Various approaches can be applied in order to discover this semantics from many sources [70]. The analysis of folksonomies attached to social image repositories has been proposed to uncover geospatial knowledge [71]-[75].

It is important to differentiate spatially explicit vs. implicit (such as Flickr) social sharing sources [76]. With implicit sources, the distribution of geographic information is clustered in popular locations (related to entertainment and tourism). It suggests that traditional clustering and aggregation methods used to discover spatial semantics and knowledge are not appropriate in other places, where the density of images is low. In [77] Flickr tags are aggregated at multiple scales to study their spatial and thematic properties. The study confirms that for sources like Flickr, tags are clustered in touristic and entertainment areas. It also identifies a strong interaction between tag spatial semantics and the associated spatial scale.

A three place–related facets classification is proposed in [78] that includes: elements (objects and people that can be identified in photos), qualities (modify elements or suggest feelings) and activities. In the paper, annotators achieve the classification manually. A similar approach is adopted in [79], which suggests using the Pansofsky-Shatford facet matrix to provide a formal universal image description model. The matrix is based on two aspects including facets. The first aspect is organised into the Who? What? Where and When? facets and the second one according to the Specific of, Generic of and About facets. Based on this formal framework, Purves *et al.* show that it is possible to improve image descriptions (and therefore image annotation analysis) particularly with the proposed notion of place.

In this chapter we aim to validate the principle of spatial coverage by applying it to the development of a categorisation technique, which allows the identification, and the finding of the characteristics of geographic places. This technique is built for the processing of documents on places with small amount of data, on which statistical approach cannot be applied. It can also be used to pre-process batch of data before applying this kind of approach.

5.1 Method Presentation

Using the available folksonomy in VGI services we propose a knowledge-based approach to highlight and possibly discover the characteristics of geographic places. The characteristics of a place are non-circumstantial information that can qualify the place, such as the place function, its usage, its name, the material it is made of, its colour *etc.* The method is based on the notion of spatial coverage and a model of tags categorisation and semantic identification, using semantic services such as GeoNames[40], OpenStreetMap[41] or WordNet [45]. To illustrate the method, we only use photos as document; more precisely we use photos from Flickr[42].

Semantic classification techniques to retrieve the characteristics (function, usage...) of places is usually based on statistical approach [71]-[75]. This kind of approaches work well on touristic areas or places with a large number of pictures. Our method allows to pre-process photos individually, that can be later processed as batch in a statistical approach; or to process photos individually for places with a small number of pictures, and still be able to semantically analyse the content. The found characteristics allow tag disambiguation and can be used to complete the semantic gap on places and POIs such as the function of buildings, which can exist in geographic services. CityGML [61] is a data model to design 3D city models and objects with respect to their geometry, topology, semantics and appearance. There exist a large number of city or neighbourhoods

[40] http://www.geonames.org

[41] http://www.openstreetmap.org/about

[42] https://www.flickr.com

available in 3D in CityGML format. However in most cases the semantic is left empty. We could use our method and the notion of spatial coverage to fill those missing semantic parameters.

Our model is firstly based on the distinction between the spatial coverage and the spatial references of a document. This distinction helps the identification of the places that appear in the picture and for which we are trying to identify the characteristics. We define the spatial coverage for photos in section 3.2.2 as: “what spatial region we see in the resource”. It is a combination of semantic and geographic notions that describes precisely the spatiotemporal region shown in the picture. We make a distinction between the geographic and spatial content of a resource and its actual spatial coverage. The goal is to differentiate the spatial content, which is composed of places that are not part of the coverage, and the actual spatial context of the resource.

The annotations of Flickr photos consist of freeform text tags created by the owner of the photo and if possible of a location tag and GPS coordinates. We propose to identify each tag of the set semantically and organise them according to the categories described below. The categorisation allows us to identify three things: the geo-spatial characteristics of places by filtering what is in the coverage and what is not; the characteristics that are only true punctually compared to the time the picture was taken, like a bike passing by or a bird, the blue sky, *etc.*; and finally identify the permanent characteristics that will allow us to determine the semantic characteristics of the place.

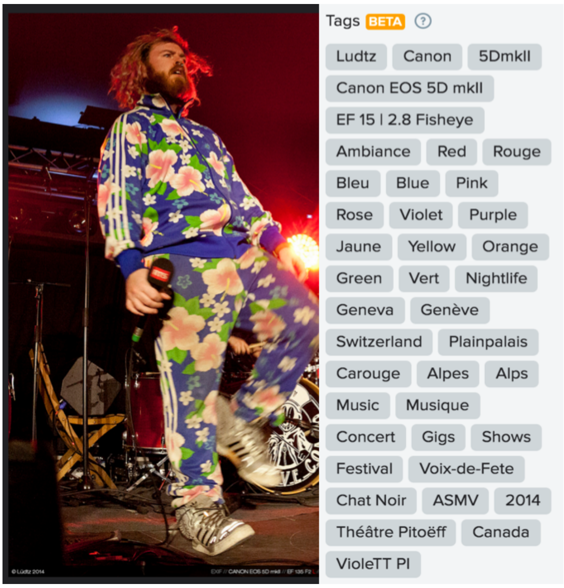


Figure 49 Photo "VioleTT Pi" from Ludtz

First, we need to sort the tags in two sets: the geographic and the non-geographic ones. The geographic tags are the geographic features and classes and their translation. Their identification is made using geographic services such as OpenStreetMap and GeoNames. The geographic features are the instances such as countries, cities, POIs, *etc.* The geographic classes are the main types categories of the geographic features for example, “tunnel”, “city”, “university”, “lake”, “forest”, “restaurant”, *etc.* [80], [81]. Those tags can be part of the picture coverage or spatial references.

For example, the photograph [https://flic.kr/p/m9ZBPB] (as shown in Figure 49) of a concert of a Canadian artist at the Théâtre Pitoëff in Geneva is tagged with both "Geneva" and "Canada". Both are geographic tags, but one is part of the spatial coverage of the picture and the other is a spatial reference. The concert took place in Geneva, so this is where the photo is true; "Geneva" is then part of the photo spatial coverage. The artist in the picture comes from Quebec, so the tag "Canada" is a spatial reference and part of the content of the photo. We should not have this photo as a result to a query for photos of Canada. The distinction between the spatial coverage and the spatial references allows clearing up this kind of ambiguity in the query result.

Once every geographic tag has been identified, the rest of the tags are all considered non-geographic tags. In order to filter the unwanted information, we have identified the following categories: Event, Temporal, Weather, Actors, Meta and Colour. Those categories have been inspired by [78], [79].

The **Event** category regroups all the tags that refer to events, or punctual things. In our example, the tag “Voix-de-fête”, as the name of a festival is part of the Event category.

The **Temporal** category regroups time-related tags such as “afternoon”, “night”, “day”, *etc.* It can be considered as a subcategory of the Event category.

The **Weather** category regroups weather tags such as “cloud”, “sun”, “rain”, “storm”, *etc.* Like the Temporal category, the Weather one can also be considered a subcategory of Event.

The **Actor** category regroups people or in movement objects references that are not always at the place the photo was taken, such as “jogger”, “tram”, “car”, “birds”, *etc.*

The **Unidentified** category regroups tags that do not refer to content or coverage information: any photographic tags such as the photographer nickname, camera brand, camera model number and photo awards. In our example, it would be: “Ludtz”, “Canon”, “5DmkII”, “Canon EOS 5D mkII”, “EF 15 | 2.8 Fisheye”.

The **Colour** category regroups all the tags that reference colours. They are often found in Flickr tags but do not refer to meaningful information in our use case.

The rest of the tags are the probable invariant characteristics that are recurrent in similar pictures and might give pertinent information on the places in the picture. For example, the photo shown in Figure 49 has the following tags left after categorisation: "Ambiance", "Music", "Musique", "Nightlife". We can deduce from this, that the "Théâtre Pitoëff" geographic feature is related to music, nightlife and ambience. We remark that the rest of the tags are those interesting in the context of finding the characteristics and functions of geographic features.

5.2 Algorithm

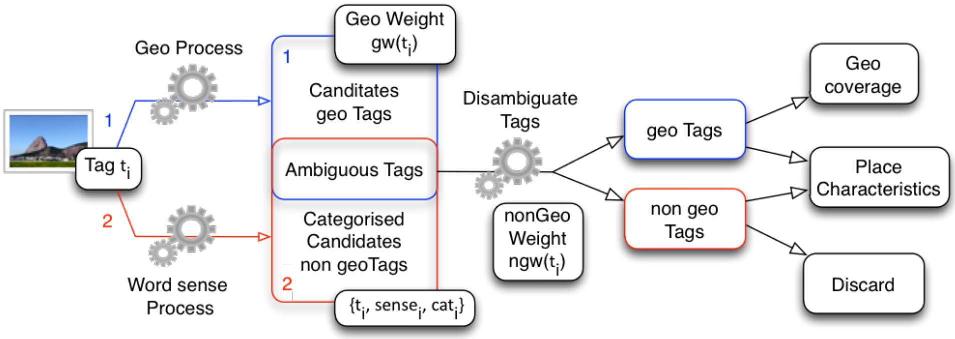


Figure 50 Methodology global schema

To identify places characteristics, we need to determine the tags that represent geographic features (geographic tags) and more particularly the coverage tags that represent features that appear in the picture. Secondly, we need to identify non-geographic tags that can hold information on the place characteristics.

The algorithm organises the tags in order to eliminate circumstantial tags, as they are not stable characteristics of a place, and in order to highlight specific geographic tags. Circumstantial tags are tags such as: "rain", "jogger", "autumn", "Tour de France"...

Our method, depicted in Figure 50, is based on the notion of spatial coverage and a model of tags categorisation and their semantic identification, using semantic services such as GeoNames, BabelFy [82] or WordNet. To illustrate the method, we only use photos as document. More precisely we use photos from Flickr.

The method is divided as follow:

1. Computing a geographic weight for each tag and associating a geo entity when possible.
2. Associating a sense (and a category) to each tag.

3. Extracting tags that represent geographic features in the picture coverage.
4. Extracting tags that characterise the geographic features.

5.2.1 Geo Process

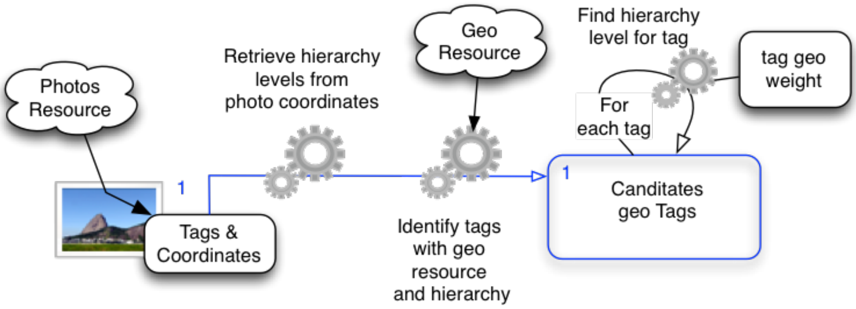

Photos Resource
Retrieve hierarchy levels from photo coordinates
Geo Resource
Find hierarchy level for tag
tag geo weight
For each tag
1
Tags & Coordinates
Identify tags with geo resource and hierarchy
1
Canditates geo Tags

Figure 51 Geo process

The geo process, as shown in Figure 51, aims to identify the geographic candidates within the tag set and to assign them a geo weight. If a tag T is the name of one or more geographic features $\{f_1, ..., f_k\}$, and f_j is the feature that is closest to the photo location p_loc, the geo weight $gw(T)$ of T is a decreasing function of the distance between p_loc and f_j. This definition is based on the fact that there is a high probability that the term used to tag a picture is one referencing a nearby geographic feature.

To simplify the subsequent processes we use discrete geo weight values that are based on a simplified administrative division hierarchy (*neighbourhood, locality, county, region, country, world)* centred at the photo location. The hierarchies are derived from Yahoo Places. The lowest the hierarchy level, the highest the weight value. Each hierarchy level determines a bounding box that is used to query a gazetteer, with each tag individually for each hierarchy level starting at the lowest (neighbourhood), until the tag is identified. If the tag is not found it is assigned a null geo weight.

The identification of a geo tag associates a geographic instance or feature to the tag and its class. The geo feature classes are feature functions like the Geonames feature codes[43] or the CityGML codes [83]. They define the function or nature of the feature.

Below is the pseudo code for the process:

[43] http://www.geonames.org/export/codes.html

```
//Retrieve hierarchy entities:
//p_loc: photo location geo entity
Retrieve_photo_location (photo) -> p_loc;
//h: array of location composing the hierarchy
Yahoo_location_hierarchy(p_loc) -> h;
//geo_h: array of location composing the hierarchy composed of
   geonames entity
Geonames_match_entites(h) -> geo_h;

//identify T:
While !found{
       Foreach geo_h as gh {
              //geoname: geonames entity
              Query_geonames(boundingbox=gh; T_label) -> geoname;
              If found{
                     //gh_level: hierarchy value of gh
                            (neighbourhood, locality, city, region…)
                     gw(T)=gh_level;
                     found == true;
              }else{
                     T is not geographic
              }
       }
}
```

5.2.2 Word Sense Process

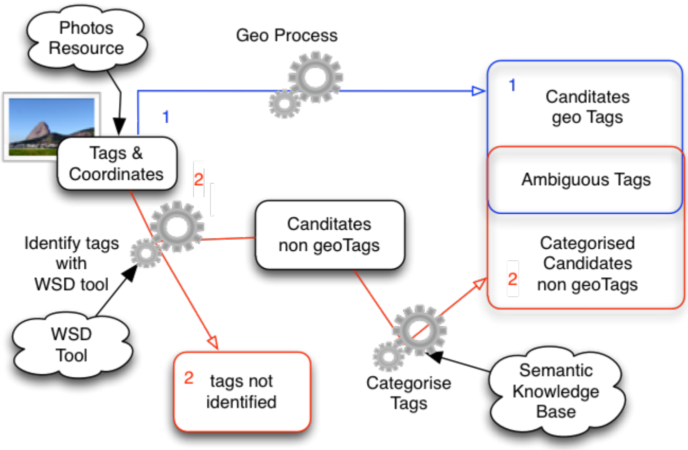


Figure 52 Word sense process

The second step is the word sense process, shown in Figure 52. It is intended to identify and classify the tags that do not designate geographic features. It is based on a word sense disambiguation (WSD) process.

All the photo's tags are concatenated to form a pseudo sentence that is passed to a WSD tool. The result is a set of pairs {(*tag_i*, *$sense_i$*)} where *$sense_i$* is the identifier of a concept in a reference knowledge base or lexical database (Dbpedia[44], WordNet, *etc.*). All the identified tags are then categorised as *Event*, *Temporal*, *Weather*, *Actor*, *Color*, *Geographic feature*, *Geographic feature class*, or *Unidentified*. The categories are inspired by the work of Purves *et al.* [78]. The classification is done by examining the class of each sense (and their superclasses if necessary). For each category, parent classes are selected form the knowledge base. If the concept is a child of those classes or an instance of them, we assume that the tag belongs to the category.

The *Geographic feature class* category contains concepts such as *mountain*, *hospital*, *etc.* whereas the *Geographic feature* category contains instances of these classes (*Matterhorn*, *Mayo Clinic*, *etc.*). The identification of geo tags during this process is done to compensate the possible missing identification from the geo process, most probably due to missing information in the geo gazetteers.

Below is the pseudo code for the process:

```
Get_all_photo_tags(photo) -> tag_list;
//call webservice (WS) babelnet
WS_Babelnet_get_senses(tag_list) -> {tag$_i$, sense$_i$};
Foreach pair tag-sense{
        wikidata_tag_identification(tag) -> wikidata_entity;
        // categorised tag according to entity parent
        Categorise_tag_wikidata(wikidata_entity) -> tag_categ;
        If tag_categ is null -> tag_categ=''unidentifed'';
}
```

[44] wiki.dbpedia.org

5.2.3 Disambiguation & Tags Extraction

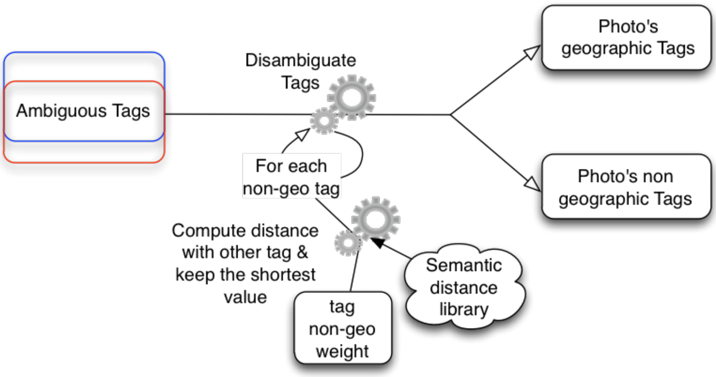

Ambiguous Tags
Disambiguate Tags
Photo's geographic Tags
Photo's non geographic Tags
For each non-geo tag
Compute distance with other tag & keep the shortest value
Semantic distance library
tag non-geo weight

Figure 53 Disambiguation process

Some of the tags might be identified both as a non-geographic concept (*e.g. Event*) and as a geographic entity (by the geo process). The disambiguation process, as depicted in Figure 53, is based on the idea that a tag that is semantically close to another non-geographic tag is probably a non-geographic tag. This is determined by using a semantic distance library tool with WordNet dictionary, to compute each semantic distance between the ambiguous tag and the other tags of the photo. The smallest distance is selected as the non-geo weight *ngw(T)* of the ambiguous tag. The disambiguation decision is then made by comparing *ngw(T)* and its geo weight *gw(T)*. If the geo identification was done using a low hierarchy level, the geo identification is kept as definition for the tag. If the non-geo weight is high then the non-geo identification is kept.

Each tag has now been processed, and is either identified as a geo tag, as a non-geo tag or not identified. The tags that have not been identified are categorised as “Unidentified” tags.

Below is the pseudo code for the process:

```
Get_all_ambiguous_tags() -> tag_list;
Foreach(tag_list as tag){
        //retrieve all photos for each tag
        Find_photolist(tag) -> photo_list;
        Foreach(photo_list as photo){
                Retieve_tag_list(photo) -> tags;
                Process_sem_distance(tags) -> smallest_distances_list;
        }
        Find_smallest(smallest_distances_list) -> non_geo_weight;
        //Compare_weights and classify tag
        Classify_tag(geo_weight, nonGeo_weight);
}
```

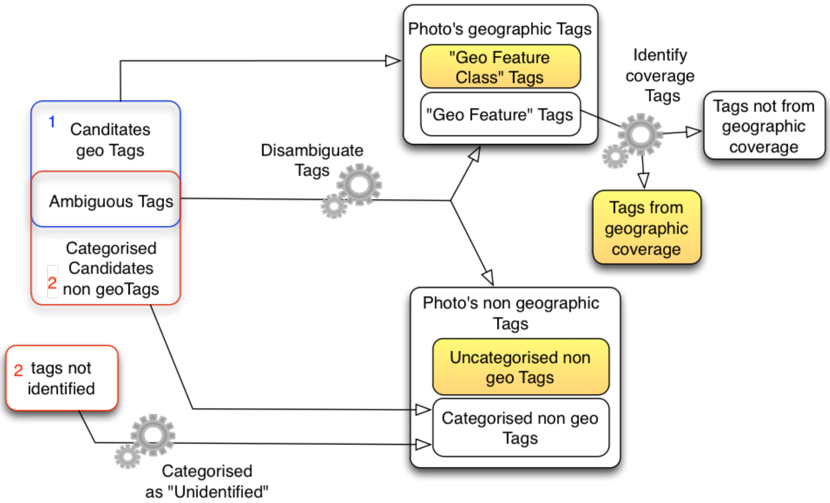


Figure 54 Extraction and selection

The final step is to extract the tags of interest and thus create the “Geo semantic” group, as shown in Figure 54. The “Geo Semantic” group of tags contains all the potential information on the place characteristics.

Within the geo tags we select the ones that belong to the photo coverage and those categorised as *geo feature class*. We consider that a feature belongs to the coverage if its geo weight is “neighbourhood” or if its geo weight is “locality” or “county”, or “region” and it belongs to a class of sufficiently large features (*ocean*, *mountain*, *desert*, *etc.*) The geo tags identified with the lowest geo weights (country, world) are considered too far to be part of the photo coverage. The tags that have been identified as non-geo and that are in the *Event*, *Temporal*, *Weather* or *Actor* categories are considered as circumstantial and are thus eliminated. The remaining tags form the “Semantic enhancement of places” (SEP) group. We consider that they potentially provide permanent characteristics of the places that appear in the photo.

The “Geo semantic” group is the final output of our algorithm.

The complete schema of the algorithm is shown in Figure 54.

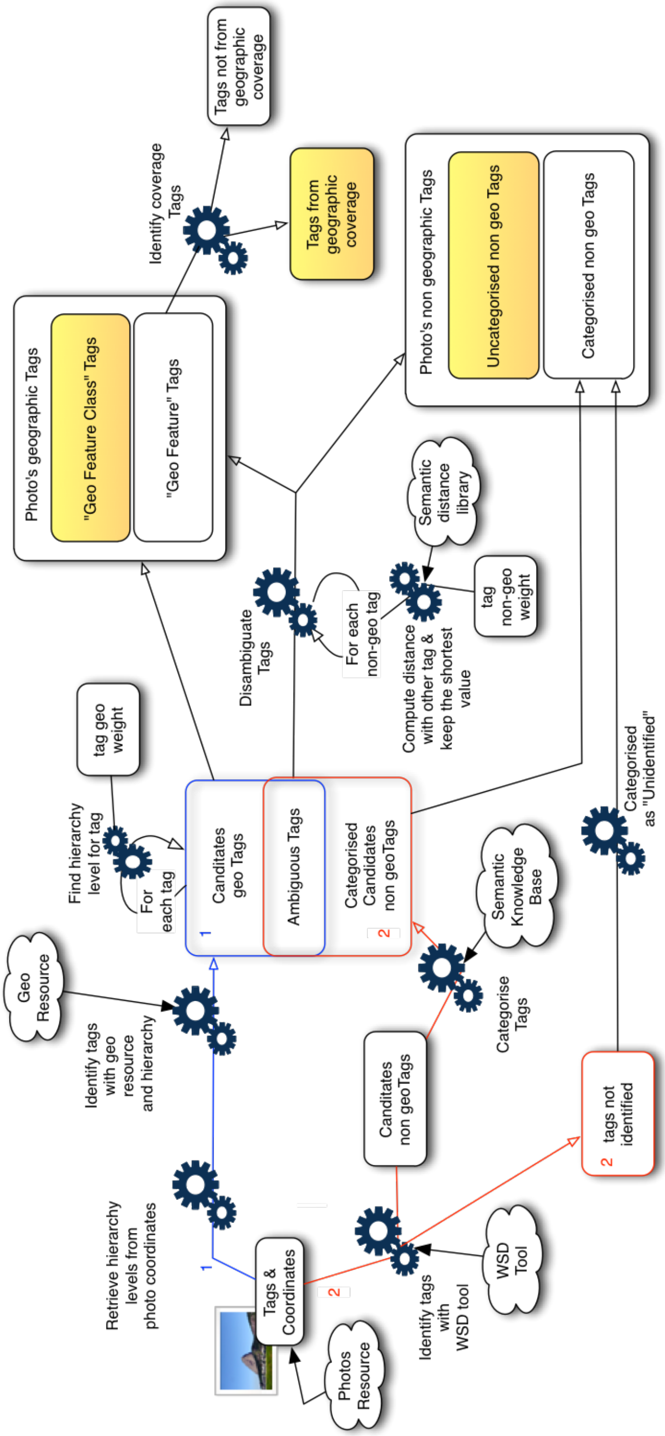

Photos Resource
Tags & Coordinates
1
2
Retrieve hierarchy levels from photo coordinates
Geo Resource
Identify tags with geo resource and hierarchy
Identify tags with WSD tool
WSD Tool
Canditates non geoTags
tags not identified
Categorised as "Unidentified"
Categorise Tags
Semantic Knowledge Base
Canditates geo Tags
Ambiguous Tags
Categorised Candidates non geoTags
Find hierarchy level for tag
For each tag
tag geo weight
Disambiguate Tags
For each non-geo tag
Compute distance with other tag & keep the shortest value
tag non-geo weight
Semantic distance library
Photo's geographic Tags
"Geo Feature Class" Tags
"Geo Feature" Tags
Identify coverage Tags
Tags not from geographic coverage
Tags from geographic coverage
Photo's non geographic Tags
Uncategorised non geo Tags
Categorised non geo Tags

Figure 55 Methodology complete schema

5.3 Implementation

As previously explained, our goal is to identify place characteristics, such as the building function or usage, and the spatial coverage of a picture. To do so, we need to determine the tags that represent geographic features (geographic tags) and more particularly the coverage tags that represent features that appear in the picture. Secondly, we need to identify non-geographic tags that can hold information on the place characteristics.

The tag categorisation algorithm uses external resources (GeoNames, Wikidata[45], BabelFy) and the WOEID (Where On Earth Id) of the photo. Each Flickr photo has a WOEID that refers to a spatial entity of Yahoo! GeoPlanet[46]. We determine the WOEID with the GPS information contained in the EXIF file. Each WOEID belongs to a hierarchy of other geographic feature. The hierarchy goes from the country down to the neighbourhood.

The algorithm is composed of the following steps:

1. Retrieve geo tags in the coverage and in the extended area: the geo process.
2. Disambiguate and classify non-geo tags: the Word Sense (WS) process.
3. Disambiguate tags identified both as geo and non-geographic entity, and extract the desired tags.

(1) This step retrieves the tags that are most probably part of the spatial coverage of the picture. We make the identification by matching the label of the tag with the name or alternate names of the geographic features that lie within a bounding box. This step uses the GeoNames API[47]. The picture location tag provides the bounding box as it contains the WOEID related to the picture and the precise coordinates.

Using the previously computed bounding box increases the correctness of the tags identification in case of homonyms. There is more probability that the term used to tag a picture is one referencing a close by geographic feature. For example for a picture located in Europe, the tag "Paris" has a greater probability to refer to the French capital than to the Texas town. The bounding box is extended to retrieve additional geo-tags that are not geographically close to the picture location. The extension is done following the WOEID hierarchy.

[45] https://www.wikidata.org

[46] https://developer.yahoo.com/geo/geoplanet /

[47] http://www.geonames.org/export/web-services.html

Increasing the bounding box usually identifies features related to the content of the resource and not its coverage. However, some distant features may appear in the picture if they are tall or massive or wide, such as mountains, lakes, *etc.* A feature is considered as potentially part of the coverage if the ratio *feature size/distance* is greater than a threshold (typically 0.04). Since the feature sizes are usually not known, the algorithm uses a rough estimate that depends on the feature class. If the geo feature was identified using one of the hierarchies' *locality*, *county*, or *region* and if its class, *e.g.* its type, is one of: ocean, mountain, desert, volcano, *etc.* then we consider it as part of coverage. The exact feature list from GeoNames we use is: 'FRST', 'VLC', 'PK', 'PKS', 'MTS', 'MT', 'ISLS', 'ISL', 'DSRT', 'SEA', 'OCN'. The algorithm also consider all the geo tags identified using the *neighbourhood* hierarchy as part of the coverage.

(2) A word sense disambiguation (WSD) algorithm such as [84]-[87], is applied to disambiguate the remaining tags. WSD is used to find the sense of a word within a given context. The system uses BabelFy API [82] as WSD tool. This tool takes as input a sentence in a given language or multilingual, and creates a graph-based semantic interpretation of the text by linking candidates meaning to fragments of the text. In BabelFy, each concept is given a semantic signature independently of the text. The semantic signature is defined as a set of related concepts. Then, when given a text, it extracts the linkable fragments and proposes the possible meanings according to the semantic network. It then selects the best candidate meaning for each fragment. The meanings are extracted from BabelNet that is both a multilingual encyclopaedic dictionary and a semantic network. BabelNet is created from the combination of multilingual web encyclopaedia with WordNet: Wikipedia, Wikidata, GeoNames, *etc.* It generates a labelled directed graph. Names or concepts refeered as nodes are connected using edges labelled with a semantic relation such as “part of” or “is a”. Each node has a set of translation for the concept it represents.

BabelNet receives as input the “sentence” formed by all the photo’s tags (even the previously identified geo tag) concatenated with the photo title and its description. Adding the photo’s description and title helps give a context to BabelFy.

The output of the algorithm is the association of a Wikidata synset (concept) to every recognised tag. This synset or named entity is then used to classify the tag in one of the categories described before (*Event, Temporal, Weather, Actors,* and *Colour*) or in none of them. Because we are working with pictures, we tended to retrieve a large number of “*transportation*” tags. Such tags do not describe place characteristics, but what appear in the picture, so to discard them we introduced this additional category.

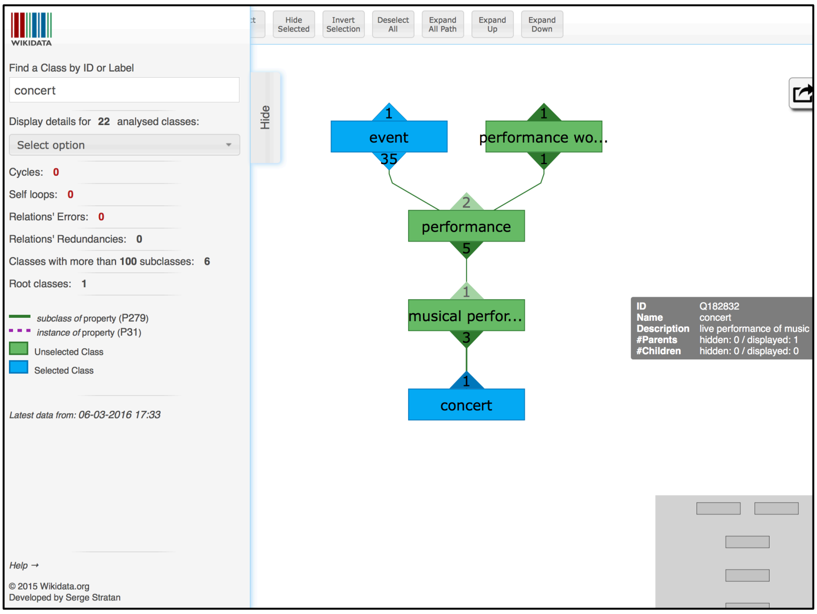
We have identified parent classes in Wikidata for each category:

- We associate the Weather category with the classes “weather” (Q11663) or “natural phenomenon” (Q1322005)
- We associate the Temporal category with the class “time interval” (Q186081)
- We associate the Event category with the classes “event” (Q1190554) or “point in time” (Q186408)
- We associate the Colour category with the class “color” (Q1075)
- We associate the Actor category with the classes “subject” (Q830077) or “organism” (Q7239)
- We associate the Transport category with the classes “transport” (Q7590) or “mode of transport” (Q334166)

Using the properties “subclass_of ”, “instance_of” or “part_of”, we search for the tag ancestors up to 5 levels. If the Wikidata id retrieved by BabelFy for the tag is children to one of the selected Wikidata class, then the tag can be categorised accordingly. In Figure 56, we can see the ancestor tree of the Wikidata concept “concert” issued from the “Taxonomy Browser” tool of Serge Stratan[48].

If the WSD algorithm has identified the tag, but none of its ancestors corresponds to the ones selected, then the tag is not categorised and is added to the “Semantic Enhancement of Places” (SEP) group. This group holds all the tags that can potentially provide the characteristics of the places that appear in the photo.

[48] http://sergestratan.bitbucket.org



Find a Class by ID or Label

concert

Display details for 22 analysed classes:

Select option

Cycles: 0

Self loops: 0

Relations' Errors: 0

Relations' Redundancies: 0

Classes with more than 100 subclasses: 6

Root classes: 1

Latest data from: 06-03-2016 17:33

Help →

© 2015 Wikidata.org
Developed by Serge Stratan

ID	Q182832
Name	concert
Description	live performance of music
#Parents	hidden: 0 / displayed: 1
#Children	hidden: 0 / displayed: 0

Figure 56 Wikidata "concert"ancestors

The WS process can also identify *Geo feature class* tags and *Geo feature* tags.

- We associate the Geo feature class category with the class “geo object” (Q618123) if found with a relation different than “instance_of”.
- We associate the Geo feature category with the classes “geo object” (Q618123), if found with the relation “instance_of”, or with the class “designation for an administrative territorial entity” (Q15617994).

This allows us to complete the identification done in the geo process. As the first process is based on a crowd-sourced resource, some feature might not be in its database. Using a second tool to identify geo tags help us fill the possible gaps in the identification process.

Tags that have not been identified by the WS process because, for example, there is no corresponding word in Wikidata can be classified two ways:

1. If there exist GeoNames entities related to these tags, they are classified as geographic tags. For those that are not identified with a GeoNames entity, they will not be assigned to a category and so will be added to the SEP group.
2. The tags that are not identified by either GeoNames or BabelFy and for which the system is not able to determine a language are assigned to the “*Unidentified*” category.

(3) The disambiguation of ambiguous tags is done comparing the geographic level and non-geographic weight. The geo-weight is the level of the bounding box that was used to identify the tag. The non-geographic weight is computed using the semantic measure library tool [88] of LGI2P laboratory from Ecole des mines d'Alès, using the pairwise measure SIM_PAIRWISE_DAG_NODE_LIN_1998 [89]. The non-geo weight is the smallest semantic distance value between the ambiguous tag and the other tags of the photo.

If tags have been identified as a geographic entity using a faraway bounding box like "*world*" or "*country*", during the geo process, and if they also are identified as non-geographic entities by the disambiguation step, with a weight greater than 0.15, then the non-geographic concept is kept as the tag disambiguation. On the other hand, if the geographic level is "*locality*", "*county*" or "*region*", and unless the non-geographic weight is high (greater than 0.8), we keep the geographic disambiguation. This is induced by the fact that many common words are also names of geographic entities in the world. For example "Pub" is the name of a mountain in Pakistan, but it also means a bar, a tavern, in English. The only exception is when we identify the tag as a hierarchy entity then we keep the geographic value.

Finally, we cleaned the database of imported pictures to remove duplicates. In our test corpora we had a large number of pictures (between 50 and 100 for each group) with the same title or the same list of tags, most probably from batch upload to Flickr. Those duplicates would have notably influenced our validation results, *i.e.* the precision and recall values. We consider duplicates the identical photos (same title and same number of tags) and similar pictures from the same author with the exact same list of tags.

5.4 Validation

The validation consists of selecting the tags that should be in the group "Semantic Enhancement of Places" and comparing the algorithm results with the validator's choices. This section represents the identified tags that belong to no categories and that can potentially bring information on the place characteristic. Five persons participated in the validation. We divided the corpus so at least two validators processed each photo. Figure 57 and Figure 58 show an example of photo and its tag validation view from our prototype. The "Geo Semantic" group of tags is the result of the algorithm, and host all the tags that contain the potential information on the place characteristics. In our example, as shown in Figure 59, we can deduce that the place in the photo "Street #9" (Figure 57) is a street, probably related to the army.

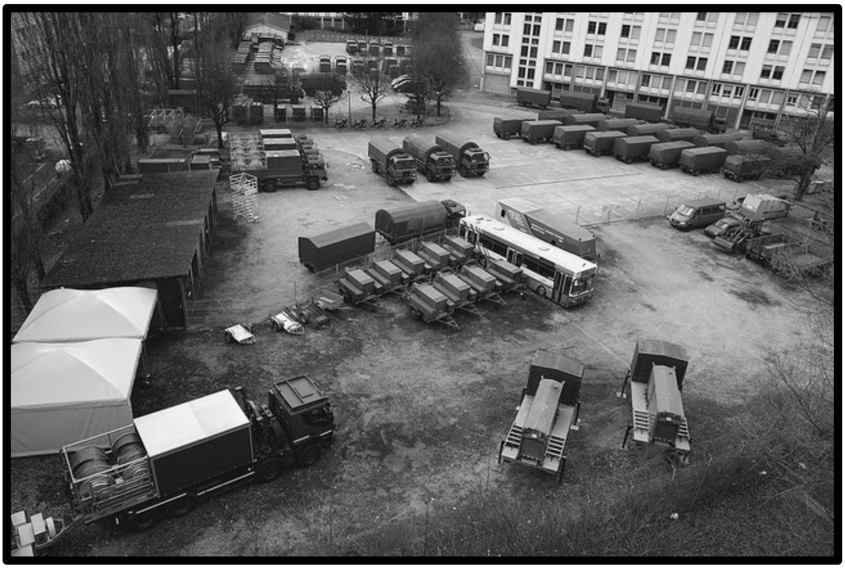
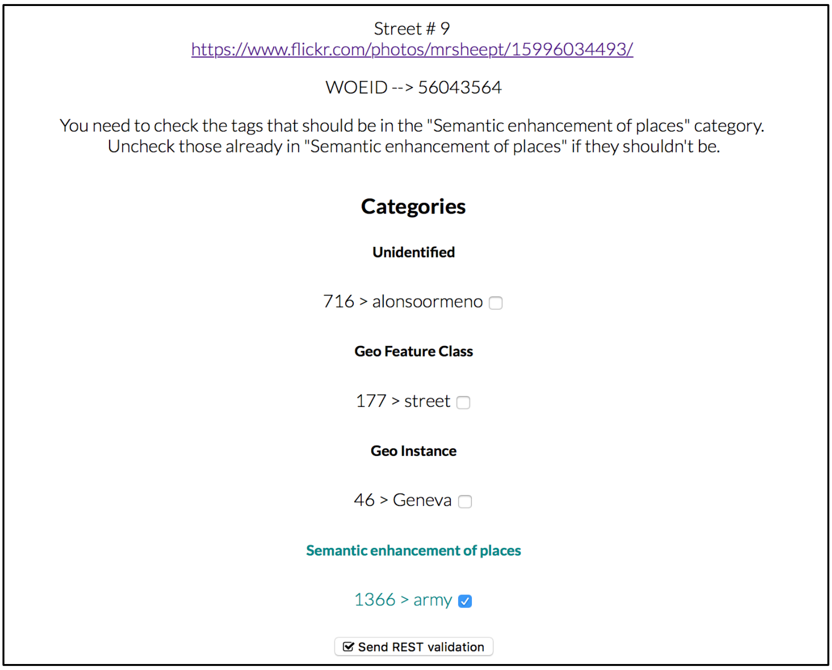


Figure 57 Photo "Street #9" from Alonso Ormeño



Street # 9
https://www.flickr.com/photos/mrsheept/15996034493/

WOEID --> 56043564

You need to check the tags that should be in the "Semantic enhancement of places" category.
Uncheck those already in "Semantic enhancement of places" if they shouldn't be.

Categories

Unidentified

716 > alonsoormeno □

Geo Feature Class

177 > street □

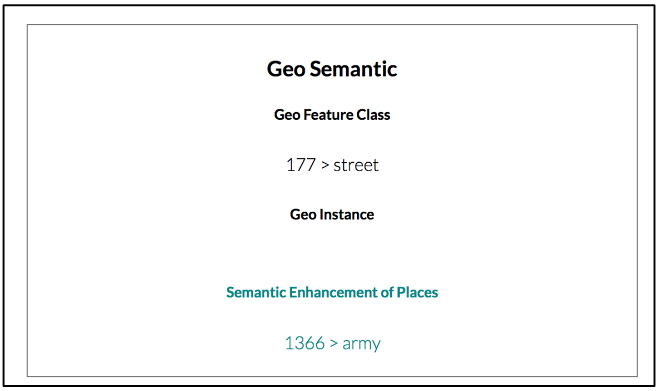
Geo Instance

46 > Geneva □

Semantic enhancement of places

1366 > army ☑

Figure 58 Validation view of photo tags



Geo Semantic

Geo Feature Class

177 > street

Geo Instance

Semantic Enhancement of Places

1366 > army

Figure 59 Prototype photo result view

Our corpus is composed of 400 documents, of which after validation, only 329 have at least one tag selected by the validators. Over the 400 photos, there is an average of 12 tags per photo, with minimum one tag and maximum 57, a median of 8 and a mode of 7. 37 photos have 3 or fewer tags, and 30 photos have more than 30 tags.

To calculate the precision and recall fore each photo, we process the number of true positives (TP) tags which are the tags that have been correctly placed in the "Semantic Enhancement of Places" by the algorithm and selected by the validator We also need the number of false positives (FP), tags placed in the section but that no validator selected, and false negatives (FN), tags selected by at least one of the validator and not categorised in the "Semantic Enhancement of Places" section.

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$F - measure = 2\times\frac{precision \times recall}{precision + recall}$$

The database has been populated using pictures from two WOEIDs:

782861: Plainpalais neighbourhood of Geneva, Switzerland

782017: Carouge city of Geneva canton, Switzerland

The global precision and recall values are computed using the multi-label precision and recall detailed in [90] n being the number of photos.

$$Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{|TPi|}{|TPi + FPi|} \qquad Recall = \frac{1}{n}\sum_{i=1}^{n}\frac{|TPi|}{|TPi + FNi|}$$

We also computed the inter-validator agreement using the Cohen's Kappa formula [91], which is specific for two validators. This Kappa uses the combination of the relative observed agreement among validators, and the hypothetical probability of chance agreement. It ranges from negative values, in rare cases, to 1, 1 being the total agreement between the validators. According to [92], the kappa value can be interpreted following this scale: < 0.20: Poor; 0.21 - 0.40: Fair; 0.41 - 0.60: Moderate; 0.61 - 0.80: Good; 0.81 - 1.00: Very good.

For all photos independently to their WOEID we found the following values:

- Precision = 72.5%
- Recall = 66.7%
- F-measure = 0.695

As a different pair of validators processed each WOEID set, we could not compute the global Cohen's Kappa value.

For each WOEID we found:

WOEID 782861

- Precision = 64.84%
- Recall = 60.68%
- F-measure = 0.764
- Cohen's Kappa= 0.77

WOEID 782017

- Precision = 80.38%
- Recall = 72.82%
- F-measure = 0.632
- Cohen's Kappa= 0.46

The precision and recall score is related to the subjective aspect of the task, where the two validators might not agree on the tags to select. For example, if a place function is to host tramways, should we select the tag with the name of the tram company to go in the "Semantic Enhancement of Places" group, as the tag "*tram*" will be in the transportation category?

The spelling of the tags also influences the pertinence of our algorithm. For example the term "transport public" is spelled three ways in our database: "transport-public", "transportpublic" and "Transports en commun". Only the last version was recognised by the algorithm, the other two were missing spaces or were using sign "–" instead.

The score is also influenced by the fact that the system is based on sources essentially filled via crowdsourcing. Those factors imply that there exist a few missing disambiguation or false evaluation. Not all places might be registered in GeoNames, or not every term might be referenced in Wikidata. BabelFy might also return a wrong result during the disambiguation process. For example for the tag #1002 "*Bridge*", BabelFy returned the concept number bn:00022229n, of the card game instead of the built structure.

We only implemented the semantic distance tool with the English WordNet dictionary, so we were not able to properly disambiguating tags in other languages. For example the tag #174 "*tramway*" was identified by both the geo and WS processes: the geonames instance 4495537, a populated place in North Carolina, USA, using the world bounding box and the BabelNet concept bn:00074613n, "streetcar" (Wikidata concept Q5641). Our language dictionary, to select the word language, categorised "tramway" as a French word so the algorithm did not use the semantic distance tool, as we did not implement it for the French language, and therefore did not assign a non-geo weight value. In result, the tag number 174 was identified as a geo tag for the photo #13253483064[49].

Finally, the parameters we choose to disambiguate the ambiguous tags are also influencing the score. To disambiguate non-geo tags, we empirically decided that the non-geo weight value must be at least 0.15 and the geo weight value to the world bounding box. But there still are some tags that get wrongly dispatched because of this value. For example the tag #1334 "*sunday*" for the photo #15330127248[50] was classified under geo tags due to its non-geo weight value of 0.1022.

We can see in our results that the inter-validator agreement values are different from one set to another. This can be explained by the fact that each set is different and so the difficulty of task was different. For each validator the interpretation of the validation can be different. For example in photo #8271078821[51] as shown in Figure 60, validators agreed on the following tags to be part of the "Semantic Enhancement of Places" group: "*building*", "*tours*", "*immeuble*". On the other hand, only one of the validator selected the tags "*window*", "*fenêtres*", "*balcon*". Should elements of the building in the picture be part of the "Semantic Enhancement of Places" group? Are they part of the building characteristics? Another example in photo #26321272793[52] as shown in Figure 61, both validators agreed on the tags "*museum*", "*exposition*", "*exhibition*", and "*art*", but only one of them selected "*artist*".

[49] https://www.flickr.com/photos/ludtz/13253483064/

[50] https://www.flickr.com/photos/128586514@N04/15330127248/

[51] https://www.flickr.com/photos/mr_brique/8271078821/

[52] https://www.flickr.com/photos/eric_g73/26321272793/

Figure 60 Photo "Où est-tu?" from Mr Brique

Figure 61 Photo "Linda Naeff exhibition @ Musée de Carouge" from Eric

5.5 Summary

Our evaluation presented in section 5.4 has shown the feasibility and validity of the methodology. However we have seen that the results are closely related to the completeness and accuracy of the crowdsourcing based tools used in the different identification process of the researched information. This method, as the precision and recall values show, is not more efficient than another but is sufficiently so for our validation. The validity is particularly evident for places with a small amount of data on which the statistical approaches are not applicable.

This prototype and its algorithm were built around our concept of geographic coverage. It actively differentiates geographic references that belong to the content and the ones that belong to the geographic coverage. We can say that it clearly illustrates the utility of geographic coverage at a conceptual level. A future possible improvement for this prototype is to implement a process to sort the found characteristics and features. The sorting process will associate each characteristic with the corresponding feature for a more precise result.

The source code of this prototype is available on GitLab at the following address: https://gitlab.com/CamilleTrd/GIEFT.

Chapter 6
Conclusion and Future Work

6.1 Contribution

In this thesis we have investigated the representation of the geo-spatial aspect of resources in a digital library by introducing a new digital library model with a definition of coverage as key concept. We have also presented the use of the model for the localisation of resources within a 3D model.

This library model is composed of a coverage model; a modular semantic annotation vocabulary that can be easily enhanced with domain ontologies; documents, which represent all the non-spatial resources hosted in the repository; and spatial resources, which are 3D city models and define the scope of the repository.

We have defined the generic coverage model and definition. We make an important distinction between the geographic and temporal content of a resource, *e.g.* mentioned place names or date, and its coverage. Accordingly, we defined the coverage as "*the geographic and temporal context of a resource refers to the spatial and temporal region in which the resource is true*". We also detailed each coverage definition applicable to each resource type (images, text, 3D models...), spatial or not, according to the specificities of the types.

The annotation vocabulary is centred on a spatio-geographic facet, and designed to be modular. Its core is composed of the alignment between generic spatial and geographic ontologies (CityGML and GeoNames) and a generic domain ontology (Urbamet). To enhance the annotation vocabulary core with new domain specific semantics, we align the generic domain ontology with the new specific domain vocabulary. Our core alignment is available online[53].

The document and spatial resource model were also defined along with the query model. The emphasis was made on the combination of the content and the coverage for each of those entities. To offer a complete information retrieval model, we propose the matching algorithm for the retrieval and ranking of documents according to a user query. The matching result is a combination of the query and document content matching with the query and document coverage matching.

[53] https://gitlab.com/CamilleTrd/GeoAnnotation-ontology

The indexing process is detailed in the implementation of the annotation model. It locates the documents within the 3D spatial resource. To match a document with the corresponding 3D objects, we apply the annotation algorithm to both the spatial and the documents. The spatial resource annotation consists in applying an identification process to each 3D object in the model. We propose an algorithm that identifies with geonames entities, CityGML objects. The documents annotation is the process of mainly identifying their coverage and expressing it using our semantic annotation vocabulary. As a result to the indexing process, we generate a matrix that references the relations and their weight between each document and each 3D object that from the spatial resource.

Another research question we approached in this work is the validation of the previously presented model, and the possible use of such models to semantically enhance geographic services or 3D city models.

We developed a categorisation system, with a non-statistical knowledge-based approach. This system aims at extracting geographic information from tags from VGI sources. This extracted information is then filtered and categorised to highlight the information that define characteristics of places. The characteristic of a place is a non-circumstantial information that qualifies a place, *i.e.* its function, its name, *etc.* The methodology used in the system is based on the coverage definition as previously explained. This system is suited to also work on places with a small amount of data where statistical approach will not work.

We divided the system into three processes: a geo process that analyses the tags with a geographic source and identifies the potential geographic tags; a word sense process to process the tags with a WSD tool [82] to identify probable geographic and non-geographic tags and categorise the non-geographic tags; a disambiguation process to clear double identification; and the extraction process to select and extract the tags that represent characteristics of places.

We implemented the system using Flickr tags, along with freely available services such as Geonames, Babelfy and Wikidata. We have validated our prototype results by running user validation of the categorisation. Validators were presented with the categorisation and selection result and were asked to validate the final selection of tags, *i.e.* those that represent or define characteristics of places. The result of our tests have shown the feasibility and validity of our process, even though it was clear that the results were partially dependant on the accuracy of the sources used in the processes.

The source code of our prototype is available online[54].

The following publications where published in relation to this thesis:

N. Ghoula, H. de Ribaupierre, C. Tardy, and G. Falquet, “Opérations sur des ressources hétérogènes dans un entrepôt de données à base d’ontologie,” presented at the 4e édition des journées francophones sur les ontologies (JFO), Montréal, Canada, 2011, pp. 203–216.

C. Tardy, L. Moccozet, and G. Falquet, “Semantic alignment of documents with 3D city models,” presented at the Usage, Usability, and Utility of 3D City Models (3U3D), 2012.

C. Tardy, L. Moccozet, and G. Falquet, “A Simple Tags Categorization Framework Using Spatial Coverage to Discover Geospatial Semantics,” presented at the Proceedings of the 25th International Conference Companion on World Wide Web, Montréal, Québec, Canada, 2016, pp. 657–660.

C. Tardy, G. Falquet, and L. Moccozet, “Semantic enrichment of places with VGI sources,” presented at the GIR'16 the 10th Workshop on Geographic Information Retrieval, Burlingame, California, USA, 2016.

6.2 Future Work

A future development for our geographic information extraction prototype is to enable the system, in the case of many places identified, to be able to associate the found characteristics with the corresponding places among the ones identified. We need to improve our results by finding better parameters for the prototype implementation and so reach a better Kappa value.

We also aim to apply this algorithm in a bigger context as a pre-process of data for statistical approaches. We think that the association of both techniques could bring good opportunities for new research directions and prospective questions in the domain of automatic extraction of geographic information.

The next stage for our repository model is its implementation, using the work described in this thesis. We have designed here the indexing, matching and querying algorithm. The visualisation and interface still needs to be addressed. How to design the querying interface within the 3D spatial resource?

The document-browsing environment will also be researched. How to translate visually the documents and 3D objects relations that we have identified in this research, in order to improve the reading and understanding of a document and facilitate decision-making task?

We also aim at identifying solutions to use our annotation model in order for users with different expertise and vocabulary, to exchange annotations and information on documents hosted in the digital library.

[54] https://gitlab.com/CamilleTrd/GIEFT

Annexe: Techniques and tools mentioned in this thesis

Bellow a table that references and describe the different technics, tools and resources that are mentioned and used in this work. The terms are issued from Chapter 1, Chapter 2, Chapter 4 and Chapter 5. The definitions are mainly extracted from the corresponding Wikipedia pages.

Term	Type	Definition
Named Entity Recognition (NER)	Technique	"Identifying references to information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions in text. It is recognized as one of the important sub-tasks of information extraction." -- Reference [10]
Part-of-Speech (POS)	Technique	Identifies similar grammatical and syntax entities such as noun, verb, adjective, adverb, pronoun, *etc.* in text. -- Reference [10]
Discounted Cumulated Gain (DCG)	Technique	Used in information retrieval to measure ranking quality. DCG measures the usefulness, or gain, of a document based on its position in the result list. It discounts the value of documents ranked further down in the result list. -- Reference [57]
Cumulated Gain (CG)	Technique	CG is the predecessor of DCG and does not include the position of a result in the consideration of the usefulness of a result set. CG is the sum of the graded relevance values of all results in a search result list. -- Reference [57]
Word Sense Disambiguation	Technique	A computational linguistic problem that consist in identifying which sense

(WSD)		of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings.
Geonames ontology	Ontology	Each Geonames toponyms has been assigned a unique identifier. The Geonames ontology represents the relations between those toponyms. Geonames proposes an RDF web service. -- Reference [60] and Note 27.
CityGML	Ontology	The City Geography Markup Language (CityGML) is a common information model for the representation of 3D urban objects. It defines the classes and relations for the most relevant topographic objects in cities and regional models with respect to their geometrical, topological, semantical and appearance properties. -- Reference [61]
OWL time	Ontology	It is an ontology of temporal concepts, for describing the temporal properties of resources in the world or described in Web pages. -- Reference [63]
WordNet	Thesaurus	A large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. -- Reference [45]
Gemet	Thesaurus	General Multilingual Environmental Thesaurus is developed by the European environment agency. It compiles 22 languages. -- Note 29
Urbamet	Thesaurus	A thesaurus that organizes the concepts form the urbamet database around the following thematics: urbanism, housing, building, architecture and equipment. -- Note 26

DBpedia	Database	An extract structured content from the information created in the Wikipedia project. It allows users to semantically query relationships and properties of Wikipedia resources, including links to other related datasets. -- Note 44
Wikidata	Database	A document-oriented database, focused on items. Each item represents a topic. It is hosted by the Wikimedia Foundation. -- Note 45
Yahoo! GeoPlanet	Tool	It coordinates world-wide geographic information, and provides both text and cartographic output, such as digital maps for any location in the world. An integral part of GeoPlanet is the WOEID that identifies any feature on Earth. -- Note 46
Volunteered Geographic Information (VGI)	Type	It is the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals. It is a special case of user-generated content. OpenStreetMap is an example of a result of VGI.
3D City Models	Type	Digital models of urban areas that represent terrain surfaces, sites, buildings, [...] as well as related objects (e.g., city furniture) belonging to urban areas. Their components are described and represented by corresponding 2D and 3D spatial data and geo-referenced data.
OpenStreetMap	Tool	A collaborative project to create a free editable map of the world. OSM is composed of open and crowd sourced data. -- Note 17
Lemur Project	Tool	It develops search engines, browser toolbars, text analysis tools, and data resources that support research and

		development of information retrieval and text mining software. -- Note 16
BabelFy	Tool	A software algorithm for the disambiguation of text written in any language. Specifically, Babelfy performs the tasks of WSD and entity linking. It is based on Babelnet semantic network to perform disambiguation. -- Reference [82]
GATE	Tool	An open source software for language processing tasks and applications. -- Reference [22]
Dublin Core	Metadata	A small set of vocabulary terms that can be used to describe digital resources and physical resources. -- Reference [43]
Adobe XMP	Metadata	An ISO standard created by Adobe. It is a data model, a serialization format and core properties for the definition and processing of extensible metadata for digital documents and data sets.
Exif	Metadata	A metadata standard for images, sound, digital camera, scanners, *etc.* -- Reference [64]
IPTC	Metadata	A metadata standard for news media (text, images, or other media). It was created to facilitate the international exchange of information. It is broadly used by photographer. -- Reference [65]
ICBM	Metadata	Intercontinental Ballistic Missile address or missile address. A metadata to exchange latitude and longitude coordinates in webpages using meta tags.
xAL (eXtensible Address Language)	Format	Part of the xNAL standard for name and address, xAL is the address standard. It differs from other address standard as it concentrates in Postal Services and other services handling

		names and addresses. xNAL is designed to handle the address structures of all countries at an abstract or detailed level.
RDF	Format	A standard model for data interchange on the Web. It was originally designed as a metadata model it is now a general method for conceptual description or modeling of information that is implemented in web resources. It uses triplets of the form subject–predicate–object.

References

[1] L. Candela, D. Castelli, P. Pagano, C. Thano, Y. Ioannidis, G. Koutrika, S. Ross, H.-J. Schek, and H. Schuldt, “Setting the Foundations of Digital Libraries: The DELOS Manifesto,” *D-Lib Magazine*, vol. 13, no. 3, p. 4, 2007.

[2] B. Haslhofer and P. Kneževié, “The BRICKS Digital Library Infrastructure,” in *Semantic Digital Libraries*, no. 11, S. R. Kruk and B. McDaniel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 151–161.

[3] M. A. Goncalves, “Streams, Structures, Spaces, Scenarios, and Societies (5S: A Formal Digital Library Framework and Its Applications,” scholar.lib.vt.edu, 2004.

[4] R. Ivanov and L. Raae, “INSPIRE: a new scientific information system for HEP,” pp. 1–8, Jun. 2009.

[5] A. Katifori, C. Nikolaou, M. Platakis, Y. Ioannidis, A. Tympas, M. Koubarakis, N. Sarris, V. Tountopoulos, E. Tzoannos, S. Bykau, N. Kiyavitskaya, C. Tsinaraki, and Y. Velegrakis, “The Papyrus Digital Library: Discovering History in the News,” in *Research and Advanced Technology for Digital Libraries. Proceedings of the International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26-28, 2011.*, vol. 6966, S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, Eds. Springer Berlin / Heidelberg, 2011, pp. 465–468.

[6] V. Albino, U. Berardi, and R. M. Dangelico, “Smart cities: Definitions, dimensions, performance, and initiatives,” *Journal of Urban Technology*, 2015.

[7] R. Sieber, “Public Participation Geographic Information Systems: A Literature Review and Framework,” *Annals of the Association of American Geographers*, vol. 96, no. 3, pp. 491–507, Sep. 2006.

[8] C. B. Jones and R. S. Purves, “Geographical information retrieval,” *International Journal of Geographical Information Science*, vol. 22, no. 3, pp. 219–228, Mar. 2008.

[9] J. Ding, G. Luis, and S. Narayanan, “Computing geographical scopes of web resources,” presented at the Proceedings of the 26th International Conference on Very Large Data Bases VLDB '00, 2000, pp. 545–556.

[10] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, Jan. 2007.

[11] C. Jones, A. Abdelmoty, D. Finch, G. Fu, and S. Vaid, “The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing,” presented at the Proceedings of the 3d International Conference in Geographic Information Science, 2004, vol. 3234, pp. 125–139.

[12] R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang, “The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet,” *International Journal of Geographical Information Science*, vol. 21, no. 7, pp. 717–745, Aug. 2007.

[13] A. Woodruff and C. Plaunt, “GIPSY: Automated Geographic Indexing of Text Documents,” *JASIS*, pp. 1–21, 1994.

[14] F. Bilhaut, T. Charnois, P. Enjalbert, and Y. Mathet, “Geographic reference analysis for geographic document querying,” presented at the HLT-NAACL-GEOREF '03: Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references, 2003, vol. 1, pp. 1–8.

[15] F. Bilhaut and A. Widlöcher, “LinguaStream: an integrated environment for computational linguistics experimentation,” presented at the EACL '06: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations, 2006.

[16] Y.-F. R. Chen, G. Di Fabbrizio, D. Gibbon, S. Jora, B. Renger, and Bin Wei, “Geotracker: geospatial and temporal RSS navigation,” presented at the WWW '07: Proceedings of the

16th international conference on World Wide Web, 2007.

[17] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling, “STEWARD: architecture of a spatio-textual search engine,” p. 25, 2007.

[18] M. Gaio, C. Sallaberry, P. Etcheverry, C. Marquesuzaa, and J. Lesbegueries, “A global process to access documents’ contents from a geographical point of view,” *sciencedirect.com*, vol. 19, no. 1, pp. 3–23, Feb. 2008.

[19] D. Buscaldi and P. Rosso, “Geooreka: Enhancing Web Searches with Geographical Information.,” vol. 9, pp. 205–212, 2009.

[20] J. Levin and B. Nalebuff, “An introduction to vote-counting schemes,” *The Journal of Economic Perspectives*, 1995.

[21] M. Á. García-Cumbreras and J. M. Perea-Ortega, “Information retrieval with geographical references. Relevant documents filtering vs. query expansion,” *Information Processing …*, 2009.

[22] Cunningham, *Developing Language Processing Components with GATE Version 8*. University of Sheffield Department of Computer Science, 2014.

[23] R. Neches, K.-T. Yao, I.-Y. Ko, A. Bugacov, V. Kumar, and R. Eleish, “GeoWorlds: integrating GIS and digital libraries for situation understanding and management,” *New Review of Hypermedia and Multimedia*, vol. 7, no. 1, pp. 127–152, Jan. 2001.

[24] M. Doerk, S. Carpendale, C. Collins, and C. Williamson, “VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery,” *Ieee T Vis Comput Gr*, vol. 14, no. 6, pp. 1205–1212, 2008.

[25] S. Ahern, M. Naaman, R. Nair, and J. Yang, “World explorer: visualizing aggregate data from unstructured text in geo-referenced collections,” *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 1–10, 2007.

[26] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne, “Okapi at TREC-4,” presented at the Proc. of the 4th Text REtrieval Conference (TREC- 4), 1995, pp. 73–96.

[27] M. Chalmers, “Using a landscape metaphor to represent a corpus of documents,” *European Conference on Spatial Information Theory*, 1993.

[28] K. Andrews and M. Pichler, “Hooking up 3-space: three-dimensional models as fully-fledged hypermedia documents,” *Multimedia, Hypermedia, and Virtual Reality Models, Systems, and Applications*, pp. 28–44, 1996.

[29] S. Havemann, V. Settgast, R. Berndt, Ø. Eide, and D. W. Fellner, “The arrigo showcase reloaded—towards a sustainable link between 3D and semantics,” *J. Comput. Cult. Herit.*, vol. 2, no. 1, pp. 1–13, Jul. 2009.

[30] K. Grønbæk, P. P. Vestergaard, and P. Ørbæk, “Towards geo-spatial hypermedia: Concepts and prototype implementation,” pp. 117–126, 2002.

[31] S. M. An, H.-Y. Lee, B. Kim, C.-Y. Yi, J.-H. Eum, and J.-H. Woo, “Geospatial spreadsheets with microscale air quality visualization and synchronization for supporting multiple-scenario visual collaboration,” *International Journal of Geographical Information Science*, Jun. 2014.

[32] J. Stoter, H. de Kluijver, and V. Kurakula, “3D noise mapping in urban areas,” *International Journal of Geographical Information Science*, vol. 22, no. 8, pp. 907–924, Aug. 2008.

[33] C. Metral, N. Ghoula, and G. Falquet, “Towards an Integrated Visualization Of Semantically Enriched 3D City Models: An Ontology of 3D Visualization Techniques,” presented at the Workshop of TU0801 Cost Action, Madrid, 2012.

[34] T. H. Kolbe, “Representing and exchanging 3D city models with CityGML,” *3D Geo-Information Sciences*, 2009.

[35] M. Goetz, “Towards generating highly detailed 3D CityGML models from OpenStreetMap,” *International Journal of Geographical Information Science*, vol. 27, no. 5, pp. 845–865, May 2013.

[36] M. Goetz and A. Zipf, “Extending OpenStreetMap to indoor environments: bringing volunteered geographic information to the next level,” *Urban and regional data*

management: UDMS Annual 2011, 2011.

[37] M. Goetz, “Proposed features/IndoorOSM - OpenStreetMap Wiki,” *wiki.openstreetmap.org*. [Online]. Available: http://wiki.openstreetmap.org/wiki/Proposed_features/IndoorOSM. [Accessed: 08-Dec-2016].

[38] M. Goetz and A. Zipf, “Towards Defining a Framework for the Automatic Derivation of 3D CityGML Models from Volunteered Geographic Information,” *International Journal of 3-D Information Modeling (IJ3DIM)*, pp. 1–16, 2012.

[39] L. Candela, D. Castelli, N. Ferro, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, M. Dobreva, V. Katifori, and H. Schuldt, “The DELOS Digital Library Reference model. Foundations for digital Libraries,” Nov. 2007.

[40] L. Candela, G. Athanasopoulos, D. Castelli, K. El Raheb, P. Innocenti, Y. Ioannidis, A. Katifori, A. Nika, G. Vullo, and S. Ross, “The Digital Library Reference Model ,” D3.2b, Apr. 2011.

[41] M. A. Goncalves, E. A. Fox, L. T. Watson, and N. Kipp, “Streams, structures, spaces, scenarios, societies (5S): a formal model for digital libraries,” *Acm Transaction on Information Systems*, vol. 22, no. 2, pp. 270–312, 2004.

[42] R. Shen, “Applying the 5S Framework To Integrating Digital Libraries,” 2006.

[43] S. L. Weibel and T. Koch, “The Dublin Core Metadata Initiative,” *D-Lib Magazine*, vol. 6, no. 12, Dec. 2000.

[44] D. L. McGuinness and F. van Harmelen, Eds., “OWL Web Ontology Language Overview,” *w3.org*. [Online]. Available: http://www.w3.org/TR/owl-features/. [Accessed: 23-Jan-2012].

[45] G. A. Miller, *WordNet: a lexical database for English*, vol. 38. Communications of the ACM, 1995, pp. 39–41.

[46] M. Dewey, *Decimal Classification and Relative Index for Libraries*. 1891.

[47] S. Gramley and M. Pátzold, *A survey of modern English*. 2004.

[48] A. Schwering, “Semantic similarity measurement including spatial relations for semantic information retrieval of geo-spatial data,” Verlag Natur & Wissenschaft, 2007.

[49] A. Schwering, “Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey,” *Transactions in GIS*, vol. 12, no. 1, pp. 5–29.

[50] J. Ge and Y. Qiu, “Concept similarity matching based on semantic distance,” presented at the Semantics, Knowledge and Grid, 2008. SKG'08. Fourth International Conference on, 2008, pp. 380–383.

[51] M. A. Rodriguez and M. J. Egenhofer, “Determining Semantic Similarity among Entity Classes from Different Ontologies,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, pp. 442–456, 2003.

[52] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Inform Process Manag*, vol. 24, no. 5, pp. 513–523, Jan. 1988.

[53] A. Markowetz, Y. Y. Chen, T. Suel, X. Long, and B. Seeger, “Design and Implementation of a Geographic Search Engine,” presented at the 8th Int. Workshop on the Web and Databases (WebDB), Baltimore, Maryland, 2005.

[54] B. Martins, M. J. Silva, and L. Andrade, “Indexing and ranking in Geo-IR systems,” presented at the the 2005 workshop, Bremen, Germany, 2005, pp. 31–34.

[55] A. Passant, “Measuring semantic distance on linking data and using it for resources recommendations,” presented at the Proceedings of the AAAI Spring Symposium” Linked Data Meets Artificial Intelligence, 2010, vol. 3.

[56] K. Janowicz, M. Raubal, and W. Kuhn, “The semantics of similarity in geographic information retrieval,” *JOSIS*, no. 2, pp. 29–57, May 2011.

[57] R. B. Yates and B. R. Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*. Addison-Wesley Professional, 2011.

[58] K. Järvelin and J. Kekäläinen, “IR evaluation methods for retrieving highly relevant documents,” presented at the SIGIR '00 Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece,

2000, pp. 41–48.

[59] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of IR techniques,” *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002.

[60] B. Vatant and M. Wick, “GeoNames Ontology - Geo Semantic Web,” *geonames.org*. [Online]. Available: http://www.geonames.org/ontology/documentation.html. [Accessed: 16-Jun-2012].

[61] T. H. Kolbe, G. Gröger, and L. Plümer, “CityGML–Interoperable access to 3D city models,” *First International Symposium on Geo-Information for Disaster Management*, 2005.

[62] G. Groger, T. H. Kolbe, A. Czerwinski, and C. Nagel, Eds., “OpenGIS® City Geography Markup Language (CityGML) Encoding Standard,” OGC 08-007r1.

[63] J. R. Hobbs and F. Pan, “Time Ontology in OWL,” *w3.org*. [Online]. Available: http://www.w3.org/TR/owl-time/. [Accessed: 05-Feb-2013].

[64] “CIPA DC- 008-Translation- 2012 ,” Camera & Imaging Products Association, CIPA DC-008-2012, Dec. 2012.

[65] “IPTC : Photo Metadata.” Core 1.2/Extension 1.3 (Oct 2016)

[66] A. Daviel and F. A. Kägi, “Geographic Registration of HTML Documents,” *Internet Draft draft-daviel-html-geo-tag-08*.

[67] Z. Liu, “A Survey on Social Image Mining,” in *Intelligent Computing and Information Science*, no. 134, R. Chen, Ed. Springer Berlin Heidelberg, 2011, pp. 662–667.

[68] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua, “Research and applications on georeferenced multimedia: a survey,” *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 77–98, 2010.

[69] M. Sester, J. J. Arsanjani, R. Klammer, D. Burghardt, and J.-H. Haunert, “Integrating and Generalising Volunteered Geographic Information,” in *Abstracting Geographic Information in a Data Rich World*, D. Burghardt, C. Duchêne, and W. Mackaness, Eds. Springer International Publishing, 2014, pp. 119–155.

[70] R. Abbasi, “Discovering and Exploiting Semantics in Folksonomies,” 2011.

[71] L. Li and M. F. Goodchild, “Constructing Places from Spatial Footprints,” New York, NY, USA, 2012, pp. 15–21.

[72] C.-A. Papadopoulou and M. Giaoutzi, “Crowdsourcing as a Tool for Knowledge Acquisition in Spatial Planning,” *Future Internet*, vol. 6, no. 1, pp. 109–125, 2014.

[73] T. Rattenbury and M. Naaman, “Methods for Extracting Place Semantics from Flickr Tags,” *ACM Trans. Web*, vol. 3, no. 1, pp. 1–30, Jan. 2009.

[74] T. Rattenbury, N. Good, and M. Naaman, “Towards Automatic Extraction of Event and Place Semantics from Flickr Tags,” presented at the the 30th annual international ACM SIGIR conference, New York, NY, USA, 2007, pp. 103–110.

[75] S. Sizov, “Latent Geospatial Semantics of Social Media,” *ACM TIST*, vol. 3, no. 4, pp. 1–20, Sep. 2012.

[76] V. Antoniou, J. Morley, and M. Haklay, “Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon,” *Geomatica*, vol. 64, no. 1, pp. 99–110, 2010.

[77] R. Feick and C. Robertson, “A multi-scale approach to exploring urban places in geotagged photographs,” *Computers, Environment and Urban Systems*, vol. 53, pp. 96–109, Sep. 2015.

[78] R. Purves, A. Edwardes, and J. Wood, “Describing place through user generated content,” presented at the In Proceedings of the workshop on Metadata Mining for Image Understanding (MMIU 2008), Sheffield, 2011, vol. 16, no. 9.

[79] R. S. Purves, A. Edwardes, and M. Sanderson, “Describing the where – improving image annotation and search through geography,” http://eprints.whiterose.ac.uk/4566/, 2008.

[80] “Map Features - OpenStreetMap Wiki,” *wiki.openstreetmap.org*. [Online]. Available: http://wiki.openstreetmap.org/wiki/Map_Features. [Accessed: 28-Dec-2015].

[81] “GeoNames Feature codes,” *geonames.org*. [Online]. Available: http://www.geonames.org/export/codes.html. [Accessed: 28-Dec-2015].

[82] A. Moro, A. Raganato, and R. Navigli, “Entity linking meets word sense disambiguation: a

unified approach," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 2, pp. 231–244, 2014.

[83] G. Groger, T. H. Kolbe, C. Nagel, and K.-H. Hafele, Eds., "OGC City Geography Markup Language (CityGML) Encoding Standard ," OGC 12-019, Apr. 2012.

[84] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet:: Similarity: measuring the relatedness of concepts," *Demonstration papers at HLT-NAACL*, 2004.

[85] O. Medelyan, D. Milne, C. Legg, and I. H. Witten, "Mining meaning from Wikipedia," *International Journal of Human-Computer Studies*, vol. 67, no. 9, pp. 716–754, Sep. 2009.

[86] R. Mihalcea, "Using Wikipedia for Automatic Word Sense Disambiguation.," *HLT-NAACL*, 2007.

[87] R. C. Bunescu and M. Pasca, "Using Encyclopedic Knowledge for Named entity Disambiguation.," *EACL*, 2006.

[88] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies," *Bioinformatics*, vol. 30, no. 5, pp. 740–742, Feb. 2014.

[89] D. Lin, "An Information-Theoretic Definition of Similarity," presented at the 15th International Conference on Machine Learning, 1998, pp. 296–304.

[90] M. S. Sorower, "A literature survey on algorithms for multi-label learning," *Oregon State University*, 2010.

[91] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, 1960.

[92] D. G. Altman, *Practical statistics for medical research*. 1990.