

PhD Dissertation

**Quality Assessment for
Semantically Close Geographical
Properties**

Walter Renteria-Agualmpia

January 2015

PhD Advisors:

Dr. Francisco J. López Pellicer

Dr. F. Javier Zarazaga Soria



**Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza**

© Copyright by Walter Renteria-Agualimpia, 2015
All Rights Reserved

*To God, my mother Rumilda, Wanner, Deicy,
Rosa, Isaac, Paulina, and my beloved wife Yuli*

Acknowledgements

I would like to thank the members and ex-members of the *Advanced Information Systems Laboratory* (IAAA) of the Computer Science and System Engineering Department of the University of Zaragoza, and its spin-off *GeoSpatiumLab* for their support in many aspects, and especially, my research advisors, F.Javier Zarazaga-Soria and Francisco J.López-Pellicer for encouraging my research and for allowing me to grow as a researcher. I would like to express my thanks to Dr. Pedro Muro, Javier Nogueras, Javier Lacasta, Ruben Béjar. and Miguel Latre. I should also include Aneta, Jesús, Covadonga, Maria J., Rodolfo, David, Alberto, Juan, Pedro, Javier, Ivan, José, Miguel, Naveen, Rocío, Borja.

Also, I would like to express my gratitude to Professor Mauro Gaio and Ludovic Moncla of the University of Pau, France, and every other person I have met there, who made every effort to make my research stay a wonderful experience, both in the professional and in the personal development. I would like to thank the people that have reviewed this thesis. Despite all of their help, I take full responsibility for any errors or omission herein.

A special thanks to my family for their unconditional support during my dreams, travels and my work on this thesis. Words cannot express how grateful I am to my mother, and grandmother for all of the sacrifices that you've made on my behalf. Your constant prayers for me were what sustained me thus far. I feel specially grateful to Pastor Néstor Suárez, his wife and the United Pentecostal Church International in Zaragoza for their help in all these last years. At the end I would like express appreciation to my beloved wife Yuli for much patience, love and support in many ways.

Resumen ejecutivo

Antecedentes. Los avances en Tecnologías de la Información centrados en la información que describe fenómenos geoespaciales han revolucionado las actividades que manejan información en los campos de la investigación y la industria. El acceso efectivo a la información geoespacial adquiere una importancia fundamental en estos contextos basados en el conocimiento. Sin embargo, el creciente volumen de datos geoespaciales diariamente hace que la búsqueda directa sea inviable. Como alternativa, muchos sistemas de información buscan a través de metadatos geoespaciales, es decir, los datos que describen los datos geoespaciales. Desde los sistemas de emergencia, rescate y localización hasta los sistemas geopolíticos, militares e industriales que usan sistemas de toma de decisiones basados en información geográfica, es esencial tener acceso a los recursos geoespaciales a través de descripciones de metadatos consistentes y con un nivel de calidad mínimo para asegurar la recuperabilidad de los recursos (Hartmann and Stuckenschmidt, 2002; Martins et al., 2007). Por otra parte, para algunos autores la calidad, la consistencia, de una descripción de metadatos podría ser la diferencia entre la vida y la muerte o entre el éxito y el fracaso (Dushay and Hillmann, 2003; Bruce and Hillmann, 2004; Hillmann et al., 2004).

Esta tesis investiga cómo asesorar la calidad de un tipo particular de metadatos: los metadatos que describen la ubicación espacial de un recurso. En particular, esta tesis investiga cómo abordar problemas que pueden surgir cuando un registro de metadatos que describe algún recurso tiene más de una propiedad que tiene la intención de describir la ubicación del recurso, es decir, el registro de metadatos que contiene propiedades geográ-

ficas semánticamente cercanos (Miles and Bechhofer, 2009). Este problema está estrechamente asociado a la facilidad con la que los recursos georeferenciados pueden ser encontrados en un sistema de información debido a que las propiedades de metadatos son clave para descubrir, acceder y recuperar los recursos en muchos sistemas basados en catálogos indexados (Goodchild and Zhou, 2003; Hill, 2006)

Objetivo. Los problemas derivados de la inconsistencia de los metadatos geoespaciales pueden ser mitigados después de realizar una evaluación de la calidad (QA) de la descripción geoespacial. Esta tesis analiza un enfoque semi-automático para detectar inconsistencias geoespaciales y sugerir posibles soluciones para una inconsistencia basado en un contexto geoespacial construido a partir del consenso de las descripciones geoespaciales que rodean el recurso inconsistente. Además, la solución analizada proporcionaría un asesoramiento para llevar a cabo procesos de calidad, tales como, curación y preservación de material cartográfico en el contexto de los repositorios digitales.

Ámbito. El análisis QA se limita a dos tipos de recursos: (1) los metadatos de servicios Web en conformidad con las especificaciones del Open Geospatial Consortium (OGC) y (2) los metadatos MARC21 ¹ que describen recursos cartográficos. El OGC lidera desde 1994 el desarrollo de las especificaciones de interfaz de servicios Web abiertos y estandarizados para el acceso a la información geoespacial. Muchas empresas, agencias gubernamentales y universidades son miembros de OGC, y participan en los consensos para desarrollar normas de interfaz de servicios Web públicamente disponibles para el acceso a la recursos geoespaciales. Muchas especificaciones de interfaces de servicios Web OGC se han convertido en estándares de la Organización Internacional para la Normalización. En relación con los estándares OGC, esta tesis se enfoca en los estándares OGC más utilizados en Infraestructuras de Datos Espaciales (SDI) (Nebert, 2004). En particular, aplicamos el análisis de QA en los metadatos de servicios Web de mapas almacenados en Servicios de Catálogos OGC (CSW,(Nebert et al., 2007)). Sin embargo, nuestro sistema pueden aplicarse a otro tipo de servicios Web y recursos SDI con propiedades geográficas semánticamente cercanas. A pesar de que nuestra

¹<http://www.loc.gov/marc/bibliographic/>

metodología podría ser utilizada para asesorar la calidad de las propiedades temáticas y temporales, vamos a considerar únicamente las propiedades geográficas semánticamente cercanos.

Por otro lado, los estándares MARC son un conjunto de formatos digitales para la descripción de los elementos catalogados por las bibliotecas, tal como los libros. Este estándar fue desarrollado durante la década de 1960 por la Biblioteca del Congreso Americano para crear registros que pueden ser utilizados por las computadoras, fue ideado para compartir archivos entre bibliotecas. Para 1971, los formatos MARC se habían convertido en el estándar nacional para la difusión de los datos bibliográficos en los Estados Unidos, y la norma internacional en 1973. Existen varias versiones de MARC en uso en todo el mundo, el más predominante es MARC21, creado en 1999 como un resultado de la armonización de los formatos de los Estados Unidos, formatos MARC canadienses y UNIMARC ampliamente utilizado en Europa. Además, en muchas bibliotecas de todo el mundo, los metadatos MARC21 son el estándar más utilizado para documentar los recursos que describen fenómenos geográficos sobre la superficie de la tierra (Furrie, 2009). En los metadatos MARC21 hay varios campos diferentes que pueden codificar diferentes aspectos de referencias espaciales directas/indirectas, incluyendo diferentes formas de asociar los códigos geográficos, o diferentes formas para expresar el método de referencia geoespacial utilizada para las coordenadas en las referencias espaciales directas. Sólo tendremos en cuenta las dos propiedades geográficas semánticamente cercanas más frecuentes que encontramos analizando los conjuntos de datos experimentales: las referencias espaciales directas (extensión geográfica/huella espacial) y las referencias espaciales indirectas (topónimos) (FGDC, 1998b). En esta tesis no tomamos en cuenta otras propiedades geográficas.

El escenario de las SDI se utiliza como caso de prueba para validar toda la arquitectura propuesta. La información descriptiva de los recursos espaciales en SDI está estructurada y procede de los expertos del dominio geográfico. Esto nos hace pensar que la descripción espacial proporcionada en el contexto de las SDI debe ser mejor que otros escenarios y contextos donde las descripciones proceden de información no estructurada o documentada por personal no experto en el dominio geográfico, por ejemplo, contextos como el de las Bibliotecas Digitales. En este sentido, la metodología y la archi-

tectura se han probado con un caso más difícil, los materiales cartográficos. Nuestro análisis se ha restringido a dos áreas geográficas, España y los Estados Unidos de América. Esta restricción garantiza que las contribuciones de esta tesis pueden beneficiar a dos de los escenarios más frecuentes que utilizan metadatos geográficos (Infraestructuras de Datos Espaciales y Bibliotecas Digitales).

En cuanto al ámbito geográfico del Sistema de Organización del Conocimiento (KOS) (Hodge, 2000; Miles and Bechhofer, 2009) utilizado en el proceso de ranking espacial, este trabajo ha analizado varios sistemas de representación del conocimiento. Estos sistemas han sido principalmente SKOS simples (Simple Knowledge Organisation System (Isaac and Summers, 2009)) vocabularios y grafos RDF/XML. Una clara restricción consiste en que el KOS utilizado en las tareas de ranking espacial debe cubrir la extensión geográfica de los recursos geográficos evaluados. Además, el nivel de granularidad de las huellas espaciales en el KOS debe estar de acuerdo con los de la colección analizada. En esta tesis nos limitamos a trabajar con KOS cuyas huellas son de dos dimensiones (2D). Sin embargo, por medio de procesos simples las huellas en 2D se pueden simplificar a los clásicos puntos (1D). La línea de investigación abierta y el reto es pasar de análisis basados en huellas geográficas en 1D a 2D para asesorar su calidad.

Método. El enfoque metodológico comprende aspectos relacionados con la ingeniería de software, ingeniería del conocimiento y la inteligencia artificial. La metodología de la ingeniería de software es un clásico desarrollo incremental de la solución (Boehm, 1988; Larman and Basili, 2003). La metodología de la ingeniería del conocimiento se basa en las medidas propuestas por la plataforma Methontology (especificación, conceptualización, formalización, integración, implementación y mantenimiento) (Fernández-López et al., 1997).

Experiencia previa, trabajo futuro. Garantizar el acceso, recuperación y visualización de los recursos en el contexto de los sistemas de información distribuidos e interoperables son objetivos comunes y prioritarios para muchos dominios, por ejemplo, SDI y Bibliotecas Digitales. Uno de los casos más consolidados es la iniciativa europea INSPIRE, cuyo objetivo es la creación de una SDI Europea. Una de las líneas de investigación del grupo

de investigación IAAA ², en la cual el autor participa como investigador, se centra en aspectos de SDI relacionados con la descripción de los datos y los servicios, el descubrimiento de estos recursos geoespaciales a través de catálogos, y los aspectos conceptuales y arquitectónicos relacionados con los datos y servicios geoespaciales. Algunos resultados de la línea de investigación en SDI donde ha participado el autor son la exploración de nuevas alternativas para garantizar la calidad de la información descriptiva de los recursos geoespaciales y la identificación de recursos geoespaciales ocultos en catálogos (Renteria-Agualimpia et al., 2013c), la exploración de avances en motores de búsqueda semántica, su integración con aspectos geoespaciales (Renteria-Agualimpia et al., 2010), y el desarrollo de modelos de recuperación de información geográfica con múltiples criterios basados en la integración semántica geoespacial (Renteria-Agualimpia and Levashkin, 2011). Estos trabajos han involucrado la identificación, el análisis y la caracterización de los errores más comunes e inconsistencias geoespaciales de los metadatos de servicios Web en el contexto de recuperación de información geográfica (Renteria-Agualimpia et al., 2013b, 2014). Además, el autor ha colaborado en el desarrollo de estudios sobre el estado y la disponibilidad de Servicios Web OGC (López-Pellicer et al., 2011, 2012b,c). Algunos resultados de la línea de investigación en Bibliotecas Digitales, donde ha participado el autor son la exploración de nuevas alternativas para garantizar la calidad de la información descriptiva de los recursos cartográficos (Renteria-Agualimpia et al., 2013a). Además, el autor ha colaborado en el estudio de nuevas formas de mejorar la visibilidad de los recursos geoespaciales en la Web (Lacasta et al., 2014b,a), y nuevas formas de mejorar la detección de inconsistencia espacial, ambigüedad de topónimos, y la detección de la existencia de problemas derivados de la falta de cobertura suficiente para los topónimos de grano fino en diccionarios geográficos/gazetteers (Moncla et al., 2014). Esta tesis está incluida en las líneas de investigación antes mencionadas y es el resultado de las investigaciones citadas. El trabajo futuro contribuirá con el Asesoramiento de la Calidad de las descripciones de recursos geoespaciales, la caracterización de otros tipos de inconsistencias y la evaluación de su impacto en los procesos de recuperación de información.

² <http://iaaa.cps.unizar.es/showContent.do?cid=presentacion.EN>

Executive summary

Background. The advances in Information Technologies focused on information describing geospatial phenomena have revolutionized the information handling activities in research and industry domains. The effective access to geospatial information acquires a critical importance in these knowledge-based contexts. However, the increasing volume of geospatial data everyday makes direct search infeasible. As alternative, many information systems search geospatial metadata, that is, data that describes geospatial data. From emergency, rescue and locating systems to geopolitical, military, and industry using decision-making systems based on geographic information, it is essential to have access to geospatial resources through consistent metadata description and a minimum level of quality in order to ensure the resource retrievability (Hartmann and Stuckenschmidt, 2002; Martins et al., 2007). Moreover, for some authors, the quality and the consistency of a metadata description could be the difference between the life and death or between the success and failure (Dushay and Hillmann, 2003; Bruce and Hillmann, 2004; Hillmann et al., 2004).

This thesis researches how to assess the quality of a particular kind of metadata: the metadata that describe the spatial location of a resource. In particular, this thesis researches how to deal with problems that may surface when a metadata record describing some resource has more than one property that intends to describe the location of resource, that is, the metadata record contains semantically close geographical properties (Miles and Bechhofer, 2009). This problem is closely associated to the facility which georeferenced resources can be found in an information system because metadata proper-

ties are key to discovery, access and retrieval of resources in many systems based on indexed catalogs (Goodchild and Zhou, 2003; Hill, 2006).

Objective.

The problems derived from the inconsistency of Geospatial metadata might be mitigated after performing a Quality Assessment (QA) of the geospatial description. This thesis analyses a semi-automatic approach to detect geospatial inconsistencies and to suggest possible solutions for an inconsistency based on a geospatial context, which is built from consensual geospatial descriptions surrounding the inconsistent resource. In addition, the analysed solution should provide assessment to perform quality processes such as, curation and preservation of cartographic material in the digital repositories field.

Scope. The QA analysis is restricted to two kinds of resources: (1) Metadata Web services compliant with the Open Geospatial Consortium (OGC) specifications and (2) MARC21³ metadata describing cartographic materials. The OGC leads the development of open and standardized Web service interface specifications for accessing geospatial information since 1994. Many companies, government agencies and universities are members of OGC, and they participate in consensual processes to develop publicly available Web service interface standards for the access to geospatial resources. Many OGC Web service interface specifications have become standards of the International Organisation for the Standardization. Related to OGC standards, this thesis has its focus in the most used OGC standards in Spatial Data Infrastructures (SDI) (Nebert, 2004). In particular, we apply our QA analysis on Web Map Services metadata stored in OGC Catalogue Service (CSW, (Nebert et al., 2007)). However, our developed systems can be applied to other kind of Web services and SDI resources with semantically close geographical properties. Although our methodology could be used to assess the quality of thematic and temporal properties, we will solely consider semantically close geographical properties.

On the other hands, MARC standards are a set of digital formats for the description of items catalogued by libraries, such as books. It was developed

³<http://www.loc.gov/marc/bibliographic/>

by the US Library of Congress during the 1960s to create records that can be used by computers, and to share those records among libraries. By 1971, MARC formats had become the national standard for dissemination of bibliographic data in the United States, and the international standard by 1973. There are several versions of MARC in use around the world, the most predominant being MARC21, created in 1999 as a result of the harmonization of U.S. and Canadian MARC formats, and UNIMARC, widely used in Europe. Additionally, in many libraries around the world, MARC21 metadata is the most used standard to document resources describing geographic phenomena over the surface of the earth (Furrie, 2009). In MARC21 there are several different fields that can encode different aspects of direct/indirect spatial references including different ways to associate geographic codes, or different ways for expressing the geospatial reference method used for the coordinates in the direct spatial references. We will solely consider the two most frequent semantically close geographical properties that we found analysing the experimental datasets: the Direct Spatial References (geographical extent/spatial footprint) and the Indirect Spatial References (place name) (FGDC, 1998b). In this thesis we do not take into account other geographical properties.

The SDI scenario is used as test case to validate all our proposed architecture. The descriptive information of the spatial resources in SDI is structured and proceeds from experts of the geographical domain. This makes us to think that the provided spatial description in SDI must be better than other scenarios and domains where the descriptions proceed from unstructured information and non-experts in the geographical domain, for example, Digital Libraries domains. In this sense, the methodology and the architecture have been tested with a more difficult case, cartographic materials. Our analysis has been restricted to two geographic areas, Spain and the United States of America. The described restriction guarantees that the contributions of this thesis can benefit two of the most frequent scenarios using geographic metadata (Spatial Data Infrastructures and Digital Libraries).

Regarding to the scope of geographical Knowledge Organisation System (KOS) (Hodge, 2000; Miles and Bechhofer, 2009) used in the spatial ranking process, this work has analysed several knowledge representation systems. These systems have been mainly simple SKOS (Simple Knowledge Organi-

sation System(Isaac and Summers, 2009)) vocabularies and RDF/XML graphs. The KOS used in the tasks of spatial ranking must cover the geographical extent of the assessed geographical resources. Also, the level of granularity of the spatial footprints in the KOS must be in accordance with those in the analysed collection. In this thesis we restrict ourself to work with KOS with footprint of two-dimensions (2D). However, by means of simple processes, the 2D footprints can be simplified to classical points (1D). The open research line and the challenge is to shift from 1D to 2D geographical footprint analysis to assess their quality.

Method. The methodological approach comprises aspects related with software engineering, knowledge engineering and artificial intelligence. The software engineering methodology is a classic incremental development of the solution (Boehm, 1988; Larman and Basili, 2003). The knowledge engineering methodology is based in the steps proposed by the Methontology framework (specification, conceptualization, formalization, integration, implementation, and maintenance) (Fernández-López et al., 1997).

Previous experience, future work.

To ensure the access, retrieval and visualization of resources in the context of distributed and interoperable information systems are common and priority goals for many domains, for example, in SDI and Digital Libraries. One of the most consolidated cases is the European INSPIRE initiative, whose aim is to create a European SDI. One of the research lines of the IAAA research group⁴ focuses on SDI aspects related with the description of geospatial data and services, the discovery of these resources through standard catalogues, and the conceptual and architectural aspects related to geospatial data and services. Some research results of the SDI research line where the author has participated are the exploration of new alternatives to ensure the quality of the descriptive information of geospatial resources and the identification of hidden geospatial resources in catalogues (Renteria-Agualimpia et al., 2013c), the exploration of the advances in semantic search engines and the integration of geospatial aspects (Renteria-Agualimpia et al., 2010), and the development of multi-criteria geographic information retrieval models

⁴<http://iaaa.cps.unizar.es/showContent.do?cid=presentacion.EN>

based on geospatial semantic integration (Renteria-Agualimpia and Levashkin, 2011). These works have involved the identification, analysis and characterization of the most common errors of geospatial inconsistencies of web services metadata in the context of Geographic Information Retrieval (Renteria-Agualimpia et al., 2013b, 2014). Additionally, the author has collaborated in the development of reality checks of the status and availability of the OGC Web Services (López-Pellicer et al., 2011, 2012b,c). Some research results of the Digital Library research line where the author has participated are the exploration of new alternatives to ensure the quality of the descriptive information of cartographic resources (Renteria-Agualimpia et al., 2013a). Additionally, the author has collaborated in the study of new ways for improving the visibility of geospatial resources on the Web (Lacasta et al., 2014b,a), and new ways for improving the detection of spatial inconsistency, ambiguous toponyms, and the detection of the existence of problems derived from the lack of enough coverage for fine-grain toponyms in gazetteers (Moncla et al., 2014). This thesis is included in the aforementioned research lines and is the result of the cited researches. Future work will improve contributions in the QA of descriptions of geospatial resources, the characterization of other kinds of inconsistencies and the evaluation of their impact in information retrieval processes.

Contents

Acknowledgements	iv
Resumen ejecutivo	v
Executive summary	x
List of Figures	xviii
List of Tables	xx
Nomenclature	xxi
1. Context and research issues	1
1.1. Background	1
1.2. Motivation	2
1.2.1. Digital Libraries	8
1.2.2. Spatial Data Infrastructures	16
1.2.3. Quality Assessment on Geographical properties	21
1.3. Research Hypothesis	22
1.4. Methodology	23
1.5. Scope	24
1.6. Contributions	26
1.7. Organisation of the Dissertation	27
2. Related Work	29
2.1. Introduction	29

2.2.	Quality Assessment in SDI	31
2.3.	QA in Digital Libraries	38
2.4.	Spatial Ranking for QA	44
2.5.	Summary	47
3.	Quality Assessment using two-dimensional Spatial Ranking	49
3.1.	Introduction	49
3.2.	Collecting	50
3.2.1.	Crawling	50
3.2.2.	Harvesting	53
3.3.	Geo-Extraction	56
3.4.	Reverse Geocoding	57
3.5.	Geospatial Clustering	61
3.6.	Clustering Validation	62
3.7.	Report Generation	64
3.8.	Summary	65
4.	Architecture and Implementation	67
4.1.	Introduction	67
4.1.1.	Core technical details of shared components	69
4.1.2.	Module based on Geonames+DBpedia approach	71
4.1.3.	Module based on Overlapping approach	71
4.1.4.	Module based on Hausdorff approach	72
4.2.	System Evaluation and Results comparisons	74
4.2.1.	Comparison between 1D and 2D approaches	75
4.2.2.	Comparison between 2D approaches	75
4.2.3.	Analysis of results	79
4.3.	Summary	80
5.	Quality Assessment for Digital Libraries	81
5.1.	Introduction	81
5.2.	Experimental details	82
5.3.	Analysis and results	84
5.4.	Discussion	92
5.5.	Summary	95

6. Conclusions	97
6.1. Summary of Contributions	97
6.2. Future Work	100
6.3. Final Conclusion	103
A. Web Map Service Metadata	105
B. Digital Library Metadata Records	107
C. Contributions	115
Bibliography	119

List of Figures

1.1.	Common problems of geospatial inconsistencies on libraries.	14
1.2.	Impact of the quality assurance information in SDIs	17
1.3.	Nested evaluation frameworks of an IR system.	27
2.1.	Different approaches to geographic information quality . . .	32
2.2.	Mapping between the Quality parameters	34
2.3.	INSPIRE geoportal problem.	35
2.4.	Example of a search in the NSGC catalogs.	36
2.5.	Example of a search in the CHDuero catalogs.	37
2.6.	Digital Libraries projects working with georeferenced	40
2.7.	Example of inconsistency in the David Rumsey Map Collection	41
2.8.	Example of inconsistency in the KartenPortal about Italy . .	42
2.9.	Example of inconsistency in the KartenPortal about Spain .	43
3.1.	Methodology to detect geospatial inconsistencies in metadata.	51
3.2.	Example of WMS Capability document	52
3.3.	Example of coordinates in MARCXML and MODS formats .	54
3.4.	Example of a XML record harvested from the LoC	55
3.5.	MARC geographic location fields most often used.	57
3.6.	Conceptual idea of the Reference Systems Transformer. . . .	58
3.7.	Example of a Reference System Transformation.	58
3.8.	Flow of the Harvester, ETL, and Reverse Geocoder processes.	60
3.9.	Inconsistent metadata record report.	65
4.1.	Architecture	68

4.2. GIR module.	70
4.3. Debepeedia Wikipedia Geonames Approach.	72
4.4. Overlapping Approach.	73
4.5. Hausdorff Approach.	74
4.6. Comparison between 1D and 2D approaches	76
4.7. Comparison between 2D approaches	77
5.1. Dataset distribution	85
5.2. Global vision of the spatial logical inconsistencies in the LoC	86
5.3. Spatial logical inconsistencies in the LoC focused on USA . .	86
5.4. Geospatial Unmatching: Cuba country	87
5.5. Geospatial Unmatching: Caledonia county	87
5.6. Contextual inconsistencies: spatial synecdoche	89
5.7. Contextual inconsistencies: South Dakota	89
5.8. Contextual inconsistencies: Ohio State	90
5.9. Example of lack of information in a KOS	90
5.10. Contextual inconsistencies: Lake county	91
5.11. Problems of separability in reduced representations	94
5.12. Conceptual differences between 1D and 2D clustering	95
B.1. Snapshot of the Website of the LoC that shows the geospatial inconsistency in the coordinate latitudes.	110
B.2. Snapshot of the Website of the LoC that shows the geospatial inconsistency in the coordinate longitudes.	114

List of Tables

1.1.	Comparative occurrence of geographic coverage	11
1.2.	Details of geospatial inconsistencies about Germany	13
1.3.	List of the most common U.S. place names	15
1.4.	The main characteristics of OGC Web Services	20
2.1.	Data quality problems versus information quality problems .	30
4.1.	Characterization	78
5.1.	Contextual inconsistency caused by systematic error	91
5.2.	Typology of the geospatial inconsistencies in the LoC	92

Nomenclature

CQL Contextual Query Language

CSDGM Content Standard for Digital Geospatial Metadata

CSW Catalogue Service for the Web

DBSCAN Density-Based Spatial Clustering of Applications with Noise

DSR Direct Spatial References

ESDIN European Spatial Data Infrastructure

ETL Extraction, Transformation and Load

FGDC Federal Geographic Data Committee

GI Geographic Information

GIR Geographic Information Retrieval

GIS Geographic Information System

GPS Global Positioning System

GWS Geospatial Web Services

HD Hausdorff Distance

INSPIRE Infrastructure for Spatial Information in the European Community

IR Information Retrieval

ISO International Organisation for the Standardization

ISO/TC 211 ISO Technical Committee 211 Geographic information/Geomatics

ISR Indirect Spatial References

KOS Knowledge Organisation System

LCCN Library of Congress Control Number

LoC Library of Congress

MARC MACHine-Readable Cataloging

MBBOX Minimum Bounding Box

MODS Metadata Object Description Schema

NSDL National Science Digital Library

OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting

OGC Open Geospatial Consortium

QA Quality Assessment

RDF Resource Description Framework

SDI Spatial Data Infrastructures

SDTS Spatial Data Transfer Standard

SKOS Simple Knowledge Organisation System

SRU Search/Retrieve via URL

URL Uniform Resource Locator

WFS Web Feature Service

WPS Web Processing Service

XML Extensible Markup Language

Let us to begin with the next history:

On Exactitude in Science . . . *In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography.*

From Jorge Luis Borges, *Collected Fictions*, Translated by Andrew Hurley
Copyright Penguin 1999 .

The short story, quoted here in its entirety, first appeared in 1946 in *Los Anales de Buenos Aires*. It was later published in *El Hacedor*, 1960, under a section titled 'Museo'. The story appears under the authorship of 'Suarez Miranda, *Viajes de varones prudentes*, Libro IV, Cap. XLV, Lerida, 1658'.

Information quality assessment (accuracy, completeness, consistency, etc.) must have a purpose. Work very hard to produce a description more complex than the resource itself is not affordable (economically sustainable, temporally viable) taking into account the volume of the available spatial information today). Resource description must have a minimum level of quality in order to know if the searched resource satisfies the users needs, and then, allow them to retrieve it.

To err is human; to try to prevent recurrence of error is science.

anonymous aphorism

Chapter 1

Context and research issues

1.1 Background

Nowadays information is increasingly becoming a critical resource. For institutional and individual processes that depend on information, the quality of information is one of the key determinants of the quality of their decisions. The popular computing saying “garbage in, garbage out” expresses the problem succinctly: when an information system processes as input unintended or nonsensical information produces as output undesired and nonsensical information. Unfortunately, as more information becomes available for use, it becomes increasingly difficult for users to identify “garbage” and many problems related to the conformance or purpose of information arise affecting providers (re-use and maintaining tasks), managers (assessment, analysis, sharing, curation and preservation tasks) and end users (accessibility, retrieval and visualization/interpretation tasks). Therefore, the success of the information exchange among information systems actors depends on the fitness for use of information. That is, it depends on the quality of information. We are using the term quality here in the sense coined by Joseph M. Juran, one of the fathers of the Quality Assessment. That is, quality means “fitness for use” (Juran, 1962).

The increasing volume of data everyday makes infeasible to search through their content directly. Many information systems use instead metadata. That is, data about data. Metadata is stored in catalogues that are used for

searching resources and eventually accessing to them. For example, Digital Libraries use metadata catalogues for locating bibliographic resources. Spatial Data Infrastructures use metadata catalogues for discovering and accessing to spatial resources. The process of metadata creation is complex, tedious and hard. Metadata production consumes enormous amounts of time (Broeder and Wittenburg, 2006), their creation is one of the major challenges (Valkeapää et al., 2007). Metadata creation is expensive and labor intensive, but the danger of hidden materials is greater. An object without metadata is for most purposes invisible and effectively lost (Register et al., 2009). Consequently, quality problems in metadata are also expensive. These problems affect directly the efficiency and the effectiveness of retrieval tasks. Also, with this large amount of available resources and poor quality information describing them it is complex to retrieve useful and relevant information for the user purposes (Duval et al., 2002; Barton et al., 2003). These challenges have opened new research lines such as research on the application of Quality Assessment mechanisms for ensuring the fulfilment of the goals of accessibility, retrievability and interoperability of information systems. This thesis belongs to the aforementioned research line. This thesis researches how to assess the quality of a particular kind of metadata, the metadata that describe the spatial location of a resource, and the problems that may surface when a metadata record describing some resource has more than one property that intends to describe the location of resource. This problem is closely associated to the facility which georeferenced resources can be found in an information system as we describe below.

1.2 Motivation

Information systems are the technological means by which people and organisations, store, gather, process, discover, explore, retrieve, visualise and disseminate/share information. In information systems, one of the essential pieces in retrieval processes is an appropriate documentation of resources. This documentation is called metadata and can be defined formally as “structured data about data” or “data which describes attributes or properties of a resource” (Miller, 1998; Baeza-Yates et al., 1999). A metadata record may offer

a description of the content, quality, condition, authorship, location and any other characteristics of a resource. It may also provide a standardized representation of information, and constitutes the mechanism to characterise the resources in order to enable other users and applications to make use of such resources. Metadata records, each one describing a specific resource, are usually published and accessible through catalogue systems, in a similar way of traditional library catalogues. Library catalogues provide users and applications with the possibility of finding the resources of their needs or interest. Thus, metadata are the basic component that facilitates accessibility, retrievability and interoperability of resources and services offered by information systems.

Metadata play a core role in the library, documentation, cataloguing, and information science profession. Particularly, in Digital Libraries, since they assist users in discovering and retrieving the useful resources for their needs or queries through enormous digital collections (Søndergaard et al., 2003; Grossner et al., 2008). Most information systems used by Digital Libraries make use of metadata as the key for discovering and retrieving tasks (Robertson, 2005). Metadata are a powerful tool that enables the user to explore and select relevant resources quickly and easily (Barton et al., 2003). This point of view is in accordance with the conception that metadata are not only used to document data (the traditional viewpoint), but also in tasks such as discovery, analysis, interpretation and efficient accessibility. Hence, metadata have a notable role and a relevant effect in the main library tasks.

The rapid advances in Digital Libraries have resulted not only in a proliferation of the amount of information and metadata available, but also in a proliferation of many semantically similar information that could satisfy the user needs, and in consequence, the challenge to find and retrieve the most appropriate information (this is the fitness for use or purpose principle). This challenge have impuled the development of disciplines and research areas in Information Retrieval to facing with the problem of efficient discovery, access, and retrieval of the searched information described by the metadata, e.g. Quality Assessment for metadata.

The assessment from the viewpoint of fitness for use implies taking into account the descriptive elements (properties) of metadata that users most

commonly search. Metadata have properties that describe the main aspects of a resource. Some of these properties are conceptually similar or semantically equivalent because they describe the same concept from different views. When it happens, then a direct mapping or close match between these properties can be done. These properties are known as ***Semantically Close Properties*** (Miles and Bechhofer, 2009).

The proliferation of data and services describing phenomena associated with the earth surface have impeded the development of specific metadata to describe this kind of information, that is, geospatial metadata (Nogueras-Iso et al., 2004). Geospatial metadata also have *Semantically Close Properties*. These properties describe the geographical context using different codifications (e.g., textual and numerical) and perspectives (e.g. official names, geographic codes) (Hill, 2006). For example, the textual place name property and the numerical coordinate property refer to the same spatial context, that is, the location or the geographical extent covered by the resource. More formally, in the context of Geographical Information, these properties can be classified as **Direct Spatial References** (for coordinates) and **Indirect Spatial References** (for place names). In the present thesis, pairs of these kinds of properties are called ***Semantically Close Geographical Properties***. Hence, be able to assess if each pair of semantically close geographical properties is consistent is relevant for the remain of this thesis, in particular, the pair formed by a Direct Spatial Reference and an Indirect Spatial Reference.

Hereafter, this thesis uses the expression *Semantically Close Geographical Properties* to identify pairs of geographical properties formed by a direct spatial reference and an indirect spatial reference that allow us to retrieve resources by means of the first property (place name), by means of the second one (coordinates) or by both.

In the modern information systems and Digital Libraries is very common to brows, query, restrict and visualise resources directly on the map. Maps are used as a geographical metaphor that facilitate the accessibility and retrieval of the geolocated resources (Grossner et al., 2008). In this modern geographical metaphor, the map is part of the user query. When a user asks

for a resource R that refer to a place in a demarcated/visualised region, implicitly, the query contains the logical "AND" operation. e.g., (*R.placeName like the place*) AND (*R.extent intersects the region*). Such queries fail (i.e. return void) when conditions do not occur simultaneously. For example, if the values of this pair of semantically close properties are inconsistent the resource will not be retrieved. That is, inconsistent values in semantically close properties may turn resources into invisible or hidden resources in retrieval tasks when the geographical metaphor is used. In Digital Libraries these inconsistency problems are crucial due to libraries using the metadata properties that describe the resources to access and retrieve them.

With the large amount of spatial information available today it is very common to use the geographical metaphor (e.g. Google Maps based applications) and to restrict the geographical area of search to the view to optimise time and costs, and also to help user to find information according to their specific needs or purposes. Unfortunately, with the amount and diversity of information available, it becomes increasingly difficult to identify quality problems in the descriptive information (metadata properties) used to retrieve the resources. These quality problems reduce the effectiveness of IR systems used in DL. Quality Assessment (QA) approaches have been proposed as an alternative to deal with these kind of problems in the context of information systems. QA for geographical data are very common in one of the most important and disseminated information systems, the Geographical Information Systems (GIS) one. For institutional and individual processes that depend on geographical information, the quality of this information is one of the key determinants of the quality of their decisions and processes. Although, Digital Libraries are taking advantage of GIS to manage cartographic materials, however the same quality problems are present in this geospatial context.

The spatial inconsistency, ambiguity, and in general, the geospatial quality problems are more evident in this context due to the possibility to be visualised on a map by users. This challenge has impeded the development of specialised research areas in Geographic Information Retrieval (GIR). GIR is an active field of the information science concerned among others on capturing user needs from the queries, processing these queries, finding and returning matching metadata in a repository, and finally evaluating the rel-

evancy of the results (Jones et al., 2002; Janowicz et al., 2010). It is important to note that GIR systems that use metadata to facilitate content-based search rely on the quality of metadata properties (Hartmann and Stuckenschmidt, 2002; Martins et al., 2007). The power of metadata properties to help retrieve resources quickly and easily can be affected by quality problems that cause inconsistency and ambiguity. Commonly, the traditional GIR systems deal with the inconsistency by means of natural languages programming techniques. These techniques evaluate the textual description of the metadata resources and mainly the conformance with standards, that is, they generally perform a syntactic analysis (Veregin, 1999; Kainz, 1995; Tolosana-Calasanz et al., 2006; Wang, 2008). However, most of these works do not take into account the importance and the advantages of the geospatial co-occurrence. In this sense, the consensus provided by geospatial phenomena whose metadata describe a common place is not exploited to ensure the consistency and then the quality of co-occurring metadata.

A scenario where the quality of geographic information is also very important is in the metadata catalogues of Spatial Data Infrastructures. Geographic information is one of the most critical elements underpinning decision making for many disciplines, organisation and agencies at the local, national, regional, and global levels. As a consequence, the number and diversity of potential users of this kind of information have increased significantly over the last few years. This has given rise to new demands for improved system of systems (infrastructures) that support discovery, access, sharing, and use of this geographic information in the decision-making processes. These systems or infrastructures are given the name of "Spatial Data Infrastructures" (SDI). As is mentioned by Abel et al. (1999), SDI initiatives are aimed at assembling digital collections of core spatial databases and at making the data available as a common resource. But, databases are not coalesced into a single warehouse, they are maintained as a distributed system of systems. For this reason, the quality of the metadata properties used to access to shared resources is critical for exchange processes. SDI promise to enable much wider use of GIS by ensuring faster and easier availability of spatial data, but previous Quality Assessment processes need to be done. Several studies (Chandler et al., 2000; European Commission, 1998; Nogueras-Iso et al., 2004) have remarked that although the value of these

infrastructures and the potential of geospatial information are recognised by governments, industries and the general public, the effective use, access and exchange of geospatial information is inhibited by poor knowledge of the quality status of the information, poorly documented information about the datasets, and inconsistencies in the meta-information used to exchange. For example, Frank (2008) recognises the need to assess the influence of data quality on the decision. Although in some scenarios it is a less important task, it is very used to make political decisions or to design constructions (e.g., higher precision data for cadastral boundary).

Although SDI and DL do not have the same historical path, they share the same philosophy of documentation, both, DL and SDI use catalogs to describe and to access to their resources. In this sense, the SDI inherited the potential quality problems derived from spatial inconsistency and ambiguity exposed above. The spatial inconsistencies and quality problems of the information in SDI environments may have higher impacts due to the kind of decision-making processes supported by the spatial infrastructures and due to the kind of users. SDI users are not common users, they are government agencies (e.g., cadastral), companies, etc. Due to the specialised, advanced and detailed spatial knowledge of the geographical domain experts to manage SDI catalogues, the findings of the SDI scenario will be used as benchmark to analyse and assess the quality problems in Digital Library catalogues. One would expect the number of quality problems in SDI scenario are minimised because SDI personnel are experts and technicians with specialised knowledge of the geographical domain. Indeed, the quality problems of the spatial information used to retrieve SDI resources (geographical data and services typically) are less, but still common, and also, they inhibit the effective of the GIR systems used in these infrastructures.

From the point of view of interoperability, SDI and DL as a systems of systems (Béjar et al., 2009) Goodchild et al. (2007) acknowledge that the development and exchange of geospatial metadata represents one of the most important practices determining the success of large-scale data sharing activities in information systems and libraries. However, the retrieval, interoperability and sharing tasks depend on the Quality Assessment of the metadata used for the information exchange. Thus, the re-use of information depends on assess the coherence and consistency of the semantically close properties

used to access the searched resources (Lutz, 2005; Piasecki et al., 2010). Beall (2006) and Park (2009) argue that inconsistent properties can create conceptual ambiguities and consequently hinder consistent resources. Even if all other aspects of a SDI or a DL system worked perfectly, poor quality metadata would degrade the quality of the searching results. Diane Hillmann, who was instrumental in the deployment of the National Science Digital Library (NSDL), has written extensively on this issue (Bruce and Hillmann, 2004; Dushay and Hillmann, 2003; Hillmann et al., 2004). Hill (2006) acknowledges that an inconsistent spatial description of geographic resources could easily generate discrepant results, bad weighted results, and even a permanent omission of results that could satisfy the queries (i.e. *invisible or hidden geographic resources*).

Advances in GIR are being used in many domains, such as, medicine (Boulos, 2005), pharmacy (Pardo et al., 2010), museums (Renteria-Agualimpia, 2009), tourism (Torres et al., 2007), photography and images (Rattenbury and Naaman, 2009). Two of the most extended uses of GIR are in the Spatial Data Infrastructures and in the Digital Libraries, we center our attention in them. In the first one, GIR systems are used to retrieve spatial information through Geospatial Web Services, in the second one, GIR systems are used to retrieve cartographic materials. Both domains depend on the Quality Assessment of the descriptive information used to access the described resources. The next sections focus the attention on these two domains.

1.2.1 Digital Libraries

As Giles (2011) exposes, in a perfect world, every data set would be fully and clearly described by a complete and consistent metadata record. The record would be maintained regularly, so that the information content remained up-to-date, accurate and consistent. Individual metadata records would take into account similar records that were already in existence and would ensure that the two records reflected their close relationship while clearly describing any distinctive features. This would guarantee that appropriate database searches would be able to recall apposite information with precision. "Recall" and "Precision" are terms originally used in this context

within the library community. Recall describes the capability to discover relevant records and precision is the proportion of the retrieved items that are relevant; a search that misses a lot of relevant information is described as having poor recall. A search that recalls relevant records along with numerous irrelevant ones is said to have poor precision. The user finds it difficult to identify the valuable records amongst the numerous returns. In the real world, individual metadata records fall far short of the ideal. Poor quality metadata can lead to misleading conclusions and costly mistakes, yet few people understand the nature of the errors associated with their own metadata.

Digital Libraries with cartographic information use digital repositories, digital collections and metadata Catalogues to managing their contents. One of the keys are the metadata. Metadata are a powerful tool that enables the user to discover and select relevant materials quickly and easily (Barton et al., 2003). But, poor quality metadata can mean that a resource is essentially invisible within a repository or archive and remains unused. Clearly metadata quality has an important role, and its assessment a significant impact in the task of information retrieval. Where metadata error exist, they can easily block access to material available through a Digital Library. These errors are most serious when metadata serves as a surrogate for resources held in a Digital Library and full text searching is not available, for example, images and maps databases are particularly vulnerable to metadata errors and inconsistencies because virtually all search access to images and maps databases is through metadata. The importance of metadata cannot be overstated. According to Robertson (2005, pp. 295) "*supporting the development of quality metadata is perhaps one of the most important roles for Library and Information Science professionals*".

Geospatial description is a core component of Digital Libraries. For example, Petras (2004) analyzed around 5 million records from the University of California library catalogue and found that approximately 35% of the records contain data in MARC21¹ fields related to geographic information. In addition to this, geospatial information can help to reveal unknown spatial patterns, increase the recall of information retrieval systems, and

¹<http://www.loc.gov/marc/bibliographic/>

enhance real world experiences of the users, since most events can be visualised, explained, and understood in geographic terms (Samulenok and Rubin, 2012). Libraries have traditionally included Geographic Information and have developed Geographic Information Retrieval systems to perform spatial queries on metadata (Buckland et al., 2007; Zong et al., 2005). In this sense, semantically close geographical properties are one of the core dimensions for many Digital Libraries and National Archives. For example, the National Archives of the United Kingdom, the Alexandria Digital Library and the Library of Congress (LoC) (Goodchild, 1995) are good examples. Many of the records that they hold make reference to spatial data. For example, the Geography and Map Division² of the LoC stores the largest and most comprehensive cartographic collection in the world with collections numbering over 5.5 million maps, 80,000 atlases, 6,000 reference works and a large number of other cartographic materials in other formats. Moreover, many user queries in Digital Libraries involve the spatial dimension. For example, approximately one fifth of queries in the National Archive of the United Kingdom involve place names (Clough et al., 2011).

Additionally, in her study of three Digital Libraries, The European Library³ (Europeana), American Memory⁴, and Opening History⁵, Zavalina (2012) exposes that subject-specific collection properties were the most consistently represented in free-text description elements across the three Digital Libraries. And the geographic coverage of a digital collection was the fourth most widely represented collection property in description metadata elements. Particularly, the Table 1.1 presents a comparative frequencies of occurrence of geographic coverage property in description.

Based on the concepts of quality exposed above, the “good” metadata reflects the degree to which it is fit for the intended functional purpose of supporting common library user tasks and services. Park (2009) analysed overlapping criteria and matrices in her survey of research on metadata quality evaluation. The study finds that completeness, accuracy, and consistency are the most commonly used criteria in measuring metadata quality.

²<http://www.loc.gov/rr/geogmap/>

³<http://www.theeuropeanlibrary.org/>

⁴<http://memory.loc.gov>

⁵<http://imlsdcc.grainger.uiuc.edu/>

Table 1.1.: Comparative frequencies of occurrence of geographic coverage property in description (Source: Zavalina (2012))

Property	Europeana	American Memory	Opening History
Geographic coverage (% of Metadata records)	55	69	81

The completeness of metadata means that individual objects are described using all metadata elements that are relevant to their full access capacity in digital repositories (Bruce and Hillmann, 2004). Accuracy concerns the degree to which the data content of metadata elements corresponds to the individual objects being described and the way that it should be represented (Stvilia et al., 2007). Consistency can be measured by looking at data value on the conceptual/semantic and structural levels, respectively. On the conceptual/semantic level, consistency is affected by the degree to which the same data values or property are chosen for representing similar concepts in resource description, that is to say, semantically close properties. On the other hand, structural consistency concerns the extent to which the same data structure or format is used to represent information chosen for given metadata elements (Park, 2009). Of the three major criteria, Caplan (2003) and others have shown that consistency especially seems to pose the greatest challenge in ensuring metadata quality in the heterogeneous, distributed context of digital repositories due to conceptual ambiguities and semantic overlaps of various metadata elements. The conceptual/semantic consistency of metadata elements is the focus of this thesis. In particular the consistency of the semantically close geographical properties, because of metadata referencing geospatial resources use these properties to provide information about the available resources. However, in addition to document or describe the available cartographic resources, semantically close geographical properties are used in tasks such as discovery, analysis, interpretation and efficient accessibility.

In accordance with the perspective of Caplan, Toy-Smith (2010) argues that among quality dimensions, metadata consistency should be one of the pri-

mary consideration in the development of digital collections. Hill (2006) points out that inconsistent spatial description of geographic resources could easily generate discrepant results, weighting and ranking problems, inaccessibility, and even a permanent omission of some records in the results. That is, *inconsistent spatial descriptions yield invisible or hidden geographical resources*. For example, a user would expect that a map about Germany should be returned either through textual queries containing the term 'Germany', spatial queries with the bounding box of Germany or both. If the term 'Germany' is not present, overly simplified (e.g., 'DE'), misleading (e.g., 'Germania') or wrong (e.g., 'Gyrnamy') in the metadata record of a resource, such resource will not be retrieved through keyword queries. Likewise, if the geometry is not present, simplified (e.g., a point), misleading (e.g., covers the geographical region named Magna Germania) or wrong (e.g., covers a different country), such resource will not be retrieved through spatial queries either.

Nowadays, such problems are quite common in library collections. Figure 1.1 shows real examples of spatial inconsistencies found in metadata records published by the LoC. This example shows metadata records that describe the location of (A) the Germany administrative divisions in the middle of the Atlantic Ocean, (B) the regional atlas of the Federal Republic of Germany over Ireland and (C) the general chart of Germany over Ukraine. All details are summarised in the Table 1.2.

The personnel involved in the preservation of the quality must revise, and prevent possible inconsistency problems. Also, they should be more careful with the spatial ambiguity and their consequences in retrieval systems. Generally, the level of spatial ambiguity related to a toponym difficulties the documentation task of geographical resources in the context of DL. Sometimes the spatial ambiguity occurs when librarian personnel assign the spatial scope (the footprint of a toponym or place name) to a resource (Ding et al., 2000; Silva et al., 2006). Commonly, this process of assigning a toponym or the place name to a location is called geocoding and the inverse process is the reverse geocoding. In general, these processes use *gazetteers*, that is, dictionaries or indexes of named geographic features (places) (Hill and Janée, 2004). When the frequency of a toponym is high, the ambiguity grows, and hence the cataloguing task is more susceptible to errors/incon-

Table 1.2.: Details of some geospatial inconsistencies in the LOC about Germany

LCCN	Place Name	Description	Case	URL
2010589731	Germany	Germany Administrative and political divisions Maps	A	http://lccn.loc.gov/2010589731
325901201x	Germany	German travel map	B	http://lccn.loc.gov/83694582
2003683097	Germany	Military geographic description of the Federal Republic of Germany	B	http://lccn.loc.gov/2003683097
3471401202	Germany	Regional atlas	B	http://lccn.loc.gov/82215282
2011585205	Germany	Federal Republic of Germany	C	http://lccn.loc.gov/2011585205
2011585214	Germany	General Chart of Germany	C	http://lccn.loc.gov/2011585214

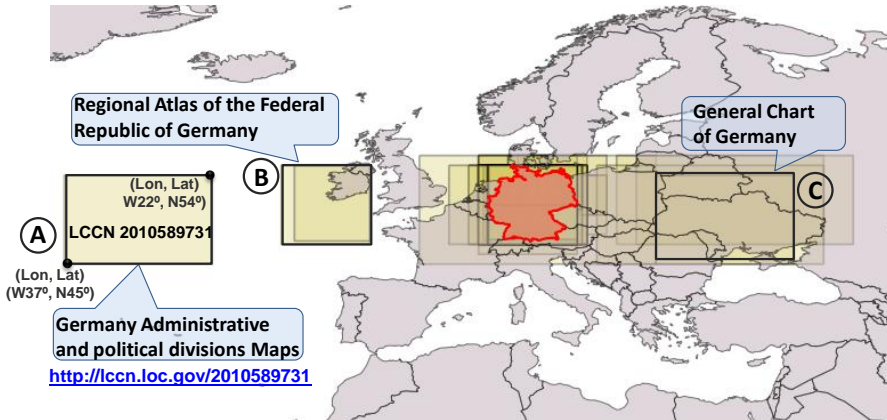


Figure 1.1.: Common problems of geospatial inconsistencies on libraries.

sistencies, one example of that is shown in Table 1.3, it shows the list of the most common U.S. place names. Even, toponyms like “Washington” may refer to geographical names, personal names, or another. Even more, the inconsistencies and ambiguities may exist in this specific kind of geographical name, that is to say, the spatial footprint and the place name may refer to different levels of spatial entities, such as a Village, a Township, a City, a State or a Country, a lake, a river or a mountain, etc. In conclusion, there are many factors causing or contributing to quality errors, for instance, the reuse of old/outdated metadata, gazetteers with low level of granularity/detail, the high level of ambiguity of some toponyms, bad transformation between coordinate systems and between reference systems, poor knowledge of the spatial, cultural and temporal context of the information, and poorly representation of the documented spatial information.

Intensive work in this area shows that the problem of resource description (metadata) consistency has drawn research interest in the past years (Servigne et al., 2000; Rodríguez, 2005; Devillers et al., 2006; Hillmann, 2008; Brisaboa et al., 2014). This is because of, the analysis, interpretation, efficient accessibility and reuse of the library materials depend on the coherence and consistency of the descriptive metadata properties. This is essential in the

⁶<http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml>

Table 1.3.: List of the most common U.S. place names (Source: American FactFinder⁶)

Place Name	Number	Place Name	Number
Washington	32	Madison	24
Greenville	32	Georgetown	23
Franklin	31	Oxford	22
Springfield	30	Ashland	22
Clinton	29	Arlington	22
Bristol	29	Jackson	21
Salem	26	Burlington	20
Fairview	26	Milton	20

context of geographic information (Lutz, 2005; Piasecki et al., 2010). Even, Quality Assessment it is more useful when an increasing amount of the content is also available digitally as it provides geo-based ways for browsing and searching resources (Powell et al., 2010).

In summary, library users want to find, access, retrieve, visualise and interpret the geographical information, but when the inconsistencies between the semantically close geographical properties difficult these tasks, then such properties violate the principle/concept of “fitness for use or purpose” from the viewpoint of information retrieval. Therefore, high quality and consistent cataloguing and use of metadata across institutions are necessary to ensure the optimum retrieval and cross-domain searching of digital resources. As Zeng and Qin (2008) expose, with increasing demands for aggregation, metadata assessment becomes more important as quality determines the success or failure of any metadata sharing projects. In addition, without Quality Assessment, metadata creation is likely to become financially inefficient as low quality metadata leads to poor resource retrieval.

1.2.2 Spatial Data Infrastructures

Discovery, visibility and accessibility problems derived from the poor quality of the geographical information are recognised in many areas, such as hydrology (Bassoullet et al., 1986), financial and business (Herrero and Ruiz, 2008), and health and disasters medicine (Aldis et al., 2005). Also, spatial data of inadequate quality may bring some social, economic, environmental and political problems as the Figure 1.2 shows. Quality assurance information may have a real impact in the main systems or infrastructures devoted to preserve the spatial information, that is to say, the Spatial Data Infrastructures. By definition, the goal of a SDI is facilitate the availability of and access to spatial data, but this goal can be altered by the poor quality of the *geographical properties* used to access the information of SDIs. Nebert (2004) provides the following definition for SDI:

“The relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data. The SDI provides a basis for spatial data discovery, evaluation, and application for users and providers within all levels of government, the commercial sector, the non-profit sector, academia and by citizens in general”.

The importance of the Quality Assessment for the geospatial information in SDI, specifically regarding the metadata standards, is recognised in the work of Wang (2008). He enumerated a list of international standardization bodies and working groups (ISO/TC 211, Federal Geographic Data Committee⁷ (FGDC), Open Geospatial Consortium⁸ (OGC)) that address spatial data quality issues. They have identified several main components of spatial data quality, which consists of seven usual quality elements: lineage, positional accuracy, attribute accuracy, semantic accuracy, temporal accuracy, completeness and logical consistency. Particularly, spatial data consistency as one component of data quality is considered as an indispensable part in an ISO metadata model.

For example, as Mäs (2009) argues, in Europe these issues have already led to political consequences: in 2001 the European Commission initiated the

⁷<https://www.fgdc.gov/>

⁸<http://www.opengeospatial.org/>

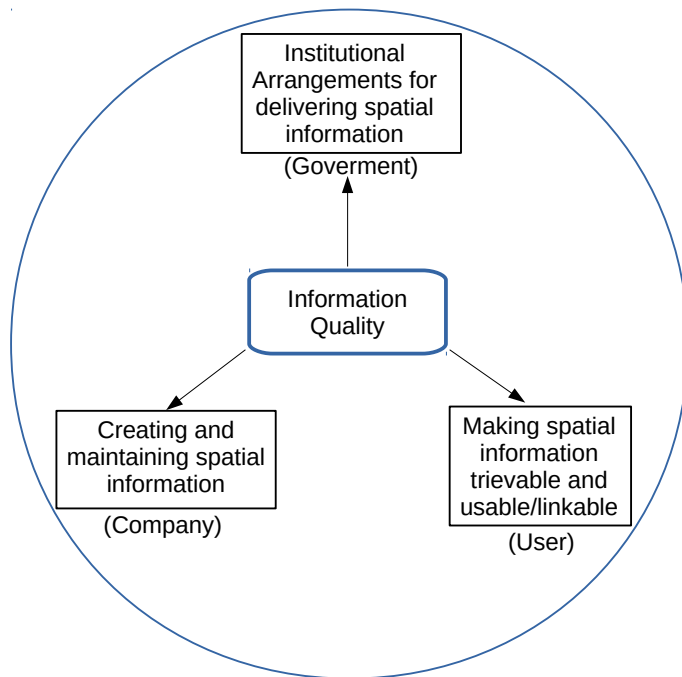


Figure 1.2.: Impact of the quality assurance information in the main actors of SDIs.

Infrastructure for Spatial Information in Europe⁹ (INSPIRE) and in 2007 the INSPIRE directive became effective. It aims at the implementation of a European wide SDI. Some main principles of INSPIRE are the provision of access to relevant, harmonised and quality Geographic Information (GI) and the support to seamlessly combining spatial information from different sources for the formulation, implementation and evaluation of EU policies. There is limited value in improving access to and sharing of geospatial information across the Web if the information quality is unknown or assumed to be assured (Sanderson et al., 2009). This makes obvious, that information quality and especially consistency are important aspects to enable such interopera-

⁹<http://inspire.ec.europa.eu/>

ble information exchange among different systems. The intensive standardisation works on handling quality within the geospatial domain confirm this perspective.

It is worth mentioning, in the line of this perspective the adoption of open OGC standards for the implementation of Geospatial Web services in SDI environment has favoured the development of a public, open and interoperable Geospatial Web (Nebert, 2004). The Geospatial Web rests on open and proprietary Web service interfaces. OGC leads the development of open specifications and interfaces to access geospatial information since 1994. This organisation has produced a set of specifications that define Web services interfaces with specific functionality. These interfaces share a common operation (*GetCapabilities*) that returns technical (e.g., the allowed literals in a parameter) and functional metadata (e.g., the description of the data it operates on) encoded in a XML document (Whiteside and Greenwood, 2010). Some of the most relevant OGC Web service specifications are for catalogue services (Catalogue Service for the Web¹⁰, CSW), map portrayal services (Web Map Service¹¹, WMS), data download services (Web Feature Service¹², WFS), and services for access to sensor data (Sensor Observation Service¹³, SOS). The Table 1.4 summarise the main characteristics and operation of the OGC Web Service interface specifications relevant in this work.

López-Pellicer (2011) describes the Geospatial Web as “*the collection of Web services, geospatial data and metadata that supports the use of geospatial data in a range of domain applications*”. In the present thesis, Web services, geospatial data and metadata are called resources. Such definition embraces a variety of resources that bear geographic information (Goodchild and Zhou, 2003). These resources may be encoded using open standards, closed standards and proprietary formats. These resources include online systems which support discovery, retrieval, storing, analysing, managing, and presenting data with geospatial dimensions (geographical properties). Some of these are *Semantically Close Geographical Properties*. SDI resources with these Semantically Close Geographical Properties are the focus of this

¹⁰<http://www.opengeospatial.org/standards/cat>

¹¹<http://www.opengeospatial.org/standards/wms>

¹²<http://www.opengeospatial.org/standards/wfs>

¹³<http://www.opengeospatial.org/standards/sos>

thesis. The consistency between these properties need to be assessed and ensure in order to access the described SDI resources. Inconsistencies between these properties can generate problems of discovery and retrieval, these problems can cause the omission/invisibility of relevant results that could fit the user needs better. In this line, Oort (2005) identifies several reasons for concerns about Geographic Information Retrieval and spatial quality issues, as follows:

- There is an increasing availability, exchange and use of spatial data.
- There is a growing group of users less aware of spatial data quality.
- Geographic Information Systems enable the use of spatial data in all sorts of applications, regardless of the appropriateness with regard to data quality.
- Current GIS offer hardly any tools for handling spatial quality.
- There is an increasing distance between those who use/access/query the spatial data (the end users) and those who are best informed about the quality of the spatial data (the producers).

Table 1.4.: The main characteristics and operation of OGC Web Services
(Source: Florczyk (2012))

Service Specification [Operation]	Brief Description
Web Catalogue Service (CSW) [DescribeRecord] [GetRecords]	It supports the ability to publish and search collections of descriptive information (metadata) of data, services, and related resources. It allows to discover elements of the supported data model. It allows discovering resources with possibility to apply spatio-temporal constraints.
Web Map Service (WMS) [GetCapabilities] [GetMap]	It produces dynamically maps of spatially referenced data from geographic information. It enumerates layers that might be rendered and supported parameters (e.g. graphic format). It produces maps.
Web Coverage Service (WCS) [GetCapabilities] [DescribeCoverage] [GetCoverage]	It supports electronic interchange of coverages (values or properties of a set of geographic locations) that represents space-varying phenomena. It enumerates coverages that might be rendered and supported parameters. It provides a full description of a coverage. It returns a coverage.
Web Feature Service (WFS) [GetCapabilities] [DescribeFeatureType] [GetFeature]	It allows direct fine-grained access to geographic info. at the feature and feature property level. It lists the features that might be requested. It returns a schema description of the requested feature. It operation returns a document that contains selection of features (retrieved from a relatively static data store), which satisfy the query specified in the request.
Web Processing Service (WPS) [GetCapabilities] [DescribeProcess] [Execute]	It allows invoke processing functionality at the feature and feature property level. It lists the processes that might be executed. It returns the description of the requested process. It executes requested process.

1.2.3 Quality Assessment on Geographical properties

Heywood et al. (1998) affirms that two factors are relevant in addressing quality and error issues: first, the terminology used for describing problems, and second, the sources, propagation and management of errors. However, Duckham and Drummond (2000) note that an obvious criticism about many spatial data quality standards and research is that these focus only on the storage, management and propagation of data quality information rather than *how to use* such information. We need a more specific way of identifying, measuring and correcting the quality problems found. For example, by identifying inconsistencies between the semantically close geographical properties used to access and retrieve the resources.

In the field of Geographic Information Retrieval the Spatial Ranking can provide a graded way to assess the quality for geographic information. The spatial ranking uses scoring functions that can measure the degree with which the available information satisfies the user needs. A generalised and widely adopted perspective about the spatial data quality is the concept of "fitness for use or purpose" (Juran, 1962), that is to say, to determine if a data is proper for the needs of a particular application or user. In conclusion, in order to assess the quality of resources, we need to have information about the resource to be used as well as the actual user need. (Wang and Strong, 1996; Veregin, 1999; Juran, 1999). At this point, data providers should supply enough information about the quality of a data set to help a data users make a proper decision in a particular situation (Chrisman, 1991). To meet "fitness for use", the producer's role has shifted to data quality documentation or "truth-in-labeling". According to the truth-in-labeling paradigm, errors are inevitable and the data quality problem results from incomplete knowledge of data limitations (Veregin, 1999). The errors are not just a bad thing, but an inevitable thing. The errors are another attribute of the spatial data. Thus, the characteristics of the errors in spatial data should be clarified/measured so that good quality results and output can be produced. For instance, the spatial ranking techniques can help to measure the level of error in the spatial data, it is done by means of the level of spatial matching between the required and available information.

Two perspectives of information quality arise here. From the viewpoint of information consumer, Wang and Strong (1996) define the information quality as the information that is fitness for use by information consumers. They argue that ultimately it is the consumer who will judge whether or not an information product is fitness for use. However, information consumers are not very capable of finding errors in information and altering the way they use the information (Klein et al., 1997).

From the data perspective, information quality can be defined as the information that meets the requirements or specifications (Kahn et al., 2002). By combining the two perspectives, Redman (2001) points out that information is of high quality if it is free of defects, inconsistencies and possesses desired features. In the context of GIR, it is when the semantically close geographical properties have a high level of geospatial semantic similarity (Renteria-Agualimpia et al., 2014).

1.3 Research Hypothesis

With the increasing volume of geospatial information and the increasing use of such information across ever more heterogeneous user groups and domains, the need to assess the “fitness for use” becomes ever more complicated (Triglav et al., 2011). Unfortunately, as more information becomes available for use, it becomes increasingly difficult to identify quality problems including, from inconsistencies to “garbage”. Garbage in the sense of the popular computing saying “garbage in garbage out”. The most of the final processes and decisions (e.g., analysis, re-use, preservation, access and retrieval) depend on the Information Quality. Thus, assess the quality is a priority task in many domains and it is the basis for the problem statement of this thesis. There is a hypothesis on retrieval improvement within the digital collections in the context of GIR, particularly, in SDI and Digital Libraries that this thesis addresses.

“In order to improve the accessibility, retrieval, and visualization for geospatial information resources in the context of digital repositories in general, it

is necessary to develop systems which are able to take advantages of the semantically close geographical properties to provide Quality Assessment in a semi-automatic manner."

From this hypothesis come the following research questions:

How geospatial inconsistencies in semantically close properties affect the discovery, accessibility and retrieval for geospatial information resources in the context of digital repositories?

Given a large collection of geospatial metadata, is it possible for a system to detect semi-automatically geospatial inconsistencies in semantically close properties using geospatial clustering?

Under which conditions, is it possible for a system to fix semi-automatically such inconsistencies?

1.4 Methodology

A systematic methodology is proposed in order to provide a context of the issues related with the inconsistency of the geospatial resources and its consequences, and the approach proposed to mitigate this issue and its consequences. The methodology comprises aspects related with software engineering, knowledge engineering and artificial intelligence.

The following methodology is applied in each of the research modules related with the problems identified in the previous sections. It starts with the analysis of the problem. The solution of the problems is the result of a cyclic incremental development process, which is decomposed in problem specification, conceptualization, implementation and evaluation.

1. **Analysis.** The analysis process reviews the existing research literature related with the problem that research module addresses.
2. **Problem specification.** The specification provides a rationale of the motivations or the challenges for the research question.

3. **Conceptualization.** The conceptualization process structures a solution identifying its key elements.
4. **Implementation.** The implementation activity develops a software platform based on the conceptualization.
5. **Evaluation.** The evaluation applies the implementation to two concrete problems in two different scenarios and evaluates its usefulness.

1.5 Scope

This research work has the following scope:

- **Knowledge representation.** In this work, several knowledge representation systems have been used. These systems are mainly simple SKOS (Simple Knowledge Organisation System (Isaac and Summers, 2009)) vocabularies and RDF/XML graphs. The KOS used in the tasks of spatial ranking must cover the geographical extent of the assessed geographical resources. Also, the level of granularity of the spatial footprints in the KOS must be in accordance with those in the analysed collection. In this thesis we restrict ourself to work with KOS with footprint of two-dimensions (2D). However, by means of simple processes the 2D footprints can be simplified to classical points (1D). The open research line and the challenge is to shift from 1D to 2D geographical footprint to assess their quality. Also, we restrict our analysis to two geographic areas, the scope of the first one is Spain and the scope of the second one is Unites Sates of America.
- **Digital Library metadata.** We focus our attention on assessing the quality of a pair of semantically close geographical properties. We do not use other fields such as the geographic area code or the place of production/creation. However our methodology could be applied to assess the quality of these properties too. It is part of the future work.
- **SDI resources.** In general, SDI resources that follow OGC specification have been in the focus of this work. We apply our Quality

Assessment analysis on Web Map Services. However our developed systems can be applied to other kind of Web services and SDI resources with semantically close geographical properties. Although, our methodology could be used to assess the quality of thematic and temporal properties, we will solely consider semantically close geographical properties. And specifically, the two most frequent semantically close geographical properties that we found analysing the experimental datasets: the Direct Spatial References (geographical extent/spatial footprint) and the Indirect Spatial References (place name). In this thesis we do not taking into account other geographical properties.

The SDI scenario is used as test case to validate the architecture, due to the descriptive information of the spatial resources in SDI is in a structured way and proceeds from experts of the geographical domains. This make us to think that the provided spatial description must be better than other scenarios and domains where the descriptions proceed from unstructured information and non-experts in the geographical domain, that is to say, domains with most spatial ambiguity, such as the Digital Library domains. In this sense, the methodology and the architecture are tested with a more difficult case, a collection from a well known Digital Library internationally.

- **Methodology.** The problems of spatial inconsistency detection of semantically close geographical properties can be addressed by two ways: (1) a methodology based on comparing geospatial coordinates, and (2) a methodology based on comparing sets of place names. In the first one, the place names are transformed into geospatial coordinates by a geocoder, then the consistency is measured by comparing the spatial similarity between the two spatial footprints. In the second one, the explicit geospatial coordinates are transformed in place names by a reverse geocoder, then the consistency is measured by comparing the similarity between the two sets of place names.

The first one is a problem widely discussed in the literature and practically solved.(Janée, 2003; Li and Fonseca, 2006; Frontiera et al., 2008). To compare the co-occurrence of two points, or two MBBoxes is not

complex, the main problem of this approach consisting on disambiguate first the place name. When a geocoding process is applied to a place name, many possibilities may exist. To decide the appropriate geospatial coordinates is the main issue that should be resolved, for example, in North America there are at least 30 places called Springfield, Franklin or Washington. Then, rather than detecting an inconsistency, this first approach could increase the inconsistencies because many times we do not have more information to disambiguate the place name.

Our research work is focused on the second approach. This approach provides a more secure way to detect inconsistencies than the first one. The transformation of spatial footprints into place names provides a more secure and accurate approach to detect inconsistencies because the spatial coordinates are unique on the earth. If exist a spatial inconsistency, it will be more likely to detect it by means of this second approach.

- **Evaluation method.** Research literature on evaluation of Information Retrieval (IR) systems identifies four main theories about IR systems (Järvelin, 2011): ranking theory, search theory, information access theory, and information interaction theory. Figure 1.3 presents them as part of the nested evaluation frameworks. In this work, only the two first theories (ranking and search) are considered. The algorithms and methods developed are evaluated using the ranking and search theories.

1.6 Contributions

This work aims to develop a methodology to assess the quality of the semantically close geographical properties. This methodology is used to show the need for systems, methods and tools that analyse the geospatial semantic consistency of these properties in order to improve the discovery, accessing, retrieval and visualization tasks of geographical information from different perspectives, in particular, from the SDI and Digital Library perspec-

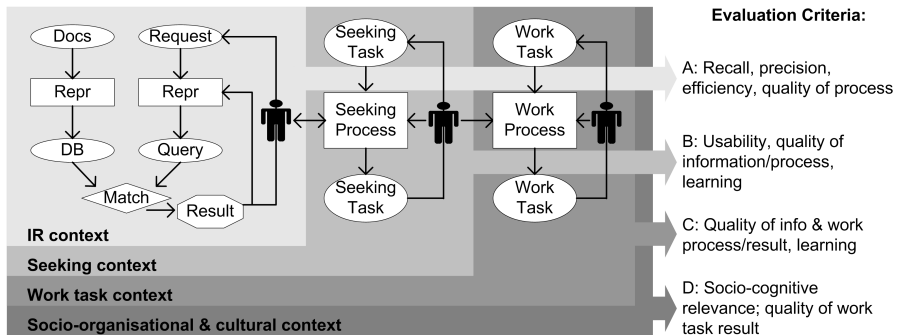


Figure 1.3.: Nested evaluation frameworks of an IR system (source: Järvelin (2011)).

tive. Starting from this aim, the main contributions of this thesis are the following:

- First, this thesis presents a study of the consistency of semantically close geographical properties from the point of view of GIR.
- Second, this thesis develops a methodology that enables the discovery of inconsistencies and systematic spatial errors.
- Third, this thesis describes a characterization of the most popular problems of geospatial inconsistencies in Digital Library and SDI.
- Fourth, this thesis develops a Quality Assessment tool for tasks related to the preservation/curation of geospatial metadata records.

1.7 Organisation of the Dissertation

Chapter 2 presents a survey of the related literature. This chapter discusses related work on exploiting location, and on assessing the quality of spatial information in the context of Digital Libraries and Spatial Data Infrastructure

Chapter 3 presents a methodology for the semi-automatic inconsistency detection between the Semantically Close Geographical Properties of spatial resources. First, the methodology uses *Knowledge Organisation Systems* combined with *geospatial ranking functions* for finding the most relevant toponyms associated with a footprint, and then it compares them with the explicit place names in the resource description. The methodology integrates these ideas with the concepts of *two-dimensional spatial clustering* to refine the detection of spatial inconsistencies and potential disagreements with the co-occurring resources. A method for spatial enrichment of spatial resources is proposed as well.

Chapter 4 presents an architecture which is intended to support Quality Assessment for spatial resources. Spatial inconsistency detection for spatial resources is discussed in more detail from the SDI community perspective, and the issue of invisibility and retrieval for these metadata resources is treated as a starting point. The architecture is implemented and tested by three different systems. Their implementation are contrasted using a dataset of Web services from the Spanish SDI. The results of the SDI scenario constitute a validation of the architecture to be extended to the Digital Libraries scenario.

Chapter 5 presents an empirical and quantitative study of the spatial quality of the semantically close geographical properties in the context of Digital Libraries. Moreover, the empirical study provides an overview of the characteristics of the common errors in metadata resources published, and reveals common errors in the current practices in the Digital Library community in the provision of metadata for cartographic resources.

Finally, Chapter 6 summarises the central contributions of this thesis, discusses our conclusions and possibilities for future work in this area.

The principal summary point to make is that the major problems in future information systems will resolve around the processes of reducing the amount of and raising the quality of information brought to the attention of the user.

Robert S. Taylor, 1986

Chapter 2

Related Work

2.1 Introduction

Because of the increasing use of decentrally held data and networked services, detailed knowledge about the meta-information quality becomes more and more important. The availability of such meta-information (descriptive metadata properties used for access to the resource) and the right evaluation of the fitness for use based on these meta-information are vital. Future information systems will resolve around the processes of reducing the amount of and raising the quality of information brought to the attention of the user needs (Taylor, 1986).

Spatial data quality deals with all quality aspects relating to geospatial data; however the information Quality Assessment involves measuring the quality dimensions that are relevant to the information consumer and comparing the resulting scores with the information consumer's quality requirements. The differences between the data quality and information quality problems are summarise in the Table 2.1. The key differences lie in that data quality problems can often be resolved through data cleansing algorithms, but information quality problems require fundamental analysis of business issues, a change in work community practices and process redesign, an analysis of the involved information community and its expectation and skills, an evaluation of the relevant knowledge domains and their attributes, as well as a rating of the content management process and infrastructure. Typical

remedies for information quality problems may include design guidelines, publishing policies, authoring training, source validation rules, the purchase of additional information services and infrastructures, a re-design of the review, assessment and feedback processes, etc.

Table 2.1.: Summary of the main differences between Data quality problems and Information quality problems (source: Eppler (2006))

Data quality problems	Information quality problems
Duplicates data relationships	Conflict recommendations in a study
Missing data relationships	Unclear causal effects in a diagnosis
Garbling (meaningless entries)	Wordy reports that have no logical flow
Spelling Errors	Untidy language that contains grammatical errors
Obsolete or outdated entries	An analysis is not updated according to recent discoveries or changes in the organizational structures
Inconsistent data formats or naming conventions	Inconsistent layout/navigation structures
Misplaced data that is saved in the wrong database	Lost 'buried' documents
Complicated query procedures	Difficult information navigation/retrieval
Wrong data coding or tagging	Inadequate or insufficient categorization (insufficient meta-information or context attributes)
Incorrect data entries because of lack of source validation	Unsubstantiated conclusions with inadequate evidence
Manipulation of stored data	Manipulation of decision processes

Information Quality Assessment is correctly considered difficult (Naumann, 2002). A general criticism within the information quality research field is that, despite the sizeable body of literature on conceptualizing information quality, relatively few researchers have tackled the problem of quantifying information quality dimensions (Naumann and Rolker, 2000; Knight and Burn, 2005).

Quality is defined as “totality of characteristics of a product that bear on its ability to satisfy stated and implied needs” (ISO/TC 211, 2002). Other definition of quality given by ISO is the “degree to which a set of inherent characteristics fulfils requirements”, where the requirement means “need or expectation that is stated, generally implied or obligatory” (ISO, 2005). In this work, the quality factors or dimensions of a geospatial resource are considered. Therefore, the recommendations from the geospatial community should be considered. Additionally, some specific issues related to the spatial Quality Assessment process in the context of SDI and DL will be discussed.

2.2 Quality Assessment in SDI

There are intensive standardisation works on handling quality within the geospatial domain. The ISO standards, for example, provide quality principles and define specific concepts (ISO 19113 (ISO/TC 211, 2002)), define principles for quality evaluation (ISO 19114 (ISO/TC 211, 2003)), and provide description of Quality Assessment methodologies (ISO 19138 (ISO/TC 211, 2006)). As for data quality, ISO 19157 standard revises ISO 19113, ISO 19114 and ISO 19138, and defines a set of measures for the spatial data quality elements identified in ISO 19113. Sharing and reusing spatial data require paying special attention to quality of spatial data; therefore this issue is relevant in any SDI. Quality has to be considered from different perspectives in a SDI. There might be different viewpoints used to describe quality (Garvin, 1988; Jakobsson, 2006). In his thesis, Jakobsson (2006) argues that quality management viewpoints are important from the SDI perspective. He discusses geographic quality concepts using these four viewpoints (Figure 2.1):

- **Production-centred viewpoint.** This perspective focuses on the variations in the production process where the most common measure is the number of defective or non-conforming products.
- **Planning-centred viewpoint.** This perspective is centred on the characteristics of products.

- **Customer-centred viewpoint.** This perspective focuses on the value of products and services to the customer.
- **System-centred viewpoint.** This perspective takes into account all stakeholders who are influenced by the organisation or its products oriented quality.

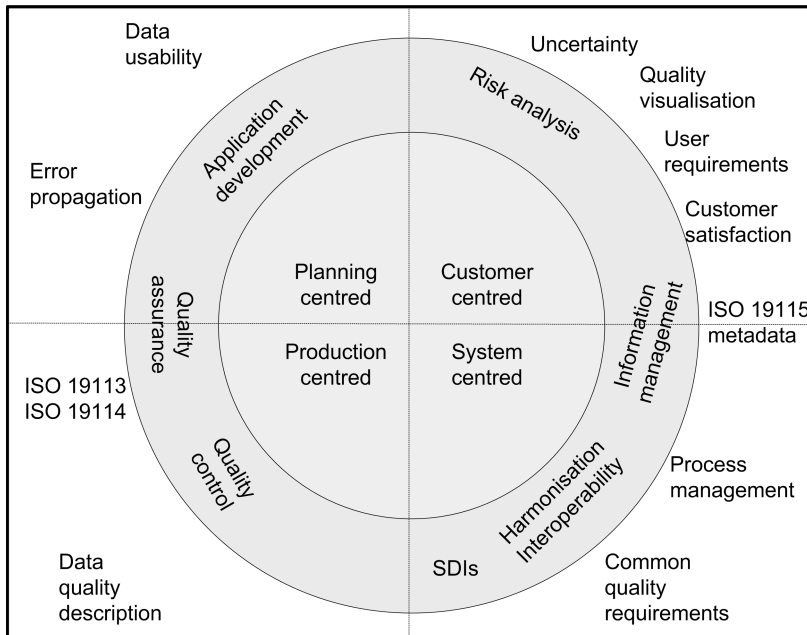


Figure 2.1.: Different approaches to geographic information quality from the quality management viewpoint (source: Jakobsson (2006)).

Data quality is also considered by the FGDC CSDGM standard (which incorporates the Spatial Data Transfer Standard (SDTS) (Moellering and Hogan, 1997) that contains a section on spatial data quality elements, parameters or dimensions) and the NAP standard. There is a remarkable agreement among the documents on the elements of spatial quality. Each of the standards that approach that question describes the same core elements:

- **Attribute (Thematic) Accuracy.** CSDGM and SDTS use term “At-

tribute Accuracy", and ISO 19115 refer to the same content as "Thematic Accuracy". It can be defined as "an assessment of the accuracy of the identification of entities and assignment of attribute values in the data set." (MDWG, 1998).

- **Completeness.** It refers to "information about omissions, selection criteria, generalization, definitions used, and other rules used to derive the data set" (MDWG, 1998).
- **Lineage.** It refers to the "information about the events, parameters, and source data which constructed the data set, and information about the responsible parties" (MDWG, 1998).
- **Logical Consistency.** It refers to "an explanation of the fidelity of relationships in the data set and tests used." (MDWG, 1998).
- **Positional Accuracy.** It is "an assessment of the accuracy of the positions of spatial objects" (MDWG, 1998).
- **Temporal Accuracy.** It is usually defined as "accuracy of the temporal attributes and temporal relationships of features" (ISO/TC 211, 2002).
- **Semantic Accuracy.** It is the "degree to which the same data values or elements/property are chosen for representing similar concepts in resource description" (Park, 2009).

Stvilia et al. (2004) group the information quality parameters in three non-exclusive dimensions: Intrinsic, Relational/Contextual and Reputational. They use this quality dimensions to develop a framework that describes 32 parameters in total. Bruce and Hillmann (2004) offer an useful examination of characteristics of metadata quality and they outline seven general characteristics of metadata quality: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. As cite Shreeves et al. (2005), they offer some possible criteria and compliance indicators for each, noting that shared metadata may require additional quality efforts. Further, it is important to note that Bruce and Hillmann (2004) devised their framework to guide human reviewers. Shreeves et al. (2005) make a mapping between the Bruce and Hillman

framework and the Gasser and Stvilia framework. The results of the comparison is shown in the Figure 2.2.

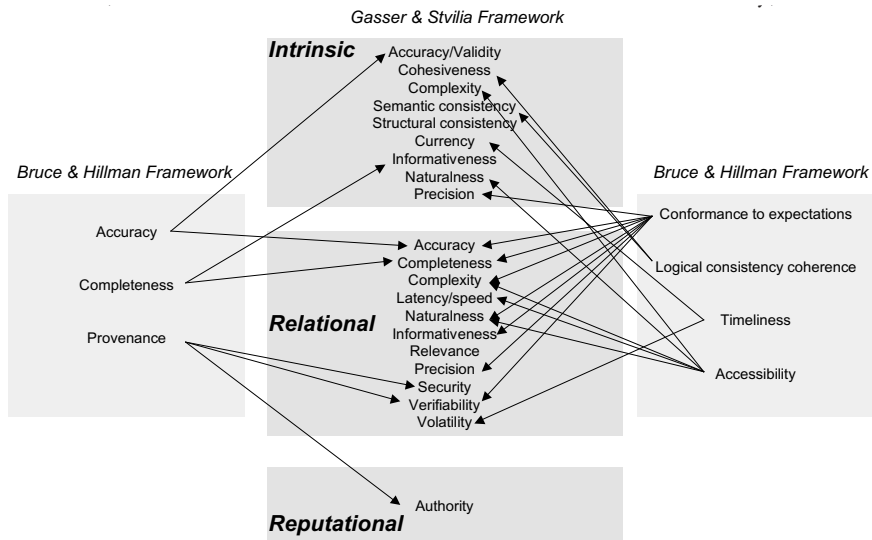


Figure 2.2.: Mapping between the Bruce and Hillman framework and the Gasser and Stvilia framework (Source: Shreeves et al. (2005)).

The impact of geospatial data and information quality problems is widely recognized (Longhorn, 2005; INSPIRE, 2013; Nogueras-Iso et al., 2005). The works of Longhorn and Nogueras-Iso et al. point out that a collection, a Spatial Data Infrastructure, or in general a repository need to ensure meta-data consistency in order to be an effective way for sharing information. For example, a user would expect in a catalogue of Geospatial Web Services (GWS) that map servers about Spain should be returned by textual queries containing the term "Spain", by spatial queries with the Minimum Bounding Box (MBBox) of Spain or by both. If the term "Spain" is not present in the metadata of a Spanish map service, the textual query will fail. If the MBBox is not present, or it is overly simplified or misleading, such resource will be hidden for trivial spatial queries. Nowadays, such problems are quite common in catalogues. For example, a real case of a problem of inconsistency is depicted in Figure 2.3. We formulated a query about Zaragoza in the area

of Colombia (South America), but the answer is a service that only returns place names in Spain. Other Web portals using this dual way to search are illustrated in the Figures 2.4 and 2.5.

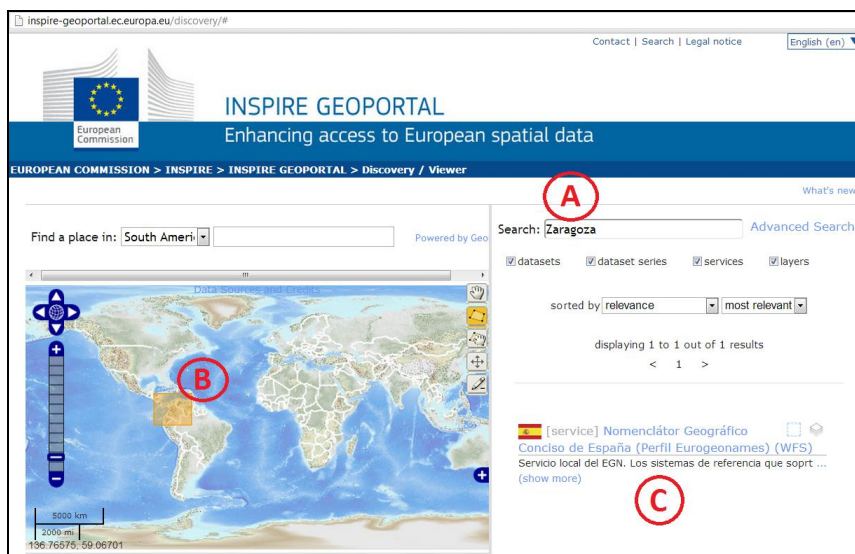


Figure 2.3.: INSPIRE geoportal problem.

In the case of INSPIRE, Article 17 of the INSPIRE Directive (Commission of the European Communities (CEC), 2007) says: “Each Member State shall adopt measures for the sharing of spatial data sets between its public authorities... for the purposes of public tasks that may have an impact on the environment.” Therefore, an effort within INSPIRE is dedicated to the development of some methods for assessing, measuring, reporting and controlling spatial quality. These aspects have been considered by the European Spatial Data Infrastructure with a Best Practice Network (ESDIN) project¹ supported by eContent+ programme.

Also the OGC, as a major standardisation body for GI, has paid attention to establish forums for describing an interoperable framework or model for OGC Quality Assurance measures and Web Services to enable access and

¹<http://www.esdin.eu/>

The screenshot shows the search interface for the NSGC catalogs. At the top, there is a navigation bar with the text "NOVA SCOTIA CANADA" and "Government of Nova Scotia | gov.ns.ca". Below this is a map of Nova Scotia. The search interface is divided into three main sections:

- Search by geography?**: This section includes a radio button for "Entire Province" and a text input for "locate your area here". Below this is a map of Nova Scotia with a red arrow pointing to it. A "Find Nova Scotia Place" button is located below the map.
- Search by subject?**: This section includes a "Full Text Search:" input field, a "Match:" section with radio buttons for "Any word", "Exact phrase", and "All words", and a "Keywords:" list with an "Add" button. The keywords list includes: aboiteaux, abundance, accommodation, acid, acreage, activities, and activity. There is also an "Organization:" input field with a "Find" button.
- Search by date?**: This section includes radio buttons for "Anytime", "Time Period for Data (yyymmdd)", and "Last published date (yyymmdd)". It also has "From:" and "To:" input fields, and an "After:" input field.

At the bottom left, there is a "Scale/Coverage:" dropdown menu.

Figure 2.4.: Example of a search in the NSGC catalogs.

sharing of high quality geospatial information, improve data analysis and ultimately influence policy decisions.

Wang (2008) compares spatial data quality elements from different standards, he concludes that ISO/TC 211 Standard 19113 defines a comprehensive one. However, ISO/TC 211 Standard 19113 does not give detailed explanation of the meaning and says nothing of how to apply them into the GIS applications, geographical information system or geographical data mining tasks in general. Therefore, in order to analysis the consistency problems in geographic information, deep and focused studies of the Quality Assessment for the geographical information need to be done.

www.mirame.chduero.es/dueroCatalog/srv/es/main/home

GOBIERNO DE ESPAÑA | MINISTERIO DE MEDIO AMBIENTE Y MEDIO RURAL Y MARINO | CONFEDERACIÓN HIDROGRÁFICA DEL DUERO

Inicio | Contáctenos | Enlaces | Acerca de | Ayuda | English | Español

Usuario Contraseña Iniciar Sesión

BÚSQUEDA DE SERVIDORES DE MAPAS, INFORMACIÓN GEOGRÁFICA, IMÁGENES DE SATÉLITES Y OTROS TIPOS DE RECURSOS.

¿Qué? (Texto)
País/Región

Los resultados de la búsqueda son: 1-10/11 (page 1/2), Ordenar por Popularidad

Lago CENTRALES TERMICAS

Resumen Capa que contiene las centrales térmicas de la parte española de la demarcación hidrográfica del Duero. A través del servicio WFS se puede descargar la geometría y los siguientes atributos: ...

Palabras clave World

[Página de Metadatos](#)

Figure 2.5.: Example of a search in the CHDuero catalogs.

In other work, Wang (2008) focuses his quality research in the spatial data consistency, it refers to the logical consistency containing semantic and temporal information. His logical consistency deals with logical rules of structure and properties for spatial data and describes the compatibility between dataset items, for example the topological relations. Semantic information indicates the pertinence of the meaning of the geographical object rather than the geometrical representation. Temporal information includes temporal attributes and temporal relationships of features. He argues that Spatial data integrity rules should consider not only the logical consistency, but also semantic information. Practical GIS applications often encounter these aspects. Moreover, different semantic notes of geospatial objects can change the meanings of spatial data integrity rules, for example, in a two dimensional map, two lineString feature types with the different semantic meaning denotes the different integrity rules: “a road is not allowed to intersect with a lake, but a bridge can be authorized to intersect with a lake”. Therewith, different kind of semantic aspects are necessary to be taken into account when investigating spatial data integrity rules.

Veregin (1999) refers to the consistency as the absence of apparent contradictions in a database. For geospatial data the term is used primarily to specify conformance with certain topological rules (Kainz, 1995). In this sense, topological consistency is one aspect of consistency in the spatial domain. Spatial inconsistencies can also be identified through redundancies in spatial attributes. For example, an entity might have the value 'Delaware' for the attribute 'state' but the value 'Lincoln' for the attribute 'county'. This is inconsistent since there is no Lincoln county in Delaware. The work of Veregin does not deal with the quality problems of the geospatial properties used to access and retrieve the database and repository resources.

Idowu and Sambo (2012) present a computer tool to determine the quality of geospatial data using horizontal coordinates of points from satellite image and large scale cadastral maps of a study area. Their spatial data Quality Assessment tool involved the use of statistical models to determine the Root Mean Square Error between the satellite image coordinates against coordinates of the same points obtained from large scale cadastral maps. Test of hypothesis was carried out using Chi-Square statistic at 95% confidence level to ascertain conformity of the variance of the satellite data with the variance obtained for large scale maps. The analysis of the results proved the efficacy of the developed tool in assessing the quality of geospatial data to determine whether or not the geospatial data are useful for further applications in GIS environment. In other words, it has been demonstrated, in their study, that quality of data acquisition in Geo-informatics could be controlled. However, the problem of identifying existing inconsistencies between the different representation of the same information in the same source is not mentioned, also the problems of inconsistencies between semantically close properties in two-dimensional dataset is not covered.

2.3 QA in Digital Libraries

A growing number of Digital Library projects are working with georeferenced data and metadata to take advantage of the ubiquity and popularity of geographic services widely available, as shows the Figure 2.6. The links in the graph represent the new Digital Library works that recognize and cite

previous works related to use of geospatial information in the DL community. An analysis of these works reveals that the most cited Digital Library projects experimenting with georeferenced data and metadata are focused on three main areas: information visualization, geographic information retrieval and information validation.

Some examples of works focused on *information visualization* are Geo-Referenced Information Network, Electronic Cultural Atlas Initiative Buchel and Hill (2010), Old Maps Online (Southall and Pridal, 2012) and the Alexandria Digital Library, probably one of the most widely cited research projects that made use of georeferencing in the context of Digital Libraries (Goodchild, 1995). They are focused on representing, exploring and browsing digital collections on a map, although some of those works also develop additional search tasks. A modern example of this kind of system is depicted in the Figure 2.7. In this example we can find an inconsistency, because the preview image shows the Curacao island in the southern Caribbean Sea, of the Venezuelan coast, but the mathematical coordinates refer to Colombia.

About the *geographic information retrieval*, they deal with the disambiguation of place names based on internal and external evidence from the text content of metadata. Internal evidence includes the use of honorifics, generic geographic labels, or linguistic environment. External evidence includes knowledge organization systems, gazetteers, biographical information, and general linguistic knowledge (Goodchild, 1995; Crane, 1998; Kanagavalli and Raja, 2013). Some works in this area are Spatially-Aware Information Retrieval on the Internet (SPIRIT) project Jones et al. (2002), GAT (Powell et al., 2010), MapRank (Oehrli et al., 2011) and Old Maps Online project (Southall and Pridal, 2012).

For example, the MapRank is a Geographical Search tool for Cartographic Materials in Libraries. It is used by the Kartenportal.ch Digital Library to supplement the text-based Conventional library catalogs OPAC² and to find cartographic material in libraries. Authors developed a cleverly devised ranking algorithm and an innovative indexing mechanism. Although MapRank is a quick and efficient tool for carrying out map searches

²<http://www.opac.net>

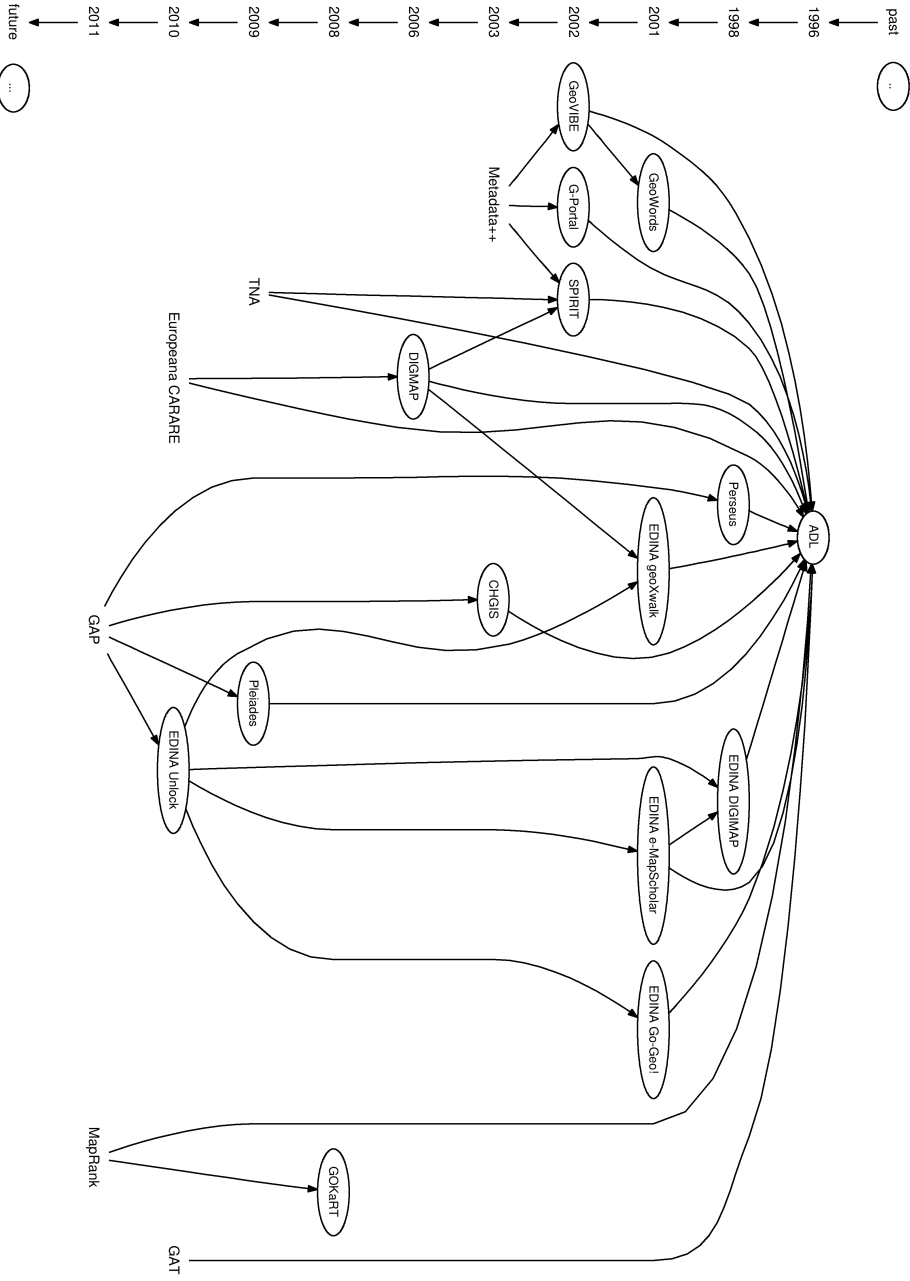


Figure 2.6.: Some of the most popular Digital Libraries working with georeferenced data and metadata

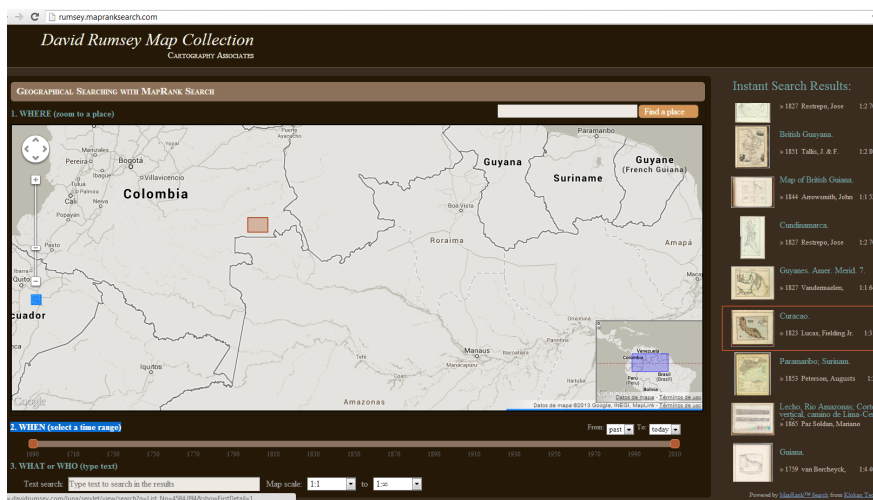


Figure 2.7.: Example of inconsistency in the David Rumsey Map Collection. The preview image shows the Curacao island in the southern Caribbean Sea, of the Venezuelan coast, but the mathematical coordinates refer to Colombia.

in meta-catalogs, we found geospatial inconsistency problem in its meta-information. Figure 2.9 and 2.8 show two examples of inconsistencies where the Direct Spatial References or footprints are inconsistent with the Indirect Spatial References or place names. The first one is an error because the Italian Somma Vesuvius volcanic complex located in the province of Nápoles is not near to the city of Barcelona in Spain. In the second one, the resource entitled “Mapa geológico de la provincia de Alava” is referring to a province of the “País Vasco” (Spain) but it is displayed among the “Province of Lleida” (Spain), France and Andorra.

With respect to *information validation*, this kind of works are focused on data and metadata quality; we center our attention on this last kind of works. Metadata quality is a semantically slippery term. Park (2009) suggests that the most commonly accepted criteria for metadata quality are completeness, accuracy, and consistence. Our work is focused on the last criterion. Relevant works in the literature during the last decades confirm this perception



Figure 2.8.: Example of an inconsistency about Italy in the KartenPortal. The resource entitled Italian Somma Vesuvius volcanic complex located in the province of Nápoles is not near to the city of Barcelona in Spain

(Moen et al., 1998; Bruce and Hillmann, 2004; Zeng et al., 2005; Shreeves et al., 2005; Shen et al., 2013)

Beall (2006) makes an interesting research, although it is not focused on the geospatial domain. His work highlights the importance of assess the quality of the metadata of Digital Libraries to ensure the accessibility and retrieval of digital resources. Beall describes the main types of data quality errors (typographical errors, scanning and data conversion errors) that occur in Digital Libraries. He argues that studying these errors is important because they can block access to online resources (Errors in metadata can also hinder access in Digital Libraries). Beall points out some types of common errors, and find and replace them. His work also discusses the responsibility for errors in digital resources and offers suggestions for managing Digital Library data quality.

Tolosana-Calasanz et al. (2006) develop a quantitative method and realize a statistical analysis for assessing the quality of geospatial metadata. The authors first formulated a list of geographic quality criteria by consulting domain experts. The identified criteria indicated tendencies of quality, The

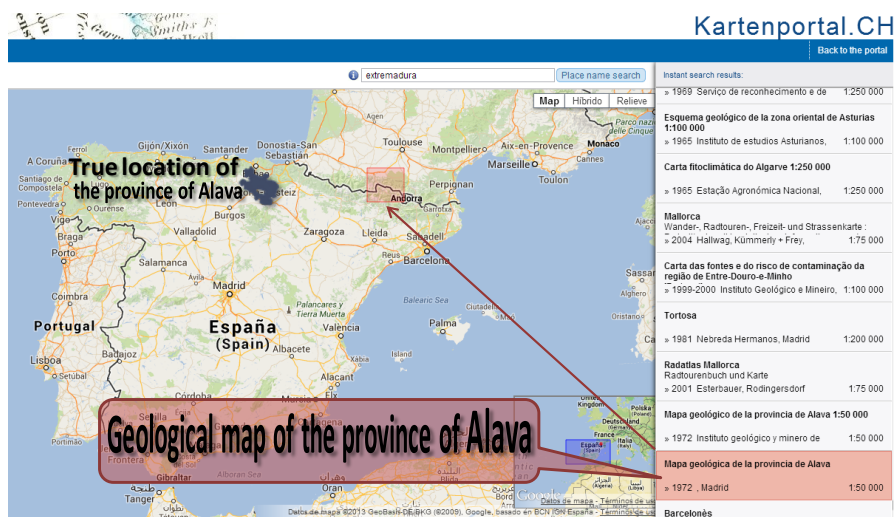


Figure 2.9.: Example of an inconsistency about Spain in the KartenPortal. The resource entitled “Mapa geológico de la provincia de Alava” is referring to a province of the “País Vasco” (Spain) but it is displayed among the “Province of Lleida” (Spain), France and Andorra.

authors also noticed the need to ensure the completeness of the spatial fields to guarantee a minimum level of quality. Their method is developed in two phases. Firstly, a list of geographic quality criteria was compiled from an opinion poll conducted to several experts of the area. The criteria were primarily classified into structural and semantic. A list of 14 geospatial metadata metrics was proposed. Secondly, a statistical analysis, Principal Component analysis, was carried out on a selection of 30 geospatial metadata record sets. Their experiment studied the relationship between the 14 metrics, which were computed for each record set, and the assessments made by some experts. As a result, it was observed that some metrics could be used as indicators of geospatial metadata quality and, within the selected 30 record sets, the geospatial metadata quality could be predicted by computing those metrics: high values of the metrics involve medium-high quality and low values of them, low quality. In a related work, Ma et al. (2009) pre-

sented a study about the Quality Assessment of metadata on the Internet Public Library. Their work is based on a combination of human evaluation (qualitative) and automatic evaluation (quantitative). This qualitative method gives an indication of the quality of information by rating accuracy, completeness, consistency and functionality.

Most of the cited works recognize in different ways the metadata quality problems, and they remark the need to span the gap between the explicit geographic information included in the metadata and the georeference information that was not explicitly labelled as such. Their main difference with our work lies in the quantitative evaluation of the problem. We present a quantitative study of the geospatial inconsistency problems in metadata focused on the libraries domain and SDI. The most cited works differ from our approach because their quantitative methods only measure the completeness of metadata in the collection, however ours is focused on evaluating the spatial consistency quantitatively, that is, we use spatial best matches for finding and measuring inconsistencies. Our Quality Assessment approach is provided to those metadata geographical fields semantically closed.

2.4 Spatial Ranking for QA

Semantic search is one of the most active field of research in Information Retrieval. Nowadays there are many efforts in SDI and Digital Library communities to develop search system integrating geospatial, semantic, and linked browsing capabilities. Renteria-Agualimpia et al. (2010) recognizes that one of the most neglected aspects of semantic search engines are the spatial properties, however, their quality assurance is a crucial aspect for the relevance of results. In this sense, the representation of the geographic data is a critical issue for indexing and retrieval Leveling (2011). According to the FGDC specifications (FGDC, 1998a), there are two main representations for geographic information, by means of (1) Indirect Spatial References (**ISR**) or by means of (2) Direct Spatial References (**DSR**). The term ISR identifies the use of geographic names (toponyms) or place names for representing geographic information. The term DSR refers to the use of geometric objects

for representing geographic information. Conceptually these two properties refer to the same real object, and it is in this conceptual sense that these two properties are *semantically close geographical properties*. The most common DSR objects are (Larson, 2011):

- Geographic point, a single latitude/longitude pair representing the centre generally.
- Minimum Bounding Rectangle or Box, a pair of latitude/longitude pairs that defines the northern, southern, east and west extremes of a geographic region.
- Complex polygon, a more accurate representation of a geographic area that may provide a faithful representation of an area's borders.

In the GIR context, the geometric objects are called geographic footprints. Many GIR techniques are based on the comparison between the footprint of the query and the resources (Hill, 2006). The literature presents three main geometric approaches to determine the spatial relevance of a resource based on:

- Topological relationships.
- Directional relationships.
- Metric characteristics.

The common principle of these approaches is the use of spatial similarity measures as a scoring function (Frontiera et al., 2008). The value of a score function is the common way to evaluate the relevancy of query results. There are three approaches to determine the score function:

Approach based on topological relationships

Geographic information retrieval works using this approach rank query results based on the spatial similarity of their footprints with respect to the query region. In this approach, the spatial similarity is focused on topological relationships such as distance, overlap, contain, nearness, and adjacency

relationships between the query and the resources. Some of the main proposals are discussed by Hill (1990), Walker et al. (1992), Beard and Sharma (1997), Janée (2003), Frontiera et al. (2008), and Martins and Calado (2010).

Approach based on directional relationships

The retrieval of geographic information makes use of a directional system, (e.g., north, northwest, west, southwest, south, southeast, east, etc) to represent directions. Then, it uses a distance function to determine a cost of transformation to establish the relevancy of resources. Some representative works are Egenhofer and Franzosa (1995), Goyal and Egenhofer (2001), Cohn and Hazarika (2001), Renz (2002) and Li and Fonseca (2006).

Approach based on metric characteristics

Main works using this approach are focused on spatial similarity measures to evaluate characteristics such as area, perimeter, length, density, dispersion, shape, among others. Some relevant works are using the Hausdorff Distance (HD) to perform retrieval based on the comparison of the shape of geographic footprints (Janée, 2003; Renteria-Agualimpia et al., 2013b). HD measures the resemblance between two point sets Q (Query) and D (Data) based on the maximum of two distances:

- The maximum distance that any point in set Q will be from the nearest point in set D .
- The maximum distance that any point in set D will be from the nearest point in set Q , where the distance metric is usually the Euclidean distance (Frontiera et al., 2008; Atallah, 1983; Huttenlocher et al., 1993).

Similarity measures have also become popular in semantic GIR systems (Egenhofer, 2002; Janowicz et al., 2011). These systems use semantics for browsing, searching and comparing concepts using Knowledge Organization Systems such as thesaurus and ontologies. The ranking in these systems is a measure of the similarity between the expected results and the

retrieved results. Numerous works in the literature combine geospatial approaches with semantic models for computing the ranking (Grütter and Bauer-Messmer, 2007; Frontiera et al., 2008; Gui et al., 2013). The scoring functions will be our way to evaluate the relevancy of query results and also the base to provide Quality Assessment for the semantically close geographical properties (DSR and ISR).

2.5 Summary

Spatial data consistency can be regarded as one of the aspects of spatial data quality as documented by the standardization bodies. Many organization, standard, and research works are focused on quality but and a general criticism within the information quality research field is that, despite the sizeable body of literature on conceptualizing information quality, relatively few researchers have tackled the problem of quantifying information quality dimensions.

Quality metadata reflect the degree to which the metadata in question perform the core bibliographic functions of discovery, access, retrieval, provenance, currency, authentication, and administration. The functional perspective is closely tied with the criteria and measurements used for assessing metadata quality. Accuracy, completeness, and consistency are the most commonly used criteria in measuring metadata quality in literature.

Although metadata guidelines and semi-automatic metadata generation tools appear to be the most frequently utilized mechanisms for quality assurance, the quality problems in the spatial dimensions (geographical properties) do not receive the required analysis.

In summary, the rapidly growing body of geographic information and digital repositories calls for further investigation of metadata quality. The identification of factors behind inconsistent metadata demand in-depth studies. Development of systems for measuring geospatial semantic consistency and for improving quality are also critical areas for further studies.

The maturity of an organization to manage the quality of its information can mean the difference between success and failure.

John R. Talbert in (Al-Hakim, 2007)

Chapter 3

Quality Assessment using two-dimensional Spatial Ranking

3.1 Introduction

This chapter presents our proposal to provide Quality Assessment by detecting inconsistencies between the semantically close geospatial properties of a resource. A general outline of the process is shown in Figure 3.1. Our methodology uses the principles proposed in Renteria-Agualimpia et al. (2013c). Its main insight is the use of *Knowledge Organization Systems* combined with *geospatial ranking functions* for finding the most relevant toponyms associated with a footprint and then compare them with the explicit place names in the resource description. We integrate this idea with the concepts of *two-dimensional spatial clustering* to refine the detection of spatial inconsistencies and potential disagreements with the co-occurring resources. The resulting methodology has six main steps:

- *Collecting*: the step to collect spatial information susceptible to be assessed (information with semantically close geographical properties)
- *Geo-Extraction*: The process of spatial information identification and extraction.
- *Reverse Geocoding*: The step to convert or transform the reference systems, from textual to mathematical specifically.

- *Geospatial Clustering*: The process to identify group according to their spatial similarities.
- *Metadata Validation*: The process of assess the quality, by comparing the semantically close geographical properties.
- *Report Generation*: The final step to summarise and inform about the Quality Assessment.

These steps are described in detail in the following subsections.

3.2 Collecting

The collecting step can be adapted to the type of accessibility of the analysed resources. Many resources are available in different ways and contexts: direct access (http-download), accessible by different protocols or technologies, etc. In this work we describe two different cases of the resource collecting applied to resources with semantically close geospatial properties.

3.2.1 Crawling

In the first case, web service metadata based on OGC standard fulfils the requirement of semantically close geospatial properties. The process of web service metadata crawling starts with a focused web crawler that crawls the Web for geospatial OGC web services and then extracts available descriptions about the services and their content (mainly: title, description, geospatial reference, keywords, creator, and date). This focused crawler uses a set of rules and heuristics that can find OGC web services that are not described in catalogues.

Figure 3.2 shows an example of WMS service capabilities crawled¹. A WMS service provides the layers contained inside (spatial resources) as map images (returned as JPEG, PNG...). The first part of the capabilities metadata

¹In this example the textual descriptions have been translated to English to facilitate its reading.

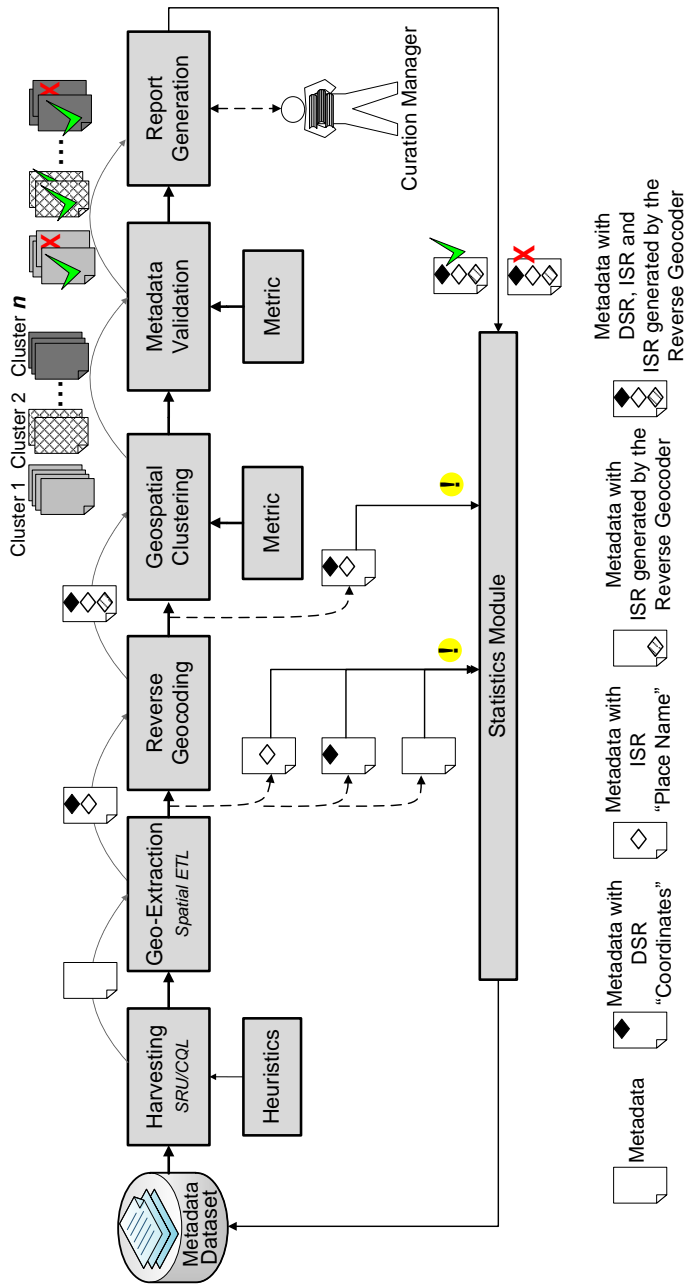


Figure 3.1.: Methodology to detect geospatial inconsistencies in metadata.

provides service details (Service tag), in this case name/title and the service URL that accepts queries. The rest (Capability tag) describes the provided resources, and the geographical extent (i.e., the Minimum Bounding Box MB-Box). The capabilities metadata in the Figure 3.2 describes a WMS service for Castilla-La Mancha region in Spain with two layers: ortophotography, and a regular grid. In this case the semantically close geospatial properties are the *name*: "Castilla-La Mancha" and the *LatLonBoundingBox*: maxx="-0.635319" maxy="40.6509" minx="-5.36468" miny="37.2951". Appendix A contains an extended example of a Web service capability document.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<Service>
  <Name>OGC:WMS</Name>
  <Title>PNOA 2006 Castilla-La Mancha 50 cm</Title>
  <OnlineResource xlink:href="http://ide.jccm.es/cgi-bin/mapserv?
    map=/usr/local/webmapping/pnoa/privado/mapserver/map/
    wms_orto.map&amp;" />
</Service>
<Capability>
  <Layer>
    <Name>PNOA_of_Castilla-La_Mancha</Name>
    <Title>PNOA 2006 Castilla-La Mancha 50 cm</Title>
    <SRS>EPSG:25830</SRS>
    <LatLonBoundingBox maxx="-0.635319" maxy="40.6509"
      minx="-5.36468" miny="37.2951" />
    <Layer cascaded="0" opaque="0" queryable="0">
      <Name>ortophoto</Name>
      <Title>Ortophoto PNOA 2006 of 0.5 m. of resolution</Title>
      <SRS>EPSG:23030</SRS>
      <ScaleHint max="49.8902848429637" min="0" />
    </Layer>
  </Layer>
</Capability>
```

Figure 3.2.: Example of WMS Capability document

To discover the OGC services available on the Web, we used the geospatial crawler proposed by López-Pellicer et al. (2012c). It first performs a set of automated queries to Bing² and Google Web Search³ APIs. The queries

²<http://api.search.live.net/json.aspx>

³<https://ajax.googleapis.com/ajax/services/search/web>

include terms associated to OGC standards (e.g., request, getCapabilities, service, endpoint, WSDL, profile ...) and geospatial tasks (e.g., coordinate transformation, interpolation, grid ...). The goal is to start the crawl from web references to OGC web services. Then, the links of the retrieved pages are examined to determine if they meet any OGC specification. If they do not meet them, they are scored to determine their exploration order. The score is computed taking into account the topic information close to the link, the parent score, and a decay value. Those links with a score greater than a selected threshold are accessed, and their content is added to the queue.

3.2.2 Harvesting

In the second case, we take advantages of the standards used in Digital Libraries community. Most libraries share their metadata using OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) and following the Z39.50⁴ protocol, and create records in MARC21 to achieve interoperability. Digital Library metadata with the MARC21 standard fulfils the requirement of semantically close geospatial properties. The process of metadata harvesting consists in collecting metadata descriptions stored in digital repositories using protocols such as OAI-PMH (Barrueco and Coll, 2003; Schindler and Diepenbroek, 2008) or Search/Retrieve via URL (SRU)⁵ In our case we are using the SRU protocol. The SRU protocol uses three types of operations: *explain*, *scan* and *searchRetrieve*. Our methodology applies the last one. The *searchRetrieve* operation allows submitting a query using the high-level Contextual Query Language known as CQL and retrieving the list of items that match the query (Denenberg, 2007). SRU has no explicit geographical information retrieval support. Hence, we formulated a heuristic based on string patterns to create queries that can retrieve metadata with information about the geographic extent of a resource, the Direct Spatial Reference specifically. This heuristic looks for metadata records that contain sub-strings that may encode geographic coordinates. For example, we formulate queries such as the URL below to retrieve metadata containing the pattern "W12*":

⁴<http://www.loc.gov/z3950/agency/>

⁵<http://www.loc.gov/standards/sru/sru-1-1.html>

http://z3950.loc.gov:7090/voyager?version=1.1&operation=searchRetrieve&maximumRecords=10&startRecord=22&recordSchema=mods&recordPacking=xml&query=W12*.

In this query, the pattern “W12*” can represent a coordinate referencing a point whose longitude is between W120 to W129 (between 120 and 129 degrees West longitude). Using this heuristic, the harvesting module retrieved all MARXML⁶ and MODS⁷ metadata records matching with this type of query in a range from 180 degrees east to 180 degrees west and from 90 degrees North to 90 degrees South. Later, the system verifies that the retrieved records contain geographic coordinates as is shown in the Figure 3.3.

```
<datafield tag="034" ind2=" " ind1="1">
  <subfield code="a">a</subfield>
  <subfield code="b">670000</subfield>
  <subfield code="d">W0830000</subfield>
  <subfield code="e">W0720000</subfield>
  <subfield code="f">N0440000</subfield>
  <subfield code="g">N0400000</subfield>
</datafield>
```

(a) MARCXML

```
<subject>
  <cartographics>
    <scale>Scale[ca.1:670,000]</scale>
    <coordinates>(W83 - -W72 / N44 - -N40)</coordinates>
  </cartographics>
</subject>
```

(b) MODS

Figure 3.3.: Example of coordinates in MARCXML and MODS formats

An example of harvested XML file is shown in the Figure 3.4. Also, Appendix B contains an extended example of the harvested XML files.

⁶<http://www.loc.gov/standards/marcxml>

⁷<http://www.loc.gov/standards/mods>

```

<?xml version="1.0"?>
<zs:searchRetrieveResponse xmlns:zs="http://www.loc.gov/zing/srw/">
<zs:version>1.1</zs:version>
<zs:numberOfRecords>3972</zs:numberOfRecords>
<zs:recordSchema>info:srw/schema/1/mods-v3.2</zs:recordSchema>
<zs:recordPacking>xml</zs:recordPacking>
<zs:recordData><mods xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.loc.gov/mods/v3" version="3.2"
xsi:schemaLocation="http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-2.xsd">
  <titleInfo>
    <title>Bottom configuration off California coast</title>
    <subTitle>depth contours at 100 fathom (600 feet) intervals</subTitle>
  </titleInfo>
  <typeOfResource>cartographic</typeOfResource>
  <genre authority="marcgt">map</genre>
  <originInfo>
    <place>
      <placeTerm type="text">California</placeTerm>
    </place>
    <dateIssued encoding="marc">1993</dateIssued>
  </originInfo>
  <physicalDescription>
    <form authority="smd">map</form> <extent>1 map ; 39 x 48 cm.</extent>
  </physicalDescription>
  <note>Depths shown by contours and soundings.</note>
  <note>Shows continental slope and oceanic basin off California coast.</note>
  <subject>
    <cartographics>
      <coordinates>W1240000 W1232000 N0384000 N0381000</coordinates>
    </cartographics>
  </subject>
  <subject authority="lcsch">
    <topic>Coasts</topic>
    <geographic>California</geographic>
    <genre>Maps</genre>
  </subject>
</mods></zs:record></zs:records></zs:searchRetrieveResponse>

```

Figure 3.4.: Example of a XML harvested from the Library of Congress (the content has been simplified for reasons of brevity).

3.3 Geo-Extraction

This step applies a second preprocessing step, the geo-extraction, to the collected and preprocessed metadata records. This step is a geospatial Extraction, Transformation and Load process (ETL) (Bédard et al., 2001). This module extracts the Direct Spatial Reference encoded in the metadata and homogenizes it. This module also extracts the Indirect Spatial Reference from textual place name fields. In MARC21 metadata, Direct Spatial Reference is the Minimum Bounding Box and it can be found in the field "*034 - Coded Cartographic Mathematical Data*" or in the coordinates field. And also, it can be found in the *LatLonBoundingBox* field of the WMS capability metadata document. A MBBBox is a pair of latitude/longitude pairs that defines the northern, southern, eastern and western extremes of a geographic region. Indirect Spatial Reference is the place name and it can be found in the field "*651 - Subject Added Entry - Geographic Name*" or sometimes it is located in the name or title field for both cases, MARC21 and WMS. MARC21 records have more fields with geographical information such as is shown in the Figure 3.5. It illustrates those MARC fields most often used in library cataloging records for georeferencing the content of the resources; for a complete view of these fields, refer to the MARC21 Web pages maintained by the Library of Congress⁸. However we selected the two more frequent field semantically closed that describe the geographical extend covered by the resource. Additionally, as Hill (2006) says, rarely are all of these fields used in a single metadata record.

The output of this process is a stream of metadata records annotated with the extracted Direct Spatial Reference (explicit MBBBox) and Indirect Spatial Reference (place name). Metadata without Indirect Spatial References, Direct Spatial References or both are counted in the Statistics Module as incomplete metadata and then they are not taken into account for further processing. Lacasta et al. (2014b) present an extended example of extraction, harmonization and cleaning focused on geospatial resources. They developed a work-flow for improving the geospatial data on the web based on

⁸<http://www.loc.gov/marc/bibliographic/>

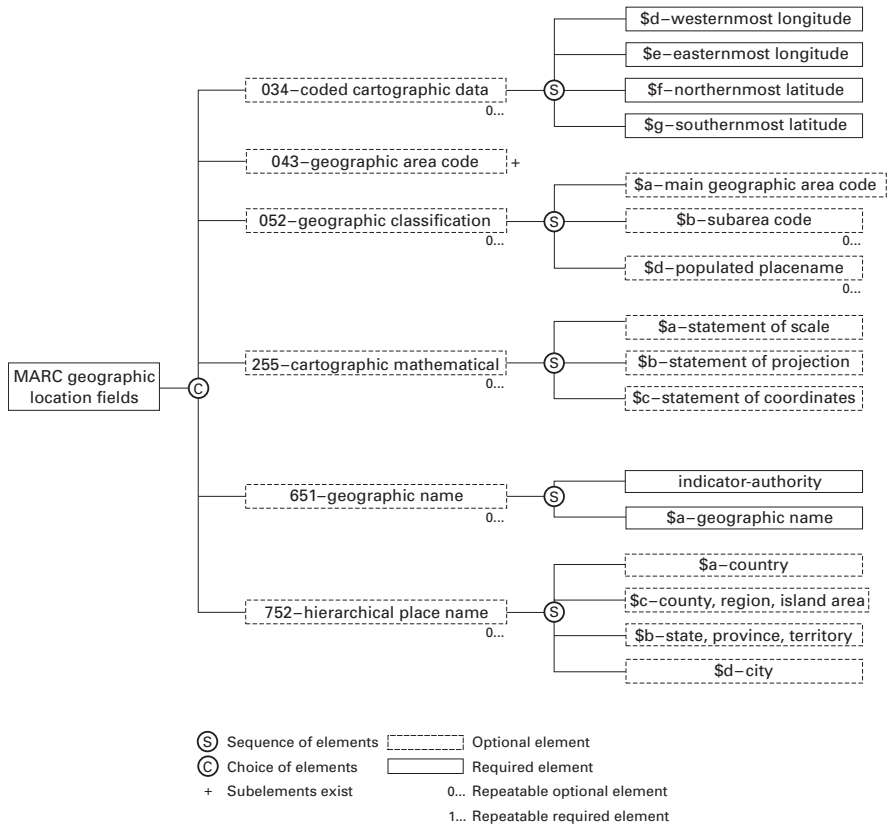


Figure 3.5.: MARC geographic location fields most often used in library cataloging (Source: Hill (2006)).

the combination of natural language processing, classification and semantic processing.

3.4 Reverse Geocoding

This step is a conversion process. It converts a reference systems based on coordinates (i.e., a footprint) into a reference systems based on geographic

identifiers.

The conceptual idea is summarized in the Figure 3.6, and a general example is shown in the Figure 3.7. The transformation process converts the ISR (the Madrid place name) into their different DSR representations (a point, a Minimum Bounding Box and a complex geometry); the inverse process is called *Reverse Geocoding*. This example is describing the main idea behind a geocoder, and illustrates the different ways of spatial representations commonly used. The right side of the Figure 3.7 shows the compression level of the information and the expressive power of the spatial representation.

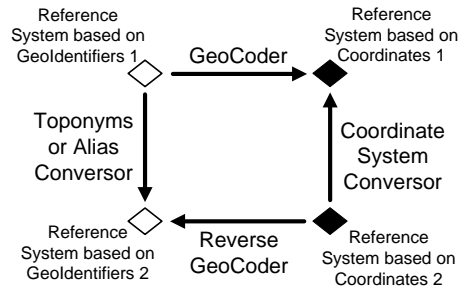


Figure 3.6.: Conceptual idea of the Reference Systems Transformer.

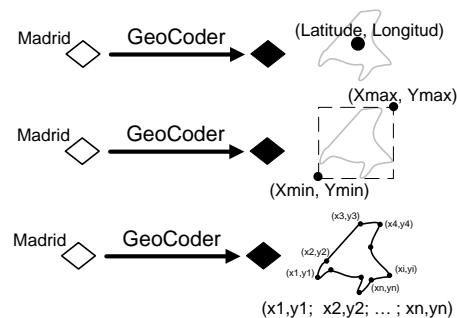


Figure 3.7.: Example of a Reference System Transformation

Then, the main goal at this point is to find the best Indirect Spatial Reference for the geographic region covered by the Direct Spatial Reference. For

this task, we use the *Reverse Geocoder* described in (Renteria-Agualimpia et al., 2013a). This *Reverse Geocoder* uses the *Hausdorff Distance* to measure the geospatial similarity between the geometrical shape of a Direct Spatial Reference and the geographic extent of entities belonging to a *geographical KOS*. The value of the *Hausdorff Distance* is used as a spatial ranking to score the most relevant entities, in a similar way to the work described in (Janée, 2003). This module annotates each processed metadata record with the list of entities that best describe its Direct Spatial Reference.

More formally, the Hausdorff distance metric can actually be adapted to different types of metric spaces, by using different types of internal distance metrics. In the case of geospatial coordinates, there are better alternatives than using the default Euclidean distance as an internal metric. In our case, the geodetic distance was used as an internal metric, since we are using geographic data projected over the Earth's surface. For this reason we use this particular approach over other methods for computing spatial similarity. The mathematical expression for the Hausdorff distance is shown in Eq. (3.1):

$$\text{dis}_H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\} \quad (3.1)$$

where X and Y are two non-empty subsets representing the points that describe a polygon, *sup* represents the supremum, *inf* the infimum, and $d(x, y)$ is the geodetic distance between a pair of latitude/longitude points.

In our approach, we normalize the *Hausdorff Distance* values to the interval $[0, 1]$, where values close to 1 mean strong similarity (high geospatial matching), and values close to 0 mean strong dissimilarity or disagreement between the compared MBBBox (Direct Spatial References). The similarity threshold value is 0.5 in all cases. The normalization function is similar to the function described in (Renteria-Agualimpia and Levashkin, 2011). They use an exponential function to discriminate better the cases really similar.

The reverse geocoding process can use different kind of geographical knowledge organization system as source to the references conversion/transformation process. The selection of the KOS is depending on the next main factors:

- The kind of footprint: The number of spatial dimensions used by the retrieved resources (point, lines, Minimum Bounding Box, multi-polygon).
- The variety of thematic: The thematics covered by the KOS selected must to cover the different thematics of the resources. It allows to establish better matchings between the footprints of the resource analysed and the footprint queried in the KOS.
- The level of detail: The geographical KOS must to have several levels of administrative divisions (geographical extents of different sizes).

The Figure 3.8 summarises the flow of these last three parts of the processes.

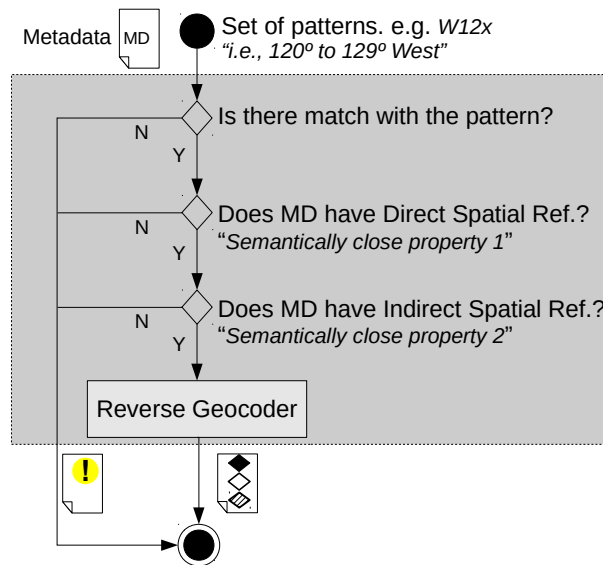


Figure 3.8.: Flow of the Harvester, ETL and Reverse Geocoder processes

3.5 Geospatial Clustering

We define a cluster as a group of resources whose spatial references co-occur in the same area and also share a high value of shape similarity (they have a similar geographical extent). This step uses this definition and works on the hypothesis that a cluster reveals a consensus among experts about the spatial references that are more likely to be used to describe textually a geographic location. This idea will serve to validate spatial descriptions in the resources and detect potential disagreements or inconsistencies. A similar concept of using consensus of spatial co-occurring resources is proposed by Hays and Efros (2008), in their work about geographic location estimation, for each query image they use an aggregate feature distances to find the nearest neighbours in an images database and then they derive geolocation estimates from those GPS tagged nearest neighbours.

To capture the spatial consensus of co-occurring metadata resources, this step uses the density-based DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) (Ester et al., 1996), it is used for computing spatial clusters using as input values of Direct Spatial References found in metadata records. DBSCAN has several advantages: it can recognize clusters with arbitrary shapes; it is not necessary to pre-define the number of clusters in the data; and it is efficient algorithm for big collection of data (Sander et al., 1998). As Wang et al. (2011) summarise it. The key idea is to define a new cluster, or extend an existing cluster, based on a neighbourhood. The neighbourhood around a point of a given radius (*Eps*) must contain at least a minimum number of points (*MinPts*). Given a dataset *D*, a distance function *dist*, and parameters *Eps* and *MinPts*, the following definitions are used to define DBSCAN.

For an arbitrary point, the neighbourhood of *p* is defined as follow:

$$N_{Eps}(p) = \{q \in D \mid dis(p, q) \leq Eps\} \quad (3.2)$$

If $\| N_{Eps}(p) \| \geq MinPts$, then *p* is a core point of a cluster. If *p* is a core point and *q* is *p*'s neighbour, *q* belongs to this cluster and each of *q*'s neighbours is examined to see if it can be added to the cluster. Otherwise, point *q*

is labeled as noise. The expansion process is repeated for every point in the neighbourhood. If a cluster cannot be expanded further, DBSCAN chooses another arbitrary unlabelled point and repeats the process. This procedure is iterated until all points in the dataset have been placed in clusters or labelled as noise. In general, DBSCAN defines a cluster as a set with a maximum number of density-connected data points, in which every core data point must have at least a minimum number of data points within a neighbour of a given radius. The input to the original algorithm can be made of points in a multi-dimensional space. The original DBSCAN algorithm assumes that the data to be clustered are points in given space, whereas in our particular application we are attempting to cluster objects that are represented as bounding rectangles instead of points. Many analysed digital collections contain a geographical extent (MBBox) which is a two-dimensional footprint. We are using an adaptation of the DBSCAN algorithm that uses the *Hausdorff Distance* as distance measurement instead of the Euclidean distance (Joshi et al., 2009; Wang et al., 2010; Renteria-Agualimpia et al., 2013c). The use of the *Hausdorff Distance* (Rockafellar et al., 1998) instead of the Euclidean distance allows computing clusters from two-dimensional data directly (bounding boxes, multi-Polygons or complex geometries).

3.6 Clustering Validation

This step computes first for each cluster two sets of places names (ISR). The first set is the union of the place names generated by the *Reverse Geocoding* module for each resource belonging to the cluster. The second set is the union of the explicit place names in the resource description belonging to a cluster. Next, this step performs in each resource belonging to a cluster a dual validation process. This validation process verifies if exist geospatial inconsistencies between the original Indirect Spatial References and the Indirect Spatial References generated by an external process (reverse Geocoding or clustering); the first validation process validates the Indirect Spatial References with respect to the *geographical Knowledge Organization System*, and the second one with respect to the geospatial consensus provided by every cluster.

Both validation processes are based on the concepts of the Vector Space Model (Salton and McGill, 1983). They measure the similarity between the spatial description of a resource and two vectors of place names associated with the cluster given a resource of a cluster. The first validation measure is the similarity between the vector of generated place names of the cluster and the vector of explicit place names of such resource. The second measure is similar, but it compares the vector of explicit place names of the cluster with the vector of explicit place names of the resource. In both cases, a resource will be considered consistent if the similarity measure is greater than 50%, and will be considered inconsistent otherwise. This step also produces the *best suggested place name*, that is, the generated place name with the best scoring match for the Direct Spatial Reference analysed. Although we use Vector Space Model for calculating the similarity, it is possible to use other metrics for measuring the similarity between them (Mihalcea et al., 2006). In our work, the similarity between two vectors is assessed by the next expression:

$$\cos(\theta_i) = \frac{g_i \cdot t_i}{\|g_i\| \cdot \|t_i\|} \quad (3.3)$$

where t_i is the vector of original place names of a resource belonging to the cluster i , and g_i is the vector of place names generated by the reverse geocoder (for the first kind of validation), or the set of the explicit place names in the cluster (for the second kind of validation). And θ is the angle between these two vectors.

$$g_i = \{g_{1,i}, g_{2,i}, \dots, g_{n,i}\} \quad (3.4)$$

$$t_i = \{t_{1,i}, t_{2,i}, \dots, t_{m,i}\} \quad (3.5)$$

Based on the not repeated place names from these two vectors, a dictionary is constructed as:

$$\{“g_1” : 1, \dots, “g_n” : n, “t_1” : n + 1, \dots, “t_m” : n + m\} \quad (3.6)$$

with $n+m=k$, where k represents the number of distinct place names. We use the indexes of the dictionary to represent each vector by a new k -entry vector, for example:

$$g'_i = \{a, b, 0, d, \dots, k-3, 0, k-1, k\} \quad (3.7)$$

$$t'_i = \{0, b, c, 0, \dots, k-3, k-2, 0, k\} \quad (3.8)$$

Then we measure the level of consistency between the two normalized vectors by calculating the cosine of the angle between vectors using the common Eq. (3.3). A high value of *consistency* for a resource indicates that the resource is consistent. It could be consistent with the geographic references contained within the geographical KOS used by reverse Geocoding (for the individual validation), or it could be consistent with the set of explicit place names in the cluster (for the collective validation). This will facilitate the detection of those resources inconsistent with co-occurring resources in the cluster.

3.7 Report Generation

This step reports the consistency of each resource with respect to its own geospatial information and with respect to its neighbours. Resources identified as consistent could be annotated as having a high quality value, and linked to the place name from the geographical Knowledge Organization Systems used by the reverse geocoder. Resources identified as inconsistent could be annotated with an alert value to advertise the need to review them. This information can be useful for preservation and curation managers (Janée, 2009). All information reported is included in a general report produced by the Statistics Module. This module counts the number of uncompleted descriptions and reports the kind of inconsistency found in individual and collective validation. This report is used for the analysis of the results. An example of the type of generated report is shown in Figure 3.9.




ISR		Place Name:	Ohio
		Title:	Ohio, major land resource areas
DSR		Bbox	(W 85°, N 42°; W 80°, N 38°)
		LCCN:	92681234
		Type:	Map
		Consistency:	0.005
ISR		Suggested Place Names:	North Dakota

Figure 3.9.: Inconsistent metadata record report.

3.8 Summary

This section has presented the details of our methodology to provide Quality Assessment to the semantically close geographical properties. This methodology is based on two-dimensional geospatial clustering, geographic knowledge organization systems, spatial ranking and information retrieval techniques that check the geospatial consistency of a metadata collection. In detail, the proposed two-dimensional reverse geocoder uses an internal topological distance to convert the Direct Spatial References into the equivalent Indirect Spatial References. This conversion is based on geographic knowledge organization systems, specifically, an ontology. The ontology acts as a gazetteer. These two spatial references (indirect and indirect) represent the two most frequently used semantically close geographical properties. Also, spatial ranking techniques have been used for measuring the level of inconsistency between these two semantically close geographical properties. Density-based clustering algorithms use the hypothesis of spatial consensus provided by the spatial co-occurrence of metadata as mean to detect and validate inconsistencies.

We have to make our science socially relevant and user friendly, and not be driven solely by technology. Therefore, geoinformatics can be considered as an agent for making our data and products useful to the public at large.

Sinha (2011)

Chapter 4

Architecture and Implementation

4.1 Introduction

Based on the methodology described above, we have implemented an architecture to provide Quality Assessment for the semantically close geographical properties. The architecture is focused on the assessment of each individual resource. Groups of resources are not evaluated in this charter, that is, the cluster stage of the methodology is considered in chapter 5. The goal of this chapter is to evaluate the suitability of three different spatial ranking approaches to provide geospatial assessment. An overview of the proposed architecture is shown in Figure 4.1.

The architecture has three main steps: *geo-Extraction* step or commonly called Spatial Extraction, Transformation and Load (ETL) (Bédard et al., 2001), *Processing* step (compounded by two parallel modules: traditional based text GIR and Reference Systems Transformer - Spatial Relevance) and finally, the *Evaluation* step.

The first step uses the spatial ETL to extract, transform, and load Direct Spatial References (DSR) and Indirect Spatial References (ISR) from resources. DSR are represented by means of a black rhombus and ISR by means of a white rhombus. When the ISR of a resource are identified, the Traditional textual GIR Module uses well-known and consolidated text GIR techniques to process (extract, clean, transform, etc.) the ISR syntactically (Leidner,

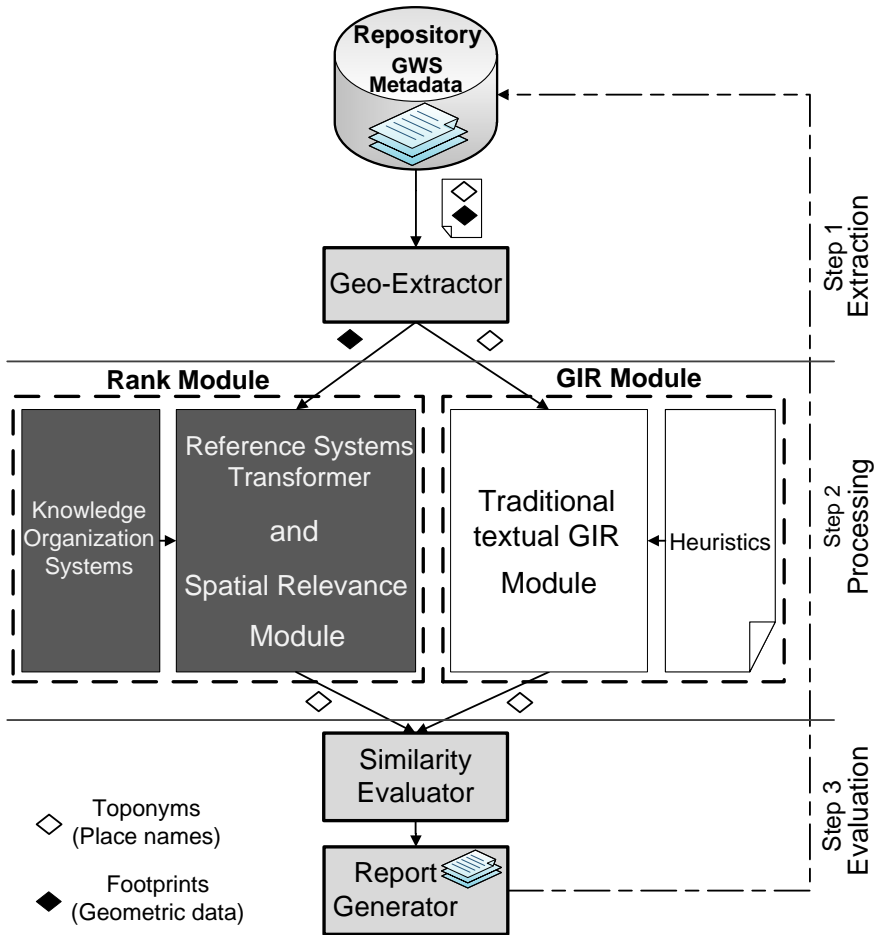


Figure 4.1.: Architecture to provide Quality Assessment for the semantically close geographical properties (source: (Renteria-Agualimpia et al., 2014)).

2004; Martins and Calado, 2010; Lacasta et al., 2014b). The Reference Systems Transformer in the middle of Figure 4.1 is combined with Spatial Relevance functions that make use of spatial similarity measures as a scoring function to retrieve the more relevant toponyms or ISR for a query region.

This module uses coordinate system transformation (Kresse and Danko, 2012) to harmonize the Reference Systems. Additionally, it converts a reference system based on coordinates (i.e., a footprint) into a reference system based on geographic identifiers (geoindentifiers).

The last step of the architecture concludes evaluating the similarity between original place names described in the resource (output of the GIR Module) and the most relevant place names suggested by the Spatial Relevance Module (Rank Module). Extracted ISR from the GIR and Rank modules are compared statistically to identify consistent and inconsistent resources. Resources identified as consistent could be annotated with a high quality value, and a link to the place name from a KOS. Resources identified as inconsistent could be annotated with an alert value to advertise the need to review these resources.

We have tested the architecture implementing three different systems. The three systems share a core module composed by the first and third steps, and also by the GIR module from the second step, but they differ in the use of KOS and how the spatial ranking is measured. The first implementation is based on points (one dimension); the second and third implementations are based on MBBBox (two dimensions). The idea is that the first approach will constitute a baseline to identify the impact of two-dimensional footprints. We formulate two hypotheses here: (1) different KOS and ranking measures detect different kind of inconsistencies, and (2) different geometric types for representing spatial footprints have a different effect on the kind of inconsistencies that can be characterized. Next sub-sections describe the technical details of the core and each of the three components.

4.1.1 Core technical details of shared components

The geospatial ETL takes DSR (the MBBBox) from the structured information described in the resource and the ISR from textual place name fields. The ISR is the input of the GIR module (module in the middle of Figure 4.2, it is shared by the three implementations). This traditional based text GIR module tokenizes the strings of the ISR candidates into individual words. The tokenizing step uses the dot, space and "/" characters as word delimiters.

It also removes the words from a stop word list. The normalizer sub-step takes care of removing case and diacritics from the ISR candidates. More detailed descriptions of these processes can be found in (Wang et al., 2005; Martins et al., 2006). The output of the GIR module is a list of the real ISR characterizing the spatial footprint of the resource.

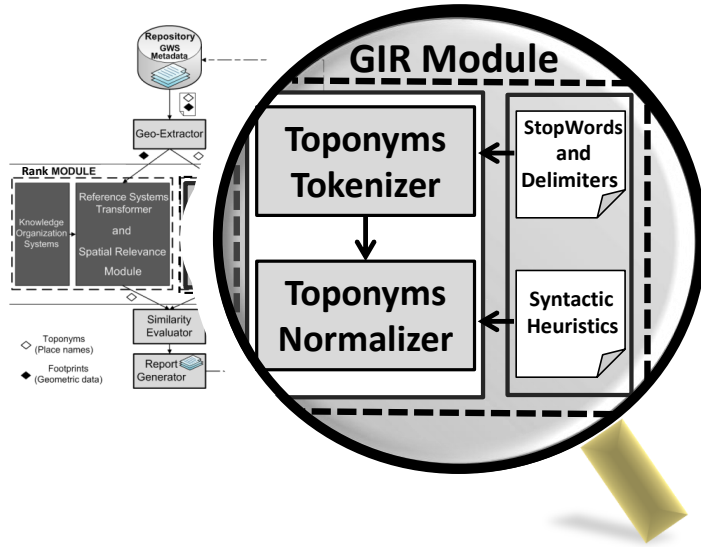


Figure 4.2.: Implementation of the GIR module (the figure has been simplified for reasons of brevity).

The other shared components of the core are the *Similarity Evaluator* and the *Reports Generator*. The first component evaluates the similarity of the ISR produced by the two modules Rank and GIR, and generates consistency reports based on the similarity. Its goal is not only to measure binary consistency (simple presence or absence of place names), but also to make a report including the degree of spatial matching, to suggest the best alternative place names, and to alert about metadata with potential problems of invisibility (relevant metadata that might be hidden to a query). We are using the concepts from the Vector Space Model (Salton and McGill, 1983) to measure the similarity between the output lists (a vector of place names) from the Rank module and GIR module. A metadata will be considered con-

sistent if the similarity (number of matches between place names suggested by an approach and the original place names in the metadata description) is greater than a threshold of 50%. And will be considered inconsistent if the similarity is less than 50%. The best suggested place name would be the place name with the best spatial ranking for the analysed footprints.

4.1.2 Module based on Geonames+DBpedia approach

The first approach implementing the Rank module will be called GeoDB-Wiki. This approach is based on the idea that the social knowledge (the usual way as people know and call a geographic location) determines how users will look for spatial resources. This social knowledge will serve to validate metadata or detect possible inconsistency between the mathematical footprints and the place name of a metadata. This implementation makes use of GeoNames and DBpedia as Knowledge Organization Systems. A GeoNames Web service is used as the Reference System Transformer. This Web service queries the Wikipedia articles in the MBBox described for the DSR of the metadata. The Web service returns Wikipedia articles whose footprint is a latitude-longitude pair (a point), and a name to identify the resource. We take the names of the nearest points to the center of the DSR. Then, a DBpedia Web service verifies if the Wikipedia article identifies a DBpedia GeoTypes, that is, a geographic resource type. The set of Wikipedia names verified as a geographic resource type is the output list of the Rank module for this implementation. The details of this implementation are shown in Figure 4.3.

4.1.3 Module based on Overlapping approach

This implementation makes use of the idea of spatial ranking to transform the DSR into the most relevant ISR. Particularity, it uses the notion of ranking query results based on the spatial similarity of the areas, that is, the areas with best overlapping matching. This module employs the ontology for the representation of Spatio-Temporal Jurisdictional Domains proposed by López-Pellicer et al. (2012a) as the geographic Knowledge Organization

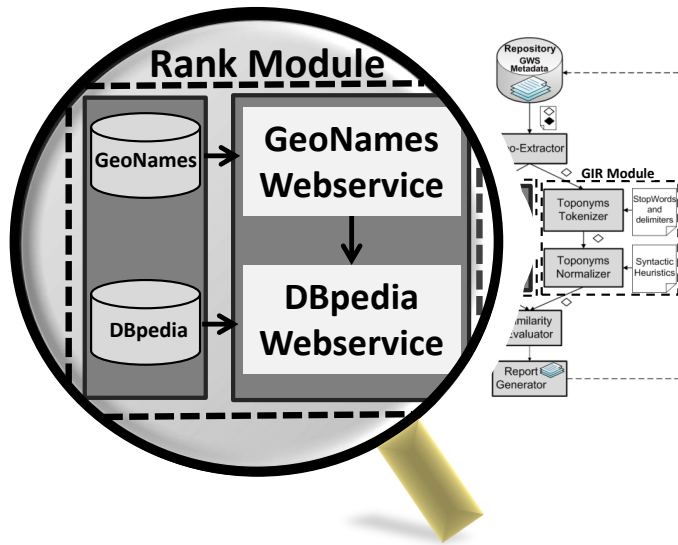


Figure 4.3.: Implementation of the Rank module using Geonames, Wikipedia and DBpedia (the figure has been simplified for reasons of brevity).

System. The ontology represents the spatial entities with a two-dimensional footprint (i.e., a MBBBox). The MBBBox described for the DSR of the metadata is used to query the ontology. This module returns the entities whose footprints are the most relevant entities (i.e., the best overlapping matching according to the function described by Janée (2003)). The ontology entity names (place names) are the output list of the Rank module. Figure 4.4 shows this specific implementation.

4.1.4 Module based on Hausdorff approach

The idea behind this implementation is based on the approaches using metric characteristics such as shape. This implementation uses the notion of ranking query results based on the spatial similarity of the shapes. This

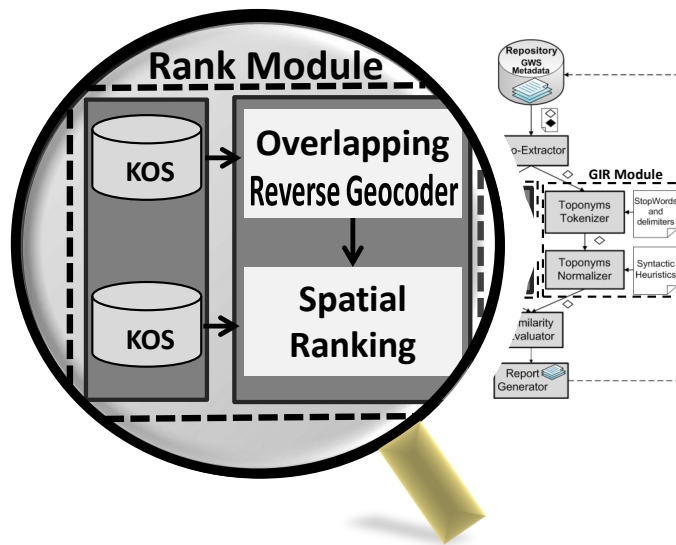


Figure 4.4.: Implementation of the Rank module using Overlapping approach.

module employs also the same ontology used in the previous implementation; it is used as the geographic Knowledge Organization System. The submodule Hausdorff Reverse Geocoder uses the Hausdorff Distance to measure the spatial similarity between the geometrical shape of the query and the shape of the entities in the ontology. The value of the Hausdorff Distance is used by the Rank module to score the most relevant entities from the spatial ontology, in a similar way to the works described in (Janée, 2003; Renteria-Agualimpia et al., 2013c). The ontology entity names (place names) are the output list of this Rank module. This implementation is shown in Figure 4.5.

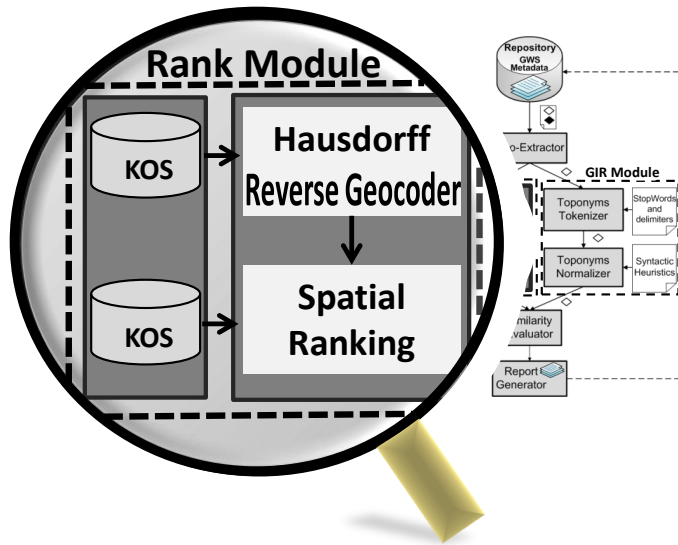


Figure 4.5.: Implementation of the Rank module using ontologies and the Hausdorff distance approach.

4.2 System Evaluation and Results comparisons

We tested the architecture and the three systems with an experiment focused on Spain. The experiment uses as resources a metadata collection of more than 1,000 Web service metadata based on the OGC standard. A focused crawler collected the resources as described the section 3.2. The Web services and the ontology for the representation of Spatio-Temporal Jurisdictional Domains have an extent covering Spain. However, the approach can be applied to other places with a proper geographical Knowledge Organization Systems describing them. The next subsections describe the results and the comparison among these approaches.

4.2.1 Comparison between 1D and 2D approaches

Our main objective was to identify if there is an advantage in shifting from one to two dimensions in spatial data. Secondly, we also wanted to know if there is any advantage due to the increase in the dimension and any advantage derived from the application of a specific approach.

The Figure 4.6. shows the difference in performance of the approaches based on two dimensions above the one-dimensional approach. Two-dimensional approaches *Overlapping* and *Hausdorff* (represented by means of the continuous line) show the ability to detect geospatial inconsistency in the metadata. Dashed line reveals that the inconsistencies are not well detected with traditional points (one dimensional approach). According to the one-dimensional approach, only the 34% of the data are consistent (consistency values greater than 50%, the threshold in the figures). However, when we use the two-dimensional approaches to analyse the other 66% of metadata, we found that only the 24% of these metadata services have a low level of consistency really (consistency values less than threshold). That is, the one dimensional approach was not able to capture all the nature of the problem and detect existing agreement between the analysed ISR.

4.2.2 Comparison between 2D approaches

At this point, we identified a remarkable gain due to the increase in the dimension. But, what is the gain due to a specific two-dimensional approach? To answer this question, we compared the performance of the two-dimensional approaches to evaluate the consistency. The two approaches are capable to detect the inconsistencies between the ISR (place names) in the metadata and the ISR generated by the reverse geocoder (based on a geographical KOS). Additionally, every approach can suggest/generate a *best/ideal place name* associated with the MBBBox of the metadata. To detect the best possible place name, we consider that if the *best place name suggested* by an approach has a consistency value greater than 50% (threshold of 0.5 in the figure) with respect to the original place name described in

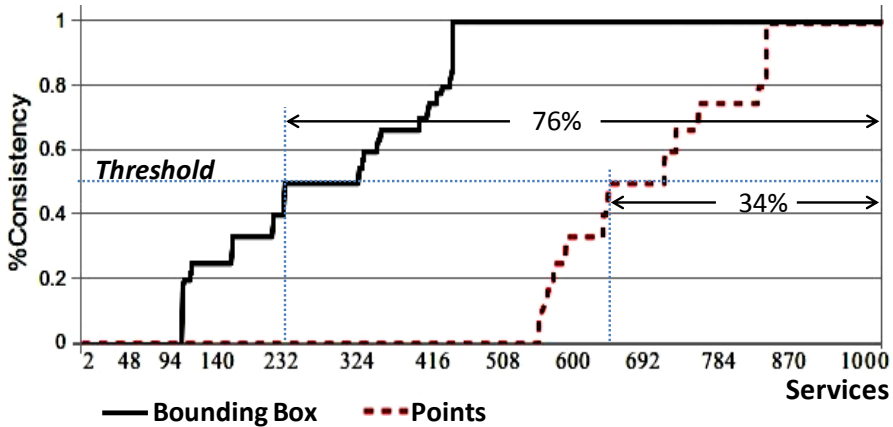


Figure 4.6.: Comparison between two and one-dimensional approaches. The y-axis represents the percentage of all suggested place names matching with the place names in the original metadata description.

the metadata, then the metadata is reported as consistent with major likelihood. We compared the *best place names* generated by the two dimensional approaches, and obtained noteworthy differences. The differences between the *Hausdorff Distance* and the *Overlapping* approach are shown in Figure 4.7. *Hausdorff Distance* approach (continuous line) suggests better *best place names* than the *Overlapping* approach to characterize a metadata, it means, many names suggested by the *Overlapping* approaches are different to the original/true place name according with the geographical KOS used by both approaches. This validation was performed manually for every metadata

We reviewed manually all inconsistent metadata and compared the original place names with the *place names suggested* by every approach. There were identified three main types of inconsistencies:

- *Systematic inconsistency caused by the writing errors.* Misspelling, typos, letter transposition of the place name in the metadata, and numerical errors using scientific notation instead of the format recom-

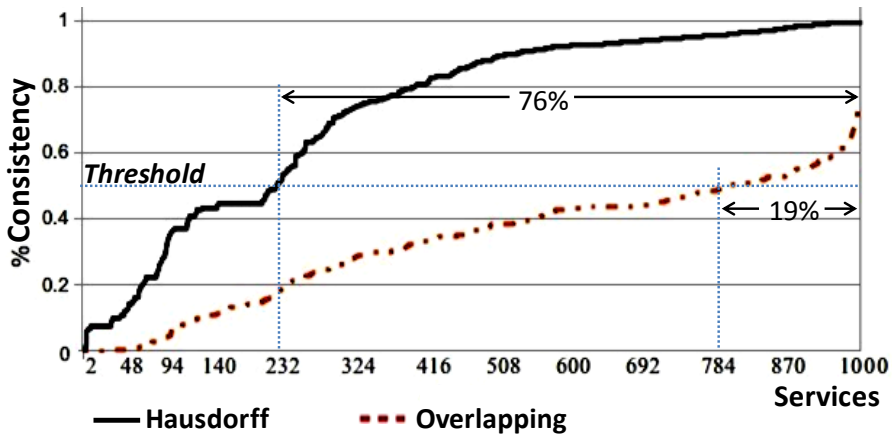


Figure 4.7.: Comparison of the consistency values of the *best place names* suggested by approaches based on two dimensions. In this plot the y-axis quantifies if the *best place name* suggested by an approach matches with the place name in the metadata description.

mended by the standards.

- *Spatial inconsistency caused by fails in the reference system transformation.* The bad transformation generated troubles such as enormous MBBboxes. In other cases, the transformation generated a triangle or a line to represent the service coverage.
- *Overestimation of the service coverage.* Sometimes spatial services have a large background, but their true spatial information is focused on a smaller area. The original place name in the metadata description refers to this small area, but the MBBbox refers to all extent. This MBBbox is used by our approaches to generate/suggest the *best place names* associated with an extent. The overestimation could causes inconsistencies between the original and the suggested place names. This kind of problem is a type of spatial synecdoche, that is, “a term for a part of something refers to the whole of something, or vice versa”.

Generally, resources with an anomalous MBBbox (i.e., MBBbox bigger than the extent of the earth) cannot be detected using the one-dimensional ap-

proach. For example, a service could have an anomalous enormous extent but its point location is consistent. *Overlapping and Hausdorff Distance* approaches detected 65 services with this third type of inconsistency. The Table 4.1 presents a summary of the capabilities of the three sub-approaches implemented.

Table 4.1.: Characterization of the capabilities of the three sub-approaches implemented

Characterize	GeoDBWiki	Overlapping	Hausdorff
Inconsistencies Type 1	Yes	Yes	Yes
Inconsistencies Type 2		Yes	Yes
Inconsistencies Type 3			Yes
% of matches ^a	326/968=34%	736/968=76%	736/968=76%
Well suggested name ^b	162/968=17%	184/968=19%	736/968=76%

^a Quantify the number of suggested place names matching with the place names in the original metadata description.

^b Measure if the best place name suggested by an approach matches with the place name in the metadata description.

According to Table 4.1, the GeoDBWiki approach only could verify the consistency of 34% of the metadata. It means that GeoDBWiki was not able to detect all matches between the place names in the metadata and the place names generated by the approach. The main problem with this approach lies in the little information represented by footprints of one dimension. On the other hand, Overlapping and Hausdorff Distance approaches generated the same number of place names because they use the same kind of footprints (the original MBBBox in the metadata). These approaches verified the consistency of 76% of the metadata and they were able to detect more matches than GeoDBWiki. The advantages of these approaches lie in the use of information spatially richer (footprints of two dimensions). However, there is a relevant difference between Overlapping and Hausdorff Distance approaches. The best place names suggested by Hausdorff Distance approaches were more accurate (76%) than those ones suggested by the Overlapping approach (19%). In this context, a high accuracy means that the best suggested place name was the same place name described in

the metadata. The disadvantage of the Overlapping approach is in its main philosophy as it is based on surface areas. Approaches like Hausdorff Distance are more accurate because they are focused on shape similarity.

4.2.3 Analysis of results

Many problems arise from the type of geographical representation. Our results show the advantages of shifting from one-dimensional to two-dimensional representations of the spatial footprint. However, our approach has problems in some special cases when the two-dimensional data (in particular MBBboxes) are not enough representative. For example, Larson (2011) exposes the imprecision of MBBbox and their tendency to overstate the size and shape of areas (for example the often noted case of Portugal being completely subsumed by the MBBbox of Spain). It can lead to errors in retrieval, and spatial ranking. This has an effect on the analysis of inconsistency and in general, in all tasks of the geographic information retrieval system. We have verified many services reported as inconsistent using the overlapping approach, but when we analysed the services, they were not really inconsistent. It was the case of the Web services of the Canary Islands. The problem is due to the basic idea behind the overlapping approach; it is based on the percentage of areas intercepted. Due to the dispersion and the small size of isles in the Canary Islands region with respect to its MBBbox, the percentage of the area of land is much less than the percentage of sea. An approach based on the shape, such as the Hausdorff Distance, solved these problems. Approach like this can be applied to discover inconsistencies in disaster databases, geospatial health datasets, georeferenced Digital Libraries, climate databases and other domain containing georeferenced two-dimensional data. For example, it is very important to ensure the consistency of air pollution datasets to improve the urban decision support systems.

4.3 Summary

This section presented a semi-automatic method to check the geospatial consistency of metadata records based on spatial ranking measures. Our experimental results with a dataset of more than 1,000 records about Spanish Geospatial Web Services (WMS specifically) show that the use of this approach provides not only significant advantage in terms of inaccuracy detection, but also a gain of use of social knowledge insight into the metadata. These results have shown that spatial ranking approaches for GIR can be applied to analyze the spatial consistency of geospatial metadata. The detection of the spatial inconsistency of a metadata can warn about potential problems of irretrievability and invisibility from the point of view of information retrieval. Additionally, we noticed that the types of geometric representation have a remarkable influence on the detection of the spatial inconsistencies. The common reduction of the dimensions in the representation has a considerable negative effect on the tasks of spatial ranking and, thus in spatial ranking applied on inconsistency detection. This work points out the following advantages of using two-dimensional approaches to detect problems causing the invisibility of spatial resources. It reveals latent problems of many works, systems and applications (e.g., Digital Libraries, Spatial Data Infrastructures, and so on) representing a spatial footprint as a point simply. We consider that this research may provide a way for retrieving deep hidden metadata records in catalogues due to geospatial inconsistencies.

Everything is related to
everything else, but near things
are more related than distant
things

Tobler (1970)

Chapter 5

Quality Assessment for Digital Libraries

5.1 Introduction

This chapter uses the third system described in the section 4.1.4, which proves to be the best spatial ranking method for applying Quality Assessment to a two-dimensional resource dataset (geospatial web services metadata). It was capable to detect the most spatial inconsistencies. Here we use Hausdorff Distance as scoring function to do the *spatial ranking*, *Knowledge Organization Systems*, and also we use *geospatial clustering* to compare co-occurring semantically close geographical properties: Indirect Spatial References (ISR) and Direct Spatial References (DSR). The hypothesis is that geospatial clusters could reveal an implicit consensus among experts that work in the Digital Library field for identifying some geographic areas. Such consensus is dependent on traditions, values, interests and particular goals to the community involved in each Digital Library, and hence it could even be specific for each cluster. Therefore, homogeneous and distinct clusters that group spatially metadata records could provide clues for validating and detecting inconsistencies among its members.

The studied dataset consists of more than 42,000 records that describe maps belonging to the American Library of Congress (LoC). Most of the maps

are referencing local, sub-national, national and regional areas, for example: cities, watersheds, forest areas, counties, states and countries.

This chapter is organized as follows. Section 5.2 introduces the detail of the experiment. Section 5.3 describes the experimental and quantitative study with its analysis results. Section 5.4 discusses the main results and Section 5.5 concludes the chapter.

5.2 Experimental details

This section presents an experiment performed in order to show the applicability of our Quality Assessment methodology to other domains and disciplines. It is shown by evaluating the spatial consistency of the semantically close geographical properties in Digital Library metadata.

Details of the six main steps (*Harvesting, Geo-Extraction, Reverse Geocoding, Geospatial Clustering, Metadata Validation, and Report Generation*) of the methodology are presented as follow:

As we mentioned in the section 3.2.2, the *collecting process* in Digital Library domains was done by means of harvesting techniques. although SRU protocol has not explicit geographical information retrieval support, we formulated a heuristic based on string patterns, it allows to create queries that can retrieve metadata with information about the geographic extent of the resource (the Direct Spatial Reference specifically). This heuristic serves to search metadata records that contain sub-strings that may encode geographic coordinates such as section 3.2.2 describes. The result was a dataset of more than 42,000 Digital Library metadata.

Regarding to the *Reverse Geocoding module*, we made an analysis of the more suitable sources of the available KOSs. Based on the main factors recommended for the selection of the KOS, we have made a KOS using several sources. The resulting KOS has geographical entities with two-dimensional footprints (Minimum Bounding Box and multi-polygon). This KOS covers the next main topics (populated places, geopolitical division, forest reserves, watershed, river basins) found in the analysed resource collection. Also,

the KOS contains geographical extents of different sizes, including: country, states, counties, districts, cities, towns and other areas. In particular, the *geographical KOS* used consists of several public models, databases and KOS. Its main sources are available online¹.

Regarding to the *Geospatial Clustering module*, it is important to have in mind the central hypothesis. This step works on the hypothesis that a cluster can reveal shared spatial information of a set of resources co-occurring in a place. In this domain, a cluster reveals a consensus among library experts about the spatial references that are more likely to be used to describe textually a geographic location. This idea will serve to validate spatial descriptions in the metadata record and detect potential inconsistencies. This step uses the density-based DBSCAN clustering algorithm using as input values the Direct Spatial References found in metadata records as section 3.5 describes.

An important issue here is the parameter setting. The DBSCAN algorithm uses three main parameters: Minimum number of elements inside the clusters *MinPts*, epsilon (*Eps*), and the *distance function*. We selected the *Hausdorff distance* as the distance function because this distance was the best suitable distance for working with two-dimensional resources (Bounding Boxes) as the chapter 4 shows. As a basic consideration, a cluster is group of at least two elements, for this reason, the *MinPts* parameter is setted with 2. The more complex selection is the *Eps* parameter. DBSCAN algorithm is very sensitive to its parameters, especially to *Eps*, the radius of the search. A small *Eps* value means that the radius of search of the algorithm is shorter, and indeed restrictive, so the results will a big number of clusters, more compact and dense, and more noise. On the other hand, using a higher *Eps* value, and the same value for *MinPts* we obtain a small number of clusters, that aggregate more number of elements each. In our work we use a restrictive (small) value of *Eps* (0.2) and *MinPts* of 2. These values provide the best separability for co-occurring spatial objects. That is to say, objects that co-occur from the one dimensional perspective (coordinates based on points),

¹<http://www.gadm.org>

<http://www.census.gov>

<http://www.naturalearthdata.com/downloads>

<http://www.nws.noaa.gov/geodata/catalog/national>

but they are georeferencing spatial entities of different levels/size (example, a city, a province and a state centered in the same point but with different extents coverage). The recommended technique for the parameter selection is described by Ester et al. (1996), the same work where DBSCAN is introduced. It consists of generating a histogram with the sorted k -neighbour distance (*Hausdorff distance* in our case), being k the desired value of *MinPts*. Then this distance is sorted (descending) and plotted. The histogram will show a descending curve. In (Ester et al., 1996), the authors suggest that the optimum value of *Eps* parameter is the distance where the curve makes its first inflexion (or “valley”). The elements located on the left of this “valley” will be noise in the resulting partition and the rest will be present on some of the resulting clusters. In (Ester et al., 1996), the authors ensure that choosing 4 as the default value of *MinPts* produces the best results in two-dimensional clusterings. In our experiments lower values, usually 2, obtained better results for the spatial ranking.

5.3 Analysis and results

We tested our methodology analysing the quality of 12,000 metadata records that describe resources in the United States of America. This is a subset of a larger collection of more than 42,000 metadata records retrieved from the LoC in May 2013. The collection was harvested by the process described in the harvesting section (3.2.2). All examples, experiments and results here are based on records available on that date. Some records may have changed since that date. Although the analysis has been restricted to the United States of America, the methodology can be applied to other places. For the experiments, we have analysed and selected just the most frequent groups of elements in the dataset; the results are metadata records on which the Direct Spatial Reference is georeferencing a State, a County, a City, a Forest and a Watershed. The distribution is shown in Figure 5.1.

The methodology has helped to detect three kinds of inconsistencies: (1) *Spatial logical inconsistency*, (2) *geospatial semantic inconsistency* or *geosemantic inconsistency*, and (3) *contextual geospatial inconsistency*.

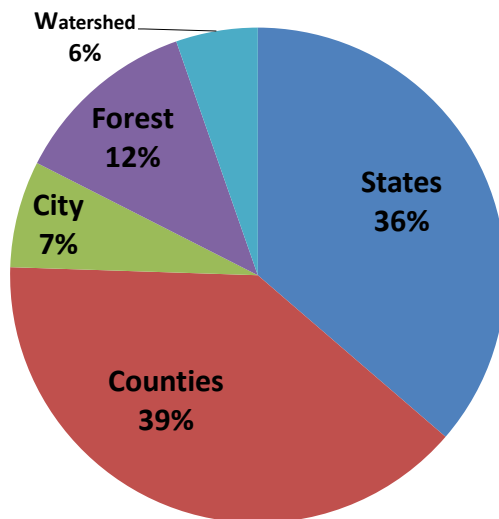


Figure 5.1.: Distribution of the type of extent coverage of the metadata used in the experiment

(1) *Spatial logical inconsistency*. This kind of inconsistency is caused by logical problems in the codification. In addition to the traditional logical consistency, a library with geospatial resources needs to verify a more complex consistency of their metadata, for example according to the international standard ISO 19113 Geographic Information - Quality Principles (Gong and Mu, 2000; Wang, 2008; Xie et al., 2010). For example, the range of latitude and longitude coordinates need to be checked: the value of latitude must be between 90° North and 90° South, and the value of longitude must be between 180° East and 180° West. In some cases, a simple query such as *Are the values of latitude coordinate always between -90° and 90° ?* can reveal a geospatial logical inconsistency. Our methodology also reveals distorted (extra long) Direct Spatial References shown in the Figures 5.2 and 5.3. In many cases the results shows that they were encoded in the description data out of range.

(2) *Geosemantic inconsistency*. This kind of geospatial semantic inconsistency is originated in the conceptual incoherency between the Direct Spatial Reference and the Indirect Spatial Reference according a specific KOS.

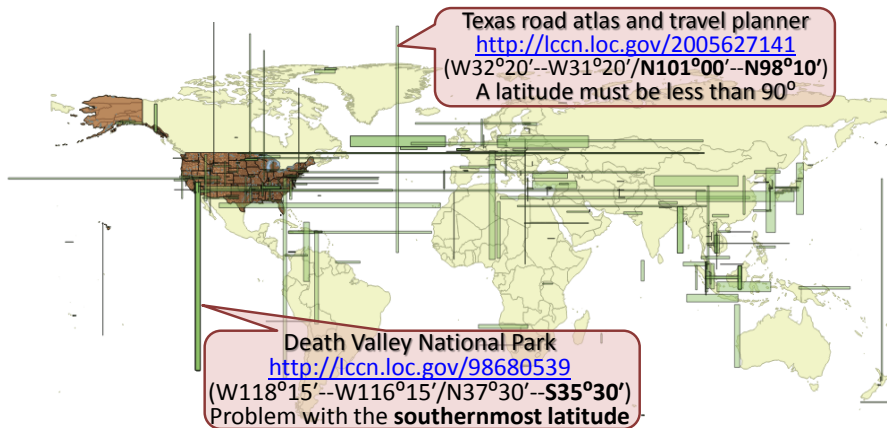


Figure 5.2.: Global vision of the spatial logical inconsistencies in the LoC metadata records.

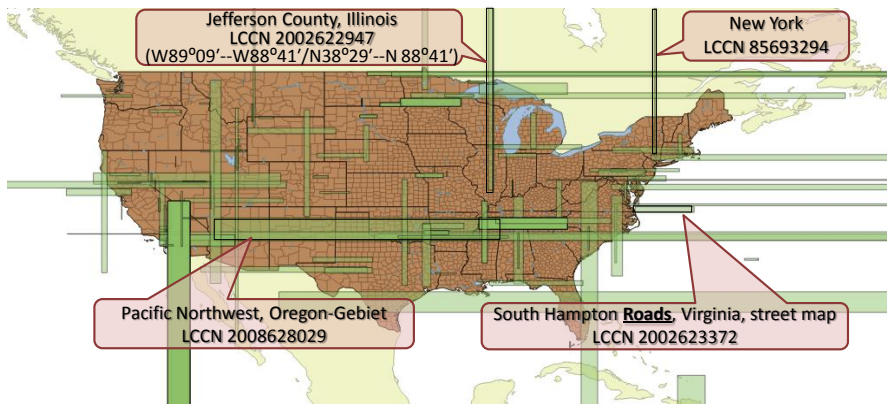


Figure 5.3.: Spatial logical inconsistencies in the LoC metadata records focused on USA.

There are three cases: micro-macro, macro-micro, and unmatched. The micro-macro case happens when the Indirect Spatial Reference (place Name) of the metadata record has a micro scope (county, town, park, forest, etc) but its Direct Spatial Reference (extent coverage) has a macro scope (state,

country, region, etc). The macro-micro case is its inverse, where the Indirect Spatial Reference has a macro scope but the Direct Spatial Reference covers a small area. This kind of problem is a type of spatial synecdoche. Finally, the unmatching case revealed by the reverse geocoder with the help of the *Hausdorff Distance*. Figures 5.4 and 5.5 show examples of unmatching cases.

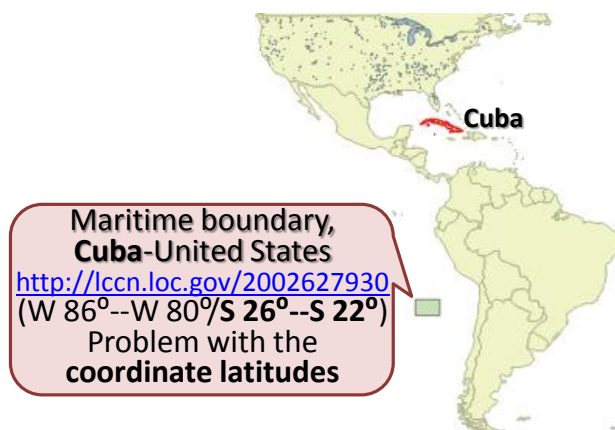


Figure 5.4.: Example of place name and footprint with geospatial unmatching of Cuba

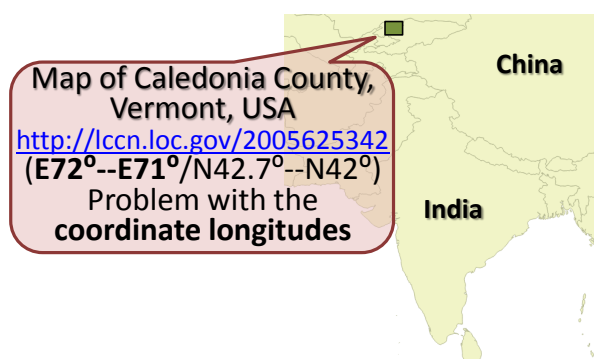


Figure 5.5.: Example of place name and footprint with geospatial unmatching of Caledonia County

(3) *Contextual geospatial inconsistency*. This kind of inconsistency is caused by a disagreement between the geospatial co-occurring Direct Spatial References, for example, a disagreement between a metadata record and the consensus of its neighbours as Figures 5.6, 5.7 and 5.8. In the first case, the methodology identifies a disagreement between a metadata (Charles County-South Dakota <http://lcn.loc.gov/99463858>) and the consensus of its neighbours (LaMoure County North Dakota). In the second example, the methodology identifies the disagreement between the metadata describing (Ohio State <http://lcn.loc.gov/92681234>) and the consensus of its neighbours (North Dakota State). Our clustering-based approach also points out groups of metadata records with potential geospatial inconsistencies. These could be caused, among other things, by systematic errors, the reuse of non-validated metadata or the lack of information about the area in the geographical KOS used to validate. When a KOS does not have information about an area, we need an alternative way to validate the consistency. For example, there are cases where the best source of information for validating is provided by the descriptions found in the cluster itself. That is, the cluster can be seen as representative of the collective knowledge of an area, some of these cases can occur with native and unofficial places names or offshore fishing ground names, etc. Two examples are shown in Figure 5.9. The contextual geospatial inconsistency differs to the geosemantic inconsistency in the sense of the individual or group evaluation and in the presence or absence of external information to validate the consistency of an evaluated metadata. Geosemantic inconsistency is applied on individual metadata and makes use of KOS, while contextual geospatial inconsistency is applied on clusters and it could use KOS optionally.

In some cases, we have found that most of the metadata records in a cluster are inconsistent. In such cases, we have applied a dual validation procedure, collective and individual one. We use the reverse geocoder to validate every metadata record and the contextual consistency of all metadata belonging to the cluster. An example is shown in Figure 5.10. In this case, 8 out of 14 elements in the cluster are inconsistent, thus the cluster is inconsistent. Table 5.1 shows these inconsistent elements. A consistent cluster could be employed in assessment tasks. For instance, it could be used to validate the geospatial consistency of new records.

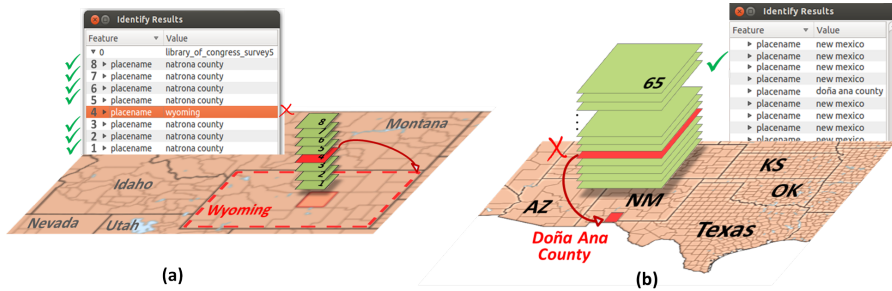


Figure 5.6.: Spatial synecdoche: Disagreement between a metadata (Wyoming <http://lccn.loc.gov/2011593232>) and the consensus of their neighbourhood (Natrona County). (b) Disagreement between (Doña Ana <http://lccn.loc.gov/93682208>) and the consensus of their neighbourhood (New Mexico).

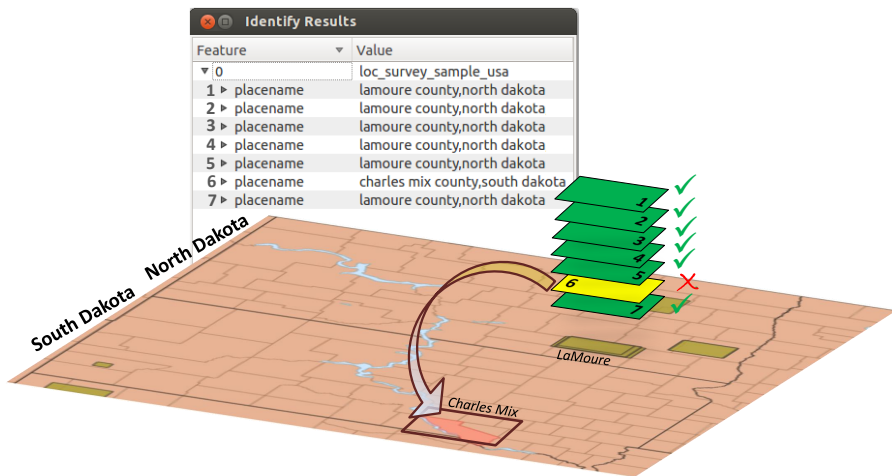


Figure 5.7.: Disagreement between a metadata (Charles County-S. Dakota <http://lccn.loc.gov/99463858>) and the consensus of its neighbours (LaMoure County N. Dakota).

The results are summarized in Table 5.2. We have found geospatial inconsistencies in 870 out of 10575 metadata records. Our methodology identified

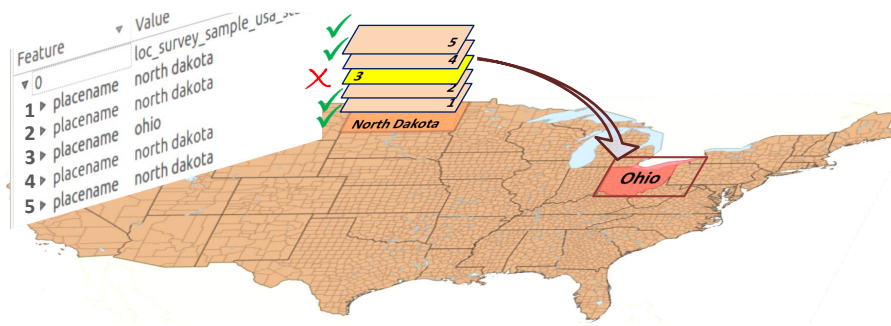


Figure 5.8.: Results showing the disagreement between a metadata record(Ohio State <http://lcn.loc.gov/92681234>) and the consensus of its neighbours (North Dakota).

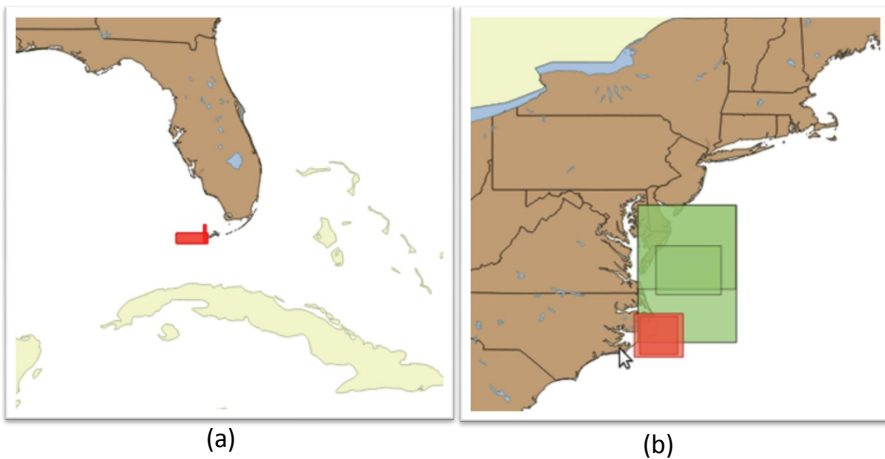


Figure 5.9.: Example of lack of information about the area in the geographical KOS, (a) Lower Keys fishing map in Florida and (b) Hatteras offshore fishing chart in North Carolina.

212 (2%) metadata records with logical inconsistencies and 802 (7.6%) metadata records with geosemantic inconsistencies. Also, 93 (0.9%) metadata records presented a disagreement with their neighbours. The administrative types (states and counties) present fewer inconsistencies than types with

Table 5.1.: Contextual inconsistency caused by systematic error probably

Current location of the DSR	Real location According to the ISR	LCCN
Lake County, Cook County, Minnesota Stae.	Ward County, North Dakota State.	http://lccn.loc.gov/00553926
		http://lccn.loc.gov/00553927
		http://lccn.loc.gov/00553928
		http://lccn.loc.gov/00553929
		http://lccn.loc.gov/00553930
		http://lccn.loc.gov/00553934
		http://lccn.loc.gov/00553935
		http://lccn.loc.gov/00553936

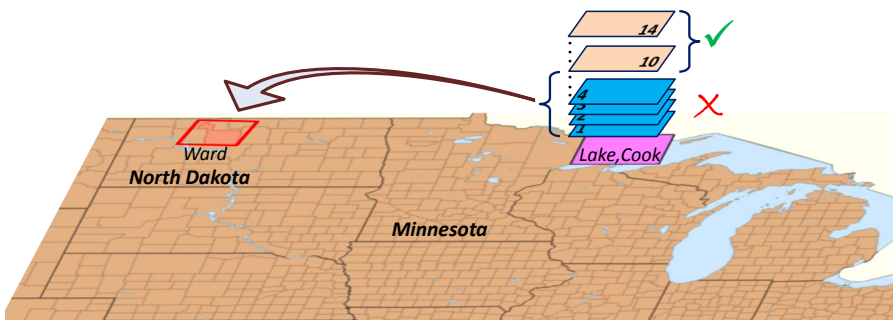


Figure 5.10.: Results showing an inconsistency due to the disagreement among geospatial co-occurring Direct Spatial References belonging to a cluster.

imprecise boundaries (cities and forests). However, it is surprising that a man-made feature (cities) has proportionally more inconsistency issues than other types analysed (24.6%). That is to say, more geospatial disagreements among these records. Proportionally, the records georeferencing states are the most consistent (96.8%) and also they present better geospatial consensus than other categories.

Table 5.2.: Typology of the geospatial inconsistencies found in the LoC metadata dataset

Type	No	Geospatial inconsistency			Total
		Logical	Semantic	Context.	
State	3840	84 (2.2%)	112 (2.9%)	15 (0.4%)	123 (3.2%)
County	4146	81 (1.9%)	305 (7.4%)	21 (0.5%)	324 (7.8%)
City	737	20 (2.7%)	162 (21.9%)	24 (3.3%)	181 (24.6%)
Forest	1287	24 (1.9%)	158 (12.3%)	14 (1.1%)	163 (12.7%)
Watershed	565	3 (0.5%)	75 (13.3%)	19 (3.4%)	79 (13.9%)
Total	10575	212 (2.0%)	802 (7.6%)	93 (0.9%)	870 (8.3%)

5.4 Discussion

There are five issues that deserve to be discussed with respect to the methodology and its results in the Digital Library domain: *A Inverse Methodology; Geographic Knowledge Organization System selection; Outlier detection and inconsistencies; The dimension in the spatial representation and Metadata reuse.*

Regarding the Inverse Methodology point, we propose a methodology based on comparing sets of place names. Alternatively, a methodology based on comparing geospatial coordinates could be developed. However, the main difficulty of this last approach is the high level of uncertainty generated by the ambiguity in the toponym transformation process (geocoding). Without additional information is complicated to convert very ambiguous terms/-toponyms in their equivalent coordinates. Furthermore, two-dimensional footprint obtained by geocoding the place names that are mentioned in the metadata descriptions is more complex. Regarding the second point, a crucial part of the *Reverse Geocoding is the geographic Knowledge Organization System selection.* We need to take into account the next requirements: it needs to be a two dimensional dataset with different levels of details (geographical extents of different sizes). And also, the *geographic Knowledge Organization System* needs to have several topics. Regarding the third point,

outlier and inconsistency detection is not an easy task. We took advantages of DBSCAN to detect outliers in our geospatial domain. Outliers are candidates to be inconsistent according to the clustering algorithm. In this case, however, we need an additional way to verify the record. In cases when a metadata record spatially consistent is alone in an area (it does not belong to any cluster), the clustering approach needs to be complemented with an individual validation, for example, by using the two-dimensional reverse geocoder. Thus, metadata validation by means of clustering can be applied when we have additional information about neighbours with a good spatial consensus. Regarding the *dimension in the spatial representation*. We have identified many cases where the one-dimensional representation generates problems. All these problems are due to MBB with incomparable areas, for example, when a metadata is georeferencing macro areas (countries, states) and another metadata is georeferencing micro-local areas (cities, towns, parks) and both are represented and centered in the same point. Thus, for these cases, a good solution could be the use of representations, algorithms and methodologies focused on two-dimensional data. This is the main idea behind our approach, this kind of techniques provides separability for co-occurring spatial objects in one dimension, but georeferencing spatial entities of different levels (example, a city, a province and a state centered in the same point but with different extents coverage) as the Figure 5.11 shows. Figure 5.12 illustrates this situation in the context of clustering, (a) a clustering process with one-dimensional representation generates 3 clusters only, while (b) a two-dimensional process generates six clusters and gets a better separability between co-occurring MBB with differentiated extent coverage.

Regarding the final point, *Metadata reuse* is an essential task in Digital Library domain. To understand the importance of reviewing the consistency of metadata first we need to understand the proper importance of the metadata such as the FGDC argues: "If you think the cost of metadata production is too high - you have not compiled the costs of not creating metadata: loss of information with staff changes, data redundancy, data conflicts, liability, misapplications, and decisions based upon poorly documented data" (FGDC, 1998a). Even if we accept the importance of metadata, we need to worry about its quality. For example, metadata sharing and reuse is a common practice in Digital Libraries. These practices should include a richer

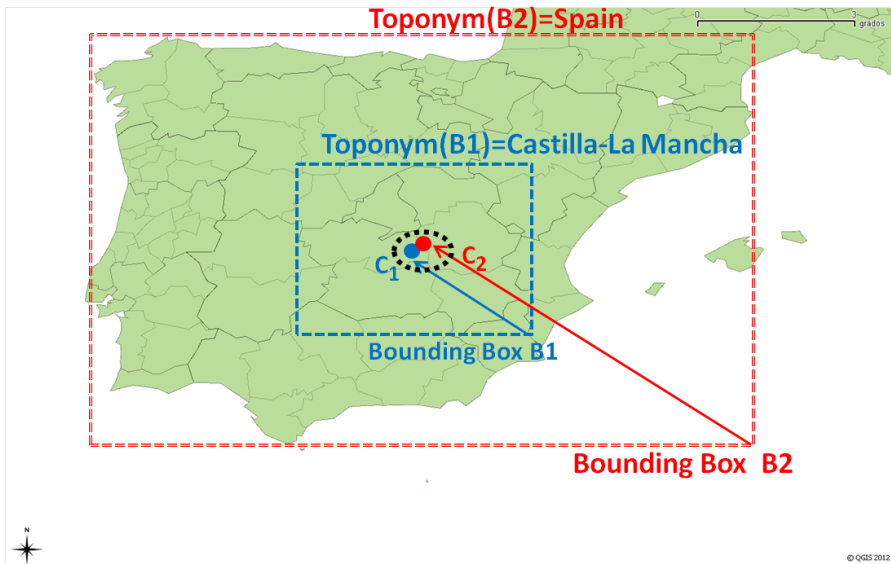


Figure 5.11.: Problems of separability when we reduce the representation of two-dimensional spatial object B_1 and B_2 to their central points C_1 and C_2 .

geospatial consistency validation in order to ensure the data be retrieved and the quality of the entire library processes. We believe that before applying interoperability and sharing in Digital Libraries, we need to revise the geospatial consistency between the semantically close geographical properties, that is, the fields of spatial references (Direct Spatial References and Indirect Spatial References) used in tasks such as retrieval, exploration and visualization of spatial information. The omission of these aspects can lead to problems of information retrieval and invisibility of geospatial resources, such as maps and other materials spatially referenced by the metadata in a Digital Library.

In some domains metadata have a poor and negative reputation. It can be changed, as Giles (2011) cites, one way to improve this reputation is to recognize and criticize existing published records that do not meet the needs

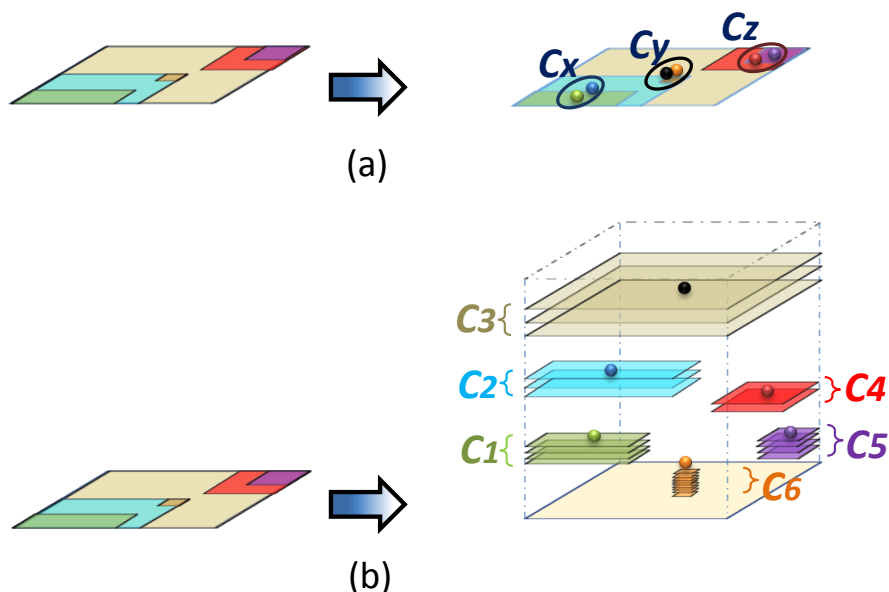


Figure 5.12.: Conceptual differences between 1D and 2D clustering. (a) Only three of the six clusters are identified with 1D clustering, but (b) the six well-differentiated clusters are identified with 2D clustering.

of the users. Take time to assess the quality, e-mail the record, warn and explain metadata shortfalls or why the records are inconsistent/inaccurate or obsolete. Use every opportunity to peer-review existing metadata records. And finally, develop more ways of feedback.

5.5 Summary

This chapter has presented the application of our methodology to the Digital Libraries domain. Our experimental results with a collection of more than 12,000 records about United States maps from the Library of Congress show that the use of this approach provides not only significant advan-

tage in terms of inaccuracy detection, but also a gain of the use of spatial co-occurrence and the geospatial consensus (spatial neighbourhood knowledge) insight into the metadata. Experimental results show that this methodology can be applied to detect the spatial inconsistency of metadata records and assess potential problems of information retrieval and invisibility of georeferenced resources. We have shown several concrete examples of serious incidents caused by inconsistencies between the semantically close geographical properties. Properties commonly used to discover, explore, access, retrieve and geo-visualised resources.

The utility of this methodology in a Digital Library is recognized, since fixing an error in a resource is appropriate when the Digital Library is responsible for the content of the resource. That is, the Digital Library authored the resource or bears intellectual responsibility for it. For metadata, it is always appropriate for a Digital Library to improve metadata errors and correct the inconsistencies. This is true whether the metadata was created by the Digital Library or whether the metadata was harvested from an external source.

"And you shall know the truth,
and the truth shall make you free."

Jesus Christ [S. John 8:32]

Chapter 6

Conclusions

This chapter summarises the work presented in this thesis, evaluates its contributions, mentions some limitations, and conclude with ideas for further research.

6.1 Summary of Contributions

This thesis has researched how to assess the quality of metadata that describe the spatial location of a resource, and the problems that may surface when a metadata record describing a resource has semantically close geographical properties, that is, pair of properties that describe its location using different reference systems (e.g. text and coordinates). This problem is closely associated to the facility which georeferenced resources can be retrieved in an information system. This approach has been used to show the need for methods and tools that analyse the geospatial semantic consistency of these properties in order to improve the discovery, accessing and retrieval processes of geographical information from different perspectives. Starting from this aim, the main contributions of this thesis are the following:

- **An approach that takes advantage of the spatial co-occurrence of the large volume of geospatial information:** We have presented a methodology that takes advantage of spatial co-occurrent metadata and their cumulative knowledge describing a same place to

validate a particular resource description or to find discrepancies with respect to its neighbourhoods. The increasing volume of geospatial data everyday makes infeasible to search through their content directly. Many information systems use instead geospatial metadata. However, we have shown that large volume of spatial information can be exploited to provide Quality Assessment for co-occurring metadata. With this methodology, an adapted two-dimensional clustering algorithm has been proposed to capture the geospatial co-occurrence, and to discriminate when a co-occurrent metadata just overlaps, and instead it belongs to an inferior or superior cluster. One of the particularities of our methodology is its flexibility. This methodology has shown the capability to integrate different clustering algorithms, reverse geocoders, and two-dimensional ranking methods.

- **A comparative study of spatial ranking approaches for one-dimensional and two-dimensional data:** This thesis has introduced a comparison between different ranking approaches and their ability to work with one-dimensional and two-dimensional data. The results have showed a significant advantage in geospatial inconsistency detection of the approaches based on two-dimensions. The nature of geospatial inconsistencies was detected mostly when we shift from one dimension to two dimensions. The results have revealed that macro and micro geographical extents traditionally are mixed in a point, but approaches based on two dimensions help to discover inconsistencies hidden for one-dimensional approaches. Also, in the comparison we contrasted approaches using social knowledge sources, Wikipedia and DBpedia, with approaches using official sources. In general, the accuracy was better when we used official sources to validate metadata descriptions, however, when we worked in the smallest geographical extents, the social sources provided spatial descriptions not found in official sources.
- **Two real tests in two real scenarios with two two-dimensional datasets:** This thesis has introduced an empirical and quantitative study of the spatial quality of the semantically close geographical properties in two scenarios: SDI and Digital Libraries. With these scenarios we have performed a dual validation of our methodology,

the first validation used a dataset of more than 1000 Web services from the Spanish SDI, meanwhile the second validation consisted on a dataset of more than 42,000 MARC21 metadata records from the U.S. Library of Congress. The empirical study has provided an overview of the characteristics of the common spatial inconsistencies in published metadata resources, and also reveals common and systematic errors in the current practices in these communities in the provision of metadata for cartographic resources. The study has characterised and summarised these common spatial inconsistencies. The characterization of inconsistencies in the Digital Library scenario is made taking into account the experiences gathered with SDI catalogues. Although, inconsistency problems exist in the SDI scenario, they were less frequent than the Digital Libraries scenario. We have found that SDI quality problems are minimised because SDI personnel are experts and technicians with specialised, advanced and detailed spatial knowledge of the geographical domain, and also, it is due to the specialised geographical focus of the SDI catalogues and developed standards. For these reasons, the inconsistency problems of resource descriptions in SDI are probably caused by technical issues.

- **A semi-automatic Quality Assessment tool for Geospatial Metadata:** Correcting geospatial inconsistencies of SDI and Digital Library resources is not trivial for non-expert personnel and users in geospatial disciplines. In this line, our methodology can assist personnel with a semi-automatic Quality Assessment tool that improve the retrieval and systems interoperability by means of reducing the invisibility of the geospatial resources, specifically, the invisibility caused by geospatial inconsistencies of the semantically close geographical properties used to retrieve those resources.

Also, we have pointed out some of the implications of the geospatial inconsistency problems. A resource with a poor quality description is for most purposes invisible. Invisible resources deteriorate the effectiveness of the information system devote to manage the information. Ensure the quality of the description is vital to ensure the future access and discovery of resources held by SDI, libraries and archives. Our work has provided a mechanism to alert and generate reports

of inconsistencies and then, help in digital curation processes. This mechanism of inconsistency detection can also be used to alert about potential problems of disconnection in interoperable and distributed information systems.

6.2 Future Work

The goal of this thesis is the improvement of the accessibility, retrieval, and visualization for geospatial information resources in the context of digital repositories in general. Many open questions remain that require further research. These are the opportunities identified for following them up:

1. **Apply lessons learned to the analysis of the geospatial consistency status of other domains that use other kind of metadata.** The implementation of the approach proposed in this thesis only considers two kind of metadata document with geospatial information, the OGC Web services metadata and the MARC21 metadata. With respect to the first one, other resource metadata schemas in the geospatial domain, such as the ISO 19119 (ISO/TC 211, 2005), use also semantically close geospatial properties to access and retrieve Geospatial Web services. With respect to the second one, archives also have the custody of important geospatial resources, which are susceptible to be analysed in order to provide assessment to their semantically close geographical properties. It seems natural to perform further research on these scenarios.
2. **Apply lessons learned in assessing the quality of semantically close geographic properties to other semantically close properties.** The invisibility problems caused by geospatial inconsistencies can also be generated when users search for other facets. An open line is to measure the level of impact of other semantically close properties such as (temporal, thematic, etc.) In this sense, we need to take into account at less two additional issues: (1) The development of knowledge organization systems, such as temporal and thematic ontologies, must be in accordance with the kind of semantically close properties

to be assessed. (2) Also, similarity measures must be developed to provide Quality Assessment for each kind of properties. The knowledge organization systems and ontologies used in this thesis provide an inventory of spatial entities existing at a time. However, services and spatial data are dynamic and change along time. For example, a Web service can change part of its contents between two short periods of time. Modelling the dynamic of some geographical resources and their content (e.g. Web services) along time is a complex problem (López-Pellicer, 2011). It will be interesting to investigate and detect spatial inconsistencies related with the time (e.g. from (x,y) to (x,y,t)).

3. **Extend the analysis using more fine grained Knowledge Organization Systems.** Part of the successful of the reverse geocoding process depends on the accuracy, the completeness and the level of detail of the gazetteers used, that is, the knowledge organization systems supporting the transformations (the spatial conversion between the reference systems). In our case, it has been the spatial ontologies used in the reverse geocoding process. Although, the results shown relevant results for the main cases, however, in areas where metadata documents refer to the smallest extents, it has been difficult to establish a spatial matching between the required/searched area and the spatial entities in the ontology. In one of the results of our research work (Moncla et al., 2014), we point out the need of official gazetteers and public spatial ontologies with a level of more fine-grained toponyms.
4. **Explore new Spatial Ranking methods for Reverse Geocoding in the context of two-dimensional datasets.** In this research work we use the concept of spatial ranking to transform (reverse geocoder) the Direct Spatial References into the most relevant Indirect Spatial References, which is referencing a location. Particularity, we use the notion of ranking query results based on the spatial similarity of two-dimensional footprints. Although, we have tested several measures of distance, new distances should be developed and tested to retrieve the resource with the best spatial matching. In our research we have found that search systems need improved measures for ranking better geospatial resources.

5. **Explore new two-dimensional clustering algorithms and metrics.** The collective metadata validation by means of clustering can be applied when we have additional information about neighbours with a good spatial consensus, that is to say, there must exist an agreement in the indirect spatial references that must describe the referenced location. Then the development of techniques to find this geospatial agreement in the presence of noisy and huge volumes of information is an open research issue. Many approaches deal with these problems in one dimension (resources referenced by a point), but the open research line and the challenge is to shift from 1D (a point) to 2D (MB-Box, multi-polygons and complex geometries) geographical footprint to assess their quality. The last point regarding two-dimensional clustering is the internal metric used to the co-occurrence of resources. When we have resources with two-dimensional footprints, the metric must measure the geospatial matching between the compared resources, that is to say, the spatial similarity. The clustering algorithm and the internal metric could be exchanged for another in order to find more accurate clusters, and then avoid potential errors.

6. **Apply lessons learned to Curation and Preservation processes.** Taking in mind, the notion of Digital Libraries lifetime, i.e. "a Digital Library provides access to information whose value is preserved across long periods of time" (Dragland, 2005), digital curation is a research field with many opportunities and challenges (Janée, 2009). We believe that geospatial Quality Assessment can help to the digital resource preservation across long periods of time. The increasing volume of accumulated geospatial resource in Digital Libraries will make it more necessary to ensure proper and consistent spatial descriptions. Preservation processes of datasets must include both, data and metadata, i.e. the assurance that in the future a resource will not be invisible, that is, the resource can be found among millions by means of the metadata used to describe, explore, geo-visualised and retrieve it. We hope that our research results will motivate data and metadata creator to ensure that metadata records are created and maintained consistent. The development of policies and Quality Assessment tools will help to ensure the efficient retrieval in future search systems.

6.3 Final Conclusion

Inconsistent metadata is often difficult to retrieve, especially, to complex query. The work developed in this thesis has shown that it is possible to detect inconsistencies in the context of geospatial digital repositories. It is done by applying geospatial Quality Assessment, particularly, assessment over the semantically close properties of the descriptive information. Metadata records in digital collections may become unretrievable due to inconsistencies between the semantically close properties of their metadata. In general, in information systems the efficiency of numerous tasks and processes depends on the consistency of the semantically close properties. In particular, processes such as discovery, retrieval, visualization, analysing, sharing and interoperability (e.g. Linked Data), curation, preservation, re-use, etc. In the geographical domain, geospatial Quality Assessment of the semantically close geographical properties can help to detect and fix inconsistencies.

Produce data, and in particular quality data is expensive. This explains why the re-use makes sense to reduce/share costs. However, it is required careful assessment of metadata descriptions that make the described resources easy to discover, share, and re-use for external consumers. It is often assumed by professionals that data management only entails preserving local consistency (not collective agreement or consensus about the proper description of a phenomenon). But this is not true. This thesis has shown that neglecting the quality of a pair of properties in a metadata record can cause serious problems of invisibility and retrievability. A resource without consistent metadata is for most purposes invisible and effectively lost. However, as this thesis presents, it is possible for large collections to make semi-automatic Quality Assessments able to detect those invisible records. Further research should analyse if this approach can be implement as an off-the-shelf component that can be added to popular information retrieval software.

Appendix A

Web Map Service Metadata

This appendix contains some details and examples of the crawled Web services metadata. Here, the WMS service is showed because it is the only service whose content is considered.

Example of a Web Map Service Capability document.

Listing A.1: Example of a Web Map Service Capability document.

```
1 <?xml version='1.0' encoding="ISO-8859-1" standalone="no" ?>
2 <WMS_Capabilities version="1.3.0" xmlns="http://www.opengis.net/wms"
3 xmlns:sld="http://www.opengis.net/sld"
4 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
5 xmlns:ms="http://mapserver.gis.umn.edu/mapserver"
6 xsi:schemaLocation="http://www.opengis.net/wms
7 http://schemas.opengis.net/wms/1.3.0/capabilities_1_3_0.xsd
8 http://www.opengis.net/sld http://schemas.opengis.net/sld/1.1.0/sld_capabilities.xsd
9 http://mapserver.gis.umn.edu/mapserver
10 http://ogc.larioja.org/wms/154trmb/request.php?service=WMS&
11 amp;version=1.3.0&request=GetSchemaExtension">
12
13 <Service>
14   <Name>WMS</Name>
15   <Title>IDERIOJA Torremontalbo [Spain] WMS</Title>
16   <Abstract>Servidor WMS del Municipio de Torremontalbo (La Rioja – Spain)</Abstract
17   >
18   <OnlineResource xmlns:xlink="http://www.w3.org/1999/xlink"
19     xlink:href="http://ogc.larioja.org/wms/154trmb/request.php?"/>
20   <MaxWidth>4096</MaxWidth>
```

```

20   <MaxHeight>4096</MaxHeight>
21 </Service>
22 <Capability>
23   <Request>
24     <GetCapabilities>
25       <Format>text/xml</Format>
26       <DCPType>
27         <HTTP>
28           <Get><OnlineResource xmlns:xlink="http://www.w3.org/1999/xlink"
29             xlink:href="http://ogc.larioja.org/wms/154trmb/request.
              php?"/></Get>
30           <Post><OnlineResource xmlns:xlink="http://www.w3.org/1999/xlink"
31             xlink:href="http://ogc.larioja.org/wms/154trmb/request.
              php?"/></Post>
32         </HTTP>
33       </DCPType>
34     </GetCapabilities>
35   <Layer>
36     <Name>TORREMONTALBO</Name>
37     <Title>IDERIOJA Torremontalbo [Spain] WMS</Title>
38     <Abstract>Servidor WMS del Municipio de Torremontalbo(La Rioja –
              Spain)</Abstract>
39     <CRS>EPSG:25830</CRS>
40     <CRS>EPSG:23030</CRS>
41     <CRS>EPSG:32630</CRS>
42     <CRS>EPSG:4230</CRS>
43     <CRS>EPSG:4258</CRS>
44     <CRS>EPSG:4326</CRS>
45     <EX_GeographicBoundingBox>
46       <westBoundLongitude>-2.70077</westBoundLongitude>
47       <eastBoundLongitude>-2.66832</eastBoundLongitude>
48       <southBoundLatitude>42.4924</southBoundLatitude>
49       <northBoundLatitude>42.5262</northBoundLatitude>
50     </EX_GeographicBoundingBox>
51     <MinScaleDenominator>100</MinScaleDenominator>
52     <MaxScaleDenominator>2.3e+008</MaxScaleDenominator>
53   </Layer>
54   :
55   :

```

Appendix B

Digital Library Metadata Records

This appendix contains some details and examples of the harvested metadata from the Library of Congress. This examples show the original XML documents with the inconsistencies. The next case presents an example of the common geospatial inconsistency found in the analysed collection. This example shows the geospatial inconsistency in the coordinate latitudes of the spatial footprint of the maritime boundary between Cuba and United States. These coordinate latitudes locate in the southern hemisphere the maritime boundaries of two countries belonging to the northern hemisphere. The Figure B.1 is the corresponding snapshot of the Website of the LoC.

Listing B.1: MARC21 records with inconsistency in the coordinate latitudes.

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <record xmlns="http://www.loc.gov/MARC21/slim" xmlns:cinclue="http://apache.org/
   cocoon/include/1.0" xmlns:zs="http://www.loc.gov/zing/srw/">
3   <leader>01197cem a2200325 a 4500</leader>
4   <controlfield tag="001">13004813</controlfield>
5   <controlfield tag="005">20130529075356.0</controlfield>
6   <controlfield tag="007">aj|canzn</controlfield>
7   <controlfield tag="008">021118s1978 dcuk bd a f 0 eng </controlfield>
8   <datafield tag="906" ind1=" " ind2=" " >
9     <subfield code="a">7</subfield>
10    <subfield code="b">cbc</subfield>
11    <subfield code="c">origcop</subfield>
12    <subfield code="d">u</subfield>
13    <subfield code="e">ncip</subfield>
14    <subfield code="f">20</subfield>
```

```
15     <subfield code="g">y-geogmaps</subfield>
16 </datafield>
17 <datafield tag="955" ind1=" " ind2=" ">
18     <subfield code="a">ga07 2002-11-19 sent to CMT</subfield>
19 </datafield>
20 <datafield tag="010" ind1=" " ind2=" ">
21     <subfield code="a"> 2002627930</subfield>
22 </datafield>
23 <datafield tag="034" ind1="1" ind2=" ">
24     <subfield code="a">a</subfield>
25     <subfield code="b">4650000</subfield>
26     <subfield code="d">W0860000</subfield>
27     <subfield code="e">W0800000</subfield>
28     <subfield code="f">S0260000</subfield>
29     <subfield code="g">S0220000</subfield>
30 </datafield>
31 <datafield tag="040" ind1=" " ind2=" ">
32     <subfield code="a">DLC</subfield>
33     <subfield code="c">DLC</subfield>
34     <subfield code="d">DLC</subfield>
35 </datafield>
36 <datafield tag="050" ind1="0" ind2="0">
37     <subfield code="a">G4921.F2 1978</subfield>
38     <subfield code="b">.U5</subfield>
39 </datafield>
40 <datafield tag="052" ind1=" " ind2=" ">
41     <subfield code="a">4921</subfield>
42 </datafield>
43 <datafield tag="052" ind1=" " ind2=" ">
44     <subfield code="a">3701</subfield>
45 </datafield>
46 <datafield tag="072" ind1=" " ind2="7">
47     <subfield code="a">F2</subfield>
48     <subfield code="2">lcg</subfield>
49 </datafield>
50 <datafield tag="110" ind1="1" ind2=" ">
51     <subfield code="a">United States.</subfield>
52     <subfield code="b">Department of State.</subfield>
53     <subfield code="b">Office of the Geographer.</subfield>
54 </datafield>
55 <datafield tag="245" ind1="1" ind2="0">
56     <subfield code="a">Maritime boundary, Cuba-United States.</subfield>
57 </datafield>
58 <datafield tag="255" ind1=" " ind2=" ">
```

```
59     <subfield code="a">Scale ca. 1:4,650,000. At 25N ;</subfield>
60     <subfield code="b">Mercator proj.</subfield>
61     <subfield code="c">(W86--W80/S26--S22).</subfield>
62 </datafield>
63 <datafield tag="260" ind1=" " ind2=" ">
64     <subfield code="a">[Washington, D.C. :</subfield>
65     <subfield code="b">Dept. of State, Office of the Geographer,</subfield>
66     <subfield code="c">1978]</subfield>
67 </datafield>
68 <datafield tag="300" ind1=" " ind2=" ">
69     <subfield code="a">1 map :</subfield>
70     <subfield code="b">col ;</subfield>
71     <subfield code="c">15 x 22 cm.</subfield>
72 </datafield>
73 <datafield tag="500" ind1=" " ind2=" ">
74     <subfield code="a">Includes coordinate table.</subfield>
75 </datafield>
76 <datafield tag="500" ind1=" " ind2=" ">
77     <subfield code="a">"3178 12--78 State (RGE) (2857)."</subfield>
78 </datafield>
79 <datafield tag="651" ind1=" " ind2="0">
80     <subfield code="a">Cuba</subfield>
81     <subfield code="x">Boundaries</subfield>
82     <subfield code="z">United States</subfield>
83     <subfield code="v">Maps.</subfield>
84 </datafield>
85 <datafield tag="651" ind1=" " ind2="0">
86     <subfield code="a">United States</subfield>
87     <subfield code="x">Boundaries</subfield>
88     <subfield code="z">Cuba</subfield>
89     <subfield code="v">Maps.</subfield>
90 </datafield>
91 <datafield tag="650" ind1=" " ind2="0">
92     <subfield code="a">Territorial waters</subfield>
93     <subfield code="z">Cuba</subfield>
94     <subfield code="v">Maps.</subfield>
95 </datafield>
96 <datafield tag="650" ind1=" " ind2="0">
97     <subfield code="a">Territorial waters</subfield>
98     <subfield code="z">United States</subfield>
99     <subfield code="v">Maps.</subfield>
100 </datafield>
101 </record>
```


LIBRARY OF CONGRESS | ASK A LIBRARIAN | DIGITAL COLLECTIONS | LIBRARY CATALOGS | Search **GO**

The Library of Congress > LC Online Catalog > LCCN Permalink

Print | Subscribe | Share/Save

LCCN Permalink [Browse](#) [Advanced Search](#) [Keyword Search](#)

Maritime boundary, Cuba-United States



MAP

Other Formats:
[Cite Record](#)
[MARCXML Record](#)
[MODS Record](#)
[Dublin Core Record](#)

More Information:
[LCCN Permalink FAQ](#)

2002627930

View record in the [LC Online Catalog](#) [Where to Request](#)

Corporate name [United States, Department of State, Office of the Geographer](#)

Main title Maritime boundary, Cuba-United States.

Published/Created [Washington, D.C. : Dept. of State, Office of the Geographer, 1978]

Description
1 map : col ; 15 x 22 cm.

Scale info
Scale ca. 1:4,650,000. At 25°N ; Mercator proj. (W 86°-W 80°/S 26°-S 22°).

LC classification
G4921.F2 1978 .U5

Subjects
[Cuba--Boundaries--United States--Maps.](#)
[United States--Boundaries--Cuba--Maps.](#)
[Territorial waters--Cuba--Maps.](#)
[Territorial waters--United States--Maps.](#)

Notes
Depths shown by contours.
Includes coordinate table.
"3178 12-78 State (RGE) (2857)."

LC control no.
2002627930

Geographic class no.
4921
3701

Type of material
Map

Figure B.1.: Snapshot of the Website of the LoC that shows the geospatial inconsistency in the coordinate latitudes.

The second case presents an example of geospatial inconsistency found in the Digital Library collection analysed. This example shows the geospatial inconsistency in the coordinate longitudes of the spatial footprint of the Caledonia County - Vermont -United States. These coordinate longitudes locate the Caledonia county over Asia continent, over the Kyrgyztan country specifically. The Figure B.2 is the corresponding snapshot of the Website of the LoC.

Listing B.2: MODS metadata records with a geospatial inconsistency in the coordinate longitude.

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <mods xmlns="http://www.loc.gov/mods/v3" xmlns:xsi="http://www.w3.org/2001/
  XMLSchema-instance"
3 xsi:schemaLocation="http://www.loc.gov/mods/v3
4 http://www.loc.gov/standards/mods/v3/mods-3-5.xsd" version="3.5">
5   <titleInfo>
6     <title>Map of Caledonia County, Vermont</title>
7   </titleInfo>
8   <name type="personal" usage="primary">
9     <namePart>Walling, Henry Francis</namePart>
10    <namePart type="date">1825-1888</namePart>
11  </name>
12  <typeOfResource>cartographic</typeOfResource>
13  <genre authority="marcgt">map</genre>
14  <originInfo>
15    <place>
16      <placeTerm type="code" authority="marccountry">nyu</
        placeTerm>
17    </place>
18    <place>
19      <placeTerm type="text">New York</placeTerm>
20    </place>
21    <publisher>Baker & Tilden</publisher>
22    <dateIssued>1858</dateIssued>
23    <issuance>monographic</issuance>
24  </originInfo>
25  <language>
26    <languageTerm type="code" authority="iso639">eng</languageTerm>
27  </language>
28  <physicalDescription>
29    <form authority="marccategory">electronic resource</form>
30    <form authority="marcsmd">remote</form>

```

```

31         <form authority="marccategory">map</form>
32         <form authority="marcsmd">map</form>
33         <extent>1 map : col. ; x 132 cm.</extent>
34     </physicalDescription>
35     <note type="statement of responsibility" altRepGroup="00">
36         from actual surveys under the direction of H.F. Walling, 1858.
37     </note>
38     <note>33 inset maps, views, tables, etc.</note>
39     <note type="additional physical form">
40     Available also through the Library of Congress Web site as a raster image.
41     </note>
42     <subject>
43         <cartographics>
44             <scale>Scale 1:50,000</scale>
45             <coordinates>(E072 30'00"—E071 45'00"/N42 45'00"—N42
46                 10'00").</coordinates>
47         </cartographics>
48     </subject>
49     <subject authority="lsh">
50     <geographic>Caledonia County (Vt.)</geographic>
51     <genre>Maps</genre>
52     </subject>
53     <subject>
54         <hierarchicalGeographic>
55             <country>United States</country>
56             <state>Vermont</state>
57             <county>Caledonia County</county>
58         </hierarchicalGeographic>
59     </subject>
60     <classification authority="lcc">G3753 .C3 1858 .W3</classification>
61     <location>
62         <physicalLocation>
63             Library of Congress Geography and Map Division Washington,
64             D.C. 20540—4650 USA
65         </physicalLocation>
66     </location>
67     <location>
68         <url displayLabel="Copy 1" usage="primary display">http://hdl.loc.gov/
69             loc.gmd/g3753c.la001184</url>
70     </location>
71     <relatedItem type="isReferencedBy">
72         <titleInfo>
73             <title>LC Land ownership maps,</title>
74         </titleInfo>

```



```
72         <part>
73             <detail type="part">
74                 <number>1184</number>
75             </detail>
76         </part>
77     </relatedItem>
78     <identifier type="lcn">2005625342</identifier>
79     <identifier type="hdl">hdl:loc.gmd/g3753c.la001184</identifier>
80     <recordInfo>
81         <descriptionStandard>aacr</descriptionStandard>
82         <recordContentSource authority="marcorg">DLC</recordContentSource>
83         <recordCreationDate encoding="marc">051130</recordCreationDate>
84         <recordChangeDate encoding="iso8601">20120914134540.0</
            recordChangeDate>
85         <recordIdentifier>14184235</recordIdentifier>
86         <recordOrigin>
87             Converted from MARCXML to MODS version 3.5 using
            MARC21slim2MODS3-5.xsl (Revision 1.96 2014/04/22)
88         </recordOrigin>
89     </recordInfo>
90 </mods>
```

LIBRARY OF CONGRESS

ASK A LIBRARIAN DIGITAL COLLECTIONS LIBRARY CATALOGS


Search Search Loc.gov GO

The Library of Congress > LC Online Catalog > LCCN Permalink

Print Subscribe Share/Save

LCCN Permalink LC Online Catalog Quick Search Search Browse Advanced Search Keyword Search

Map of Caledonia County, Vermont



MAP

Other Formats:

[Cite Record](#)

[MARCXML Record](#)

[MODS Record](#)

[Dublin Core Record](#)

More Information:

[LCCN Permalink FAQ](#)

2005625342

View record in the [LC Online Catalog](#) [Where to Request](#)

Personal Name [Walling, Henry Francis, 1825-1888.](#)

Main title Map of Caledonia County, Vermont / from actual surveys under the direction of H.F. Walling, 1858.

Published/Created New York : Baker & Tilden, 1858.

Description
1 map : col. ; x 132 cm.

Scale info
Scale 1:50,000 (E072°30'00"-E071°45'00"/N42°45'00"-N42°10'00").

Links
Copy 1: <http://hdl.loc.gov/loc/qmd/q3753c.la001184>

LC classification
G3753 .C3 1858 .W3

Subjects
[Caledonia County \(Vt.\)--Maps.](#)
United States--Vermont--Caledonia County.

Notes
33 inset maps, views, tables, etc.

References
LC Land ownership maps, 1184

Additional formats
Available also through the Library of Congress Web site as a raster image.

LC control no.
2005625342

Repository
Library of Congress Geography and Map Division Washington, D.C. 20540-4650 USA dcu

Type of material
Map

Figure B.2.: Snapshot of the Website of the LoC that shows the geospatial inconsistency in the coordinate longitudes.

Appendix C

Contributions

- 2014 Improving the geospatial consistency of digital libraries meta-data** (Walter Renteria-Agualimpia, Francisco Javier López-Pellicer, Javier Lacasta, Pedro Rafael Muro-Medrano, F. Javier Zarazaga-Soria), *Journal of Information Science*.
- 2014 Identifying geospatial inconsistency of web services meta-data using spatial ranking** (Walter Renteria-Agualimpia, Francisco Javier Lopez-Pellicer, Javier Lacasta, Pedro Rafael Muro-Medrano, F. Javier Zarazaga-Soria), *In Earth Science Informatics*, Springer Berlin Heidelberg, pp. 1-11, 2014.
- 2014 Agregador automático de servicios web geoespaciales** (Javier Lacasta, Javier Lopez-Pellicer, Walter Renteria-Agualimpia, Javier Nogueras-Iso), *In Scire: representación y organización del conocimiento*, volume 20, pp. 43-48, 2014.
- 2014 Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus** (Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, Mauro Gaiò), *In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, Texas, USA, November 4-7, 2014*.
- 2014 Improving the visibility of geospatial data on the Web** (Javier Lacasta, Javier Lopez-Pellicer, Walter Renteria-Agualimpia, Javier

- Nogueras-Iso), *In Proceedings of Digital Libraries 2014: ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014) and International Conference on Theory and Practice of Digital Libraries (TPDL 2014)*, London, September 8-12th 2014, 2014.
- 2014 Linked Map: maps, Linked Data and INSPIRE** (Francisco J. López-Pellicer, Javier Lacasta, Walter Renteria-Agualimpia, Jesús Barrera, Juan López-Larrinzar, Jose Agudo), *INSPIRE Conference 2014, Infrastructure for Spatial Information in the European Community, Aalborg, Denmark, 18-20 June 2014.*
- 2013 Detectando anomalías en los metadatos de cartotecas** (Walter Renteria-Agualimpia, Francisco Javier López Pellicer, Aneta J Florczyk, Juan López de Larrinzar, Javier Lacasta, Pedro R Muro-Medrano, Francisco Javier Zarazaga Soria), *In Scire: representación y organizacin del conocimiento*, volume 19, pp. 23-29, 2013.
- 2013 Aproximación geosemántica para detectar inconsistencias en los metadatos de Servicios Web Geoespaciales** (Walter Renteria-Agualimpia, Francisco Javier López Pellicer, Javier Lacasta, Pedro R Muro-Medrano, Francisco Javier Zarazaga Soria), *In GeoFocus: International Review of Geographical Information Science and Technology*, volume 13, pp. 154-176, 2013.
- 2013 Identifying hidden geospatial resources in catalogues** (Walter Renteria-Agualimpia, Francisco J Lopez-Pellicer, Javier Lacasta, F Javier Zarazaga-Soria, Pedro R Muro-Medrano), *In Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, June 12-14, Madrid, Spain*, pp. 32, 2013.
- 2012 Availability of the OGC geoprocessing standard: March 2011 reality check** (Francisco J Lopez-Pellicer, Walter Renteria-Agualimpia, Rubén Béjar, Pedro R Muro-Medrano, F Javier Zarazaga-Soria), *In Computers & Geosciences*, Elsevier, volume 47, pp. 13-19, 2012.
- 2012 Towards an active directory of Geospatial Web services** (F. J. Lopez-Pellicer, W. Renteria-Agualimpia, J. Nogueras-Iso, F. J.

Zarazaga-soria, P. R. Muro-Medrano.), *In Bridging the Geographic Information Sciences*, pp. 63-80, 2012.

- 2011 Status of INSPIRE inspired OGC Web Services** (FJ Lopez-Pellicer, R Béjar, W Renteria-Agualimpia, AJ Florczyk, PR Muro-Medrano, FJ Zaragoza-Soria), *In INSPIRE Conference 2011: INSPIRED by 2020 - Contributing to smart, sustainable and inclusive growth. Edinburgh, Scotland, 27 June - 1 July, 2011.*
- 2011 Publishing standard geospatial catalogues in the Web of Data** (Francisco J Lopez-Pellicer, Aneta J Florczyk, Walter Renteria-Agualimpia, Javier Nogueras-Iso, Pedro R Muro-Medrano), *In 14th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2011, San Cristóbal de La Laguna, Tenerife, Spain, November 7-10, 2011.*
- 2011 Implantation of OGC geoprocessing services for Geoscience** (Francisco J López-Pellicer, Walter Renteria-Agualimpia, Rubén Béjar, Juan Valiño, F Javier Zarazaga-Soria, Pedro R Muro-Medrano), *In Actas de las II Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE'2011), Barcelona, Spain, November 9-11, 2011.*
- 2011 CSW2LD: a Linked Data frontend for CSW** (Francisco J López-Pellicer, Aneta J Florczyk, Walter Rentería-Aguaviva, Javier Nogueras-Iso, Pedro R Muro-Medrano), *In Actas de las II Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE'2011), Barcelona, November 9-11, 2011.*
- 2011 Multi-criteria geographic information retrieval model based on geospatial semantic integration** (Walter Renteria-Agualimpia, Sergei Levashkin), *In Chapter in GeoSpatial Semantics*, Springer, pp. 166-181, 2011.
- 2010 Anclando "La Balsa de Piedra" digital de las IDEs al mundo de la Web Semántica** (Francisco J Lopez-Pellicer, W Rentería-Agualimpia, Juan Valiño, Rubén Béjar, Javier Nogueras-Iso, Pedro R Muro-Medrano), *In booktitle=Actas de las I Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE'2010) Lisboa Portugal October 27-29, 2010.*

2010 Exploring the advances in semantic search engines (Walter Renteria-Agualimpia, Francisco J López-Pellicer, Pedro R Muro-Medrano, Javier Nogueras-Iso, F Javier Zarazaga-Soria), *In Chapter in International Symposium on Distributed Computing and Artificial Intelligence 2010 (DCAI'2010). Advances in Intelligent and Soft-Computing. Springer, Springer, pp. 613-620, 2010.*

Bibliography

- D. J. Abel, V. J. Gaede, K. L. Taylor, and X. Zhou. Smart: Towards spatial internet marketplaces. *Geoinformatica*, 3(2):141–164, 1999.
- L. Al-Hakim. *Information quality management: theory and applications*. Idea Group Inc (IGI) Global, 2007.
- W. Aldis, G. Rockenschaub, Y. Gorokhovich, S. Doocy, P. Lumbiganon, and F. Grunewald. Panel 2.1: assessing impact and needs. *Prehospital and disaster medicine*, 20(06):396–398, 2005.
- M. J. Atallah. A linear time algorithm for the hausdorff distance between convex polygons. *Information Processing Letters*, 17(4):207–209, 1983.
- R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.
- J. M. Barrueco and I. S. Coll. Open Archives Initiative. Protocol for Metadata Harvesting (OAI-PMH): descripción, funciones y aplicaciones de un protocolo. *El profesional de la información*, 12(2):99–106, 2003.
- J. Barton, S. Currier, and J. Hey. Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. In *2003 Dublin Core Conference: Supporting Communities of Discourse and Practice - Metadata Research and Applications*, Seattle, WA., 2003. URL <http://eprints.erpanet.org/83/>. Last visited on 12.11.2014.
- P. Bassoullet, R. Djuwansah, D. Gouleau, and C. Marius. Hydrosedimentological processes and soils of the Barito estuary (South-Kalimantan, Indonesia). *Oceanologica acta*, 9(3):217–226, 1986.

- J. Beall. Metadata and data quality problems in the digital library. *Journal of Digital Information*, 6(3), 2006.
- K. Beard and V. Sharma. Multidimensional ranking for data in digital spatial libraries. *International Journal on Digital Libraries*, 1(2):153–160, 1997.
- Y. Bédard, T. Merrett, and J. Han. Fundamentals of spatial data warehousing for geographic knowledge discovery. *Geographic data mining and knowledge discovery*, 2:53–73, 2001.
- R. Béjar, P. R. Muro-Medrano, and J. Noguera-Iso. *Contributions to the modelling of spatial data infrastructures and their portrayal services*. PhD thesis, PhD Dissertation, Computer Science and Systems Engineering Department, University of Zaragoza, Spain, 2009.
- B. W. Boehm. A spiral model of software development and enhancement. *Computer*, 21(5):61–72, 1988.
- M. N. Boulos. On geography and medical journalology: a study of the geographical distribution of articles published in a leading medical informatics journal between 1999 and 2004. *International journal of health geographics*, 4(1):7, 2005.
- N. R. Brisaboa, M. R. Luaces, M. Andrea Rodríguez, and D. Seco. An inconsistency measure of spatial data sets with respect to topological constraints. *International Journal of Geographical Information Science*, 28(1):56–82, 2014.
- D. Broeder and P. Wittenburg. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1(2):119–132, 2006.
- T. R. Bruce and D. I. Hillmann. The continuum of metadata quality: defining, expressing, exploiting. In *Metadata in Practice*, Edited by Diane I. Hillmann and Elaine L. Westbrooks. Chicago: American Library Association. ALA editions, 2004.
- O. Buchel and L. L. Hill. Treatment of georeferencing in knowledge organization systems: North american contributions to integrated georeferencing. *Knowledge organization*, 37(1):72–78, 2010.
- M. Buckland, A. Chen, F. C. Gey, R. R. Larson, R. Mostern, and V. Petras. Geographic search: catalogs, gazetteers, and maps. *College & Research Libraries*, 68(5):376–387, 2007.
- P. Caplan. *Metadata fundamentals for all librarians*. American Library Association, 2003.

- A. Chandler, D. Foley, and A. M. Hafez. Mapping and Converting Essential Federal Geographic Data Committee (FGDC) Metadata into MARC21 and Dublin Core. *D-Lib magazine*, 6(1):1082–9873, 2000.
- N. Chrisman. The error component in spatial data. *Geographical information systems*, 1:165–174, 1991.
- P. Clough, J. Tang, M. M. Hall, and A. Warner. Linking archival data to location: a case study at the uk national archives. In *Aslib Proceedings*, volume 63, pages 127–147. Emerald Group Publishing Limited, 2011.
- A. G. Cohn and S. M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta informaticae*, 46(1):1–29, 2001.
- Commission of the European Communities (CEC). Directive of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Council Directive (EC) 2007/2/EC, Commission of the European Communities (CEC), 2007.
- G. Crane. The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. *D-Lib Magazine*, Jan. 1998. URL <http://www.dlib.org/dlib/january98/01crane.html>. Last visited on 02.05.2013.
- R. Denenberg. Sru (search/retrieve via url),. In *The Library of Congress, Washington, DC*, 2007. URL <http://www.loc.gov/standards/sru/>. Last visited on 05.11.2013.
- R. Devillers, R. Jeansoulin, et al. *Fundamentals of spatial data quality*. ISTE London, 2006.
- J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of Web resources. Technical report, Department of Computer Science, Columbia University, 2000. URL <http://hdl.handle.net/10022/AC:P:29391>. Last visited on 03.01.2014.
- K. Dragland. Adding a local node to a global georeferenced digital library. Master’s thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, Norway, 2005. URL <http://www.diva-portal.org/smash/get/diva2:348020/FULLTEXT01.pdf>.
- M. Duckham and J. Drummond. Assessment of error in digital vector data using fractal geometry. *International Journal of Geographical Information Science*, 14(1):67–84, 2000.

- N. Dushay and D. I. Hillmann. Analyzing metadata for effective use and re-use. In *DCMI Metadata Conference and Workshop, Seattle*. Dublin Core Metadata Initiative, 2003. URL <http://hdl.handle.net/1813/7896>. Last visited on 03.11.2014.
- E. Duval, W. Hodgins, S. Sutton, and S. L. Weibel. Metadata principles and practicalities. *D-lib Magazine*, 8(4):16, 2002.
- M. J. Egenhofer. Toward the semantic geospatial Web. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pages 1–4. ACM, 2002.
- M. J. Egenhofer and R. D. Franzosa. On the equivalence of topological relations. *International Journal of Geographical Information Systems*, 9(2):133–152, 1995.
- M. J. Eppler. *Managing information quality: increasing the value of information in knowledge-intensive products and processes*. Springer, 2006.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.
- European Commission. Public sector information: A key resource for europe. green paper on public sector information in the information society. com (98) 585 final, 20 january 1999, 1998. URL <http://aei.pitt.edu/1168/>.
- M. Fernández-López, A. Gómez-Pérez, and N. Juristo. Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering*, pages 33–40, 1997.
- FGDC. Ten most common metadata errors. Technical report, Federal Geographic Data Committee, 1998a. URL <http://www.fgdc.gov/metadata/documents/top10metadataerrors.pdf>. Last visited on 03.08.2013.
- FGDC. *Content standard for digital geospatial metadata*. Federal Geographic Data Committee, 1998b. URL <http://www.fgdc.gov/metadata/csdlgm/>. Last visited on 15.04.2013.
- A. J. Florczyk. *Search improvement within the geospatial Web in the context of spatial data infrastructures*. PhD thesis, Universidad de Zaragoza, 2012.
- A. U. Frank. Analysis of dependence of decision quality on data quality. *Journal of geographical systems*, 10(1):71–88, 2008.

- P. Frontiera, R. Larson, and J. Radke. A comparison of geometric approaches to assessing spatial similarity for gir. *International Journal of Geographical Information Science*, 22(3):337–360, 2008.
- B. Furrie. Understanding marc bibliographic: machine-readable cataloging. Cataloging Distribution Service, Library of Congress, in collaboration with the Follett Software Company, 2009. URL <http://www.loc.gov/marc/umb/>. Last visited on 14.11.2014.
- D. Garvin. *Managing Quality: The Strategic Competitive Edge*. New York: The Free Press, 1988.
- J. R. Giles. Geoscience metadata—no pain, no gain. *Geological Society of America Special Papers*, 482:29–33, 2011.
- P. Gong and L. Mu. Error detection through consistency checking. *Geographic Information Sciences*, 6(2):188–193, 2000.
- M. F. Goodchild. Report on a workshop on metadata held in santa barbara, ca, november 8, 1995. santa barbara, ca: Alexandria digital library ([www:http://alexandria.sdc.ucsb.edu](http://alexandria.sdc.ucsb.edu)), 1995. URL <http://www.geog.ucsb.edu/~good/papers/251.pdf>. Last visited on 04.07.2012.
- M. F. Goodchild and J. Zhou. Finding Geographic Information: Collection-Level Metadata. *GeoInformatica*, 7(2):95–112, 2003.
- M. F. Goodchild, P. Fu, and P. Rich. Sharing geographic information: an assessment of the geospatial one-stop. *Annals of the Association of American Geographers*, 97(2):250–266, 2007.
- R. K. Goyal and M. J. Egenhofer. Similarity of cardinal directions. In *Advances in Spatial and Temporal Databases*, pages 36–55. Springer, 2001.
- K. E. Grossner, M. F. Goodchild, and K. C. Clarke. Defining a digital earth system. *Transactions in GIS*, 12(1):145–160, 2008.
- R. Grütter and B. Bauer-Messmer. Towards spatial reasoning in the semantic Web: A hybrid knowledge representation system architecture. In *The European Information Society*, pages 349–364. Springer, 2007.
- Z. Gui, C. Yang, J. Xia, K. Liu, C. Xu, J. Li, and P. Lostritto. A performance, semantic and service quality-enhanced distributed search engine for improving geospatial resource discovery. *International Journal of Geographical Information Science*, 27(6):1109–1132, 2013.

- J. Hartmann and H. Stuckenschmidt. Automatic metadata analysis for environmental information systems. In *Proceedings of the International Symposium on Environmental Informatics*, 2002.
- J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
- A. G. Herrero and J. M. Ruiz. Do trade and financial links foster business cycle synchronization in a small economy? *Moneda y Credito*, 226(1):187, 2008.
- I. Heywood, S. Cornelius, and S. Carver. *An Introduction to Geographic Information Systems*. New York: Addison Wesley Longman, 1998.
- L. L. Hill. *Access to geographic concepts in online bibliographic files: effectiveness of current practices and the potential of a graphic interface*. PhD thesis, University of Pittsburgh, 1990.
- L. L. Hill. *Georeferencing: The Geographic Associations of Information (Digital Libraries and Electronic Publishing)*. The MIT Press, 2006.
- L. L. Hill and G. Janée. The alexandria digital library project: Metadata development and use. *Metadata in Practice: a Work in Progress*, 2004.
- D. I. Hillmann. Metadata quality: From evaluation to augmentation. *Cataloging & Classification Quarterly*, 46(1):65–80, 2008.
- D. I. Hillmann, N. Dushay, and J. Phipps. Improving metadata quality: augmentation and recombination. In *DC-2004, Shanghai, China*. Dublin Core Metadata Initiative, 2004.
- G. Hodge. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. ERIC, 2000.
- D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the Hausdorff distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(9):850–863, 1993.
- T. Idowu and J. Sambo. Development of quality assessment tool for geospatial data acquisition in surveying and geoinformatics. *Journal of Emerging Trends in Engineering and Applied Sciences*, 3(1):114–120, 2012.
- INSPIRE. Infrastructure for spatial information in european community, 2013. URL <http://inspire.ec.europa.eu/>. Last visited on 05.07.2014.

- A. Isaac and E. Summers. SKOS Simple Knowledge Organization System Primer, W3C Working Group Note 18 August 2009, 2009.
- ISO. ISO 9000: Quality management systems - Fundamentals and vocabulary. Published standard, International Organization for Standardization, October 2005.
- ISO/TC 211. ISO 19113:2002. Geographic information – Quality principles. Published standard ISO 19113:2002, International Organization for Standardization, October 2002.
- ISO/TC 211. ISO 19114:2003 Geographic information – Quality evaluation procedures. Published standard ISO 19114:2003, International Organization for Standardization, October 2003.
- ISO/TC 211. ISO 19119:2005 Geographic Information – Services. Published standard ISO 19119:2005, International Organization for Standardization, October 2005.
- ISO/TC 211. ISO/TS 19138:2006 Geographic information – Data quality measures. Published standard ISO 19138:2006, International Organization for Standardization, October 2006.
- A. Jakobsson. *On the Future of Topographic Base Information Management in Finland and Europe*. PhD thesis, Helsinki University of Technology, 2006.
- G. Janée. Spatial similarity functions, 2003. URL <http://www.alexandria.ucsb.edu/~gjanee/archive/2003/similarity.html>. Last visited on 08.10.2013.
- G. Janée. Digital curation. In *Encyclopedia of Database Systems*, pages 816–817. Springer, 2009.
- K. Janowicz, S. Schade, A. Bröring, C. Keßler, P. Maué, and C. Stasch. Semantic enablement for spatial data infrastructures. *Transactions in GIS*, 14(2):111–129, 2010.
- K. Janowicz, M. Raubal, and W. Kuhn. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, (2):29–57, 2011.
- K. Järvelin. IR research: systems, interaction, evaluation and theories. In *Proceedings of the 33rd European conference on Advances in information retrieval, ECIR'11*, pages 1–3, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-20160-8.
- C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. Van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies an overview of the spirit project. In *Proceedings of the 25th annual international ACM SIGIR*

- conference on Research and development in information retrieval, pages 387–388. ACM, 2002.
- D. Joshi, A. K. Samal, and L.-K. Soh. Density-based clustering of polygons. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, pages 171–178. IEEE, 2009.
- J. M. Juran. Quality control handbook. In *Quality control handbook*. McGraw-Hill, 1962.
- J. M. Juran. How to think about quality. *JM Juran, AB Godfrey, RE Hoogstoel, and EG, Schilling (Eds.): Quality-Control Handbook*. New York: McGraw-Hill, 1999.
- B. K. Kahn, D. M. Strong, and R. Y. Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4):184–192, 2002.
- W. Kainz. Logical consistency. *Elements of spatial data quality*, pages 109–137, 1995.
- V. Kanagavalli and K. Raja. A fuzzy logic based method for efficient retrieval of vague and uncertain spatial expressions in text exploiting the granulation of the spatial event queries. *CoRR*, 2013.
- B. D. Klein, D. L. Goodhue, and G. B. Davis. Can humans detect errors in data? impact of base rates, incentives, and goals. *MIS Quarterly*, pages 169–194, 1997.
- S.-A. Knight and J. M. Burn. Developing a framework for assessing information quality on the world wide Web. *Informing Science: International Journal of an Emerging Transdiscipline*, 8(5):159–172, 2005.
- W. Kresse and D. M. Danko. *Springer handbook of geographic information*. Springer, 2012.
- J. Lacasta, J. López-Pellicer, W. Renteria-Agualimpia, and J. Nogueras-Iso. Agregador automático de servicios Web geoespaciales. *Scire: representación y organización del conocimiento*, 20(2):43–48, 2014a.
- J. Lacasta, J. López-Pellicer, W. Renteria-Agualimpia, and J. Nogueras-Iso. Improving the visibility of geospatial data on the Web. In *Proceedings of Digital Libraries 2014: ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014) and International Conference on Theory and Practice of Digital Libraries (TPDL 2014), London, September 8-12th 2014*. ACM/IEEE, 2014b.
- C. Larman and V. R. Basili. Iterative and incremental development: A brief history. *Computer*, 36(6):47–56, 2003.

- R. R. Larson. Ranking approaches for gir. *SIGSPATIAL Special*, 3(2):37–41, 2011.
- J. L. Leidner. Towards a reference corpus for automatic toponym resolution evaluation. In *Workshop on Geographic Information Retrieval, Sheffield, UK*, 2004.
- J. Leveling. Challenges for indexing in GIR. *SIGSPATIAL Special*, 3(2):29–32, July 2011. ISSN 1946-7729. doi: 10.1145/2047296.2047303. URL <http://doi.acm.org/10.1145/2047296.2047303>.
- B. Li and F. Fonseca. Tdd: A comprehensive model for qualitative spatial similarity assessment. *Spatial Cognition and Computation*, 6(1):31–62, 2006.
- R. A. Longhorn. Geospatial standards, interoperability, metadata semantics and spatial data infrastructure. In *Proceedings NIEeS Workshop on Activating Metadata, July 6-7 2005, Cambridge, UK*, pages 1–23, 2005.
- F. López-Pellicer, R. Béjar, W. Renteria-Agualimpia, A. Florczyk, P. Muro-Medrano, and F. Zaragoza-Soria. Status of INSPIRE inspired OGC Web Services. In *INSPIRE Conference*, 2011.
- F. J. López-Pellicer. *Semantic Linkage of the Invisible Geospatial Web*. PhD thesis, Universidad de Zaragoza, 2011.
- F. J. López-Pellicer, J. Lacasta, A. Florczyk, J. Noguera-Iso, and F. J. Zarazaga-Soria. An ontology for the representation of spatiotemporal jurisdictional domains in information retrieval systems. *International Journal of Geographical Information Science*, 26(4):579–597, 2012a.
- F. J. López-Pellicer, W. Renteria-Agualimpia, R. Béjar, P. R. Muro-Medrano, and F. J. Zarazaga-Soria. Availability of the OGC geoprocessing standard: March 2011 reality check. *Computers & Geosciences*, 47:13–19, 2012b.
- F. J. López-Pellicer, W. Renteria-Agualimpia, J. Noguera-Iso, F. J. Zarazaga-soria, and P. R. Muro-Medrano. Towards an active directory of geospatial Web services. *Bridging the Geographic Information Sciences*, pages 63–80, 2012c.
- M. Lutz. *Ontology-based discovery and composition of geographic information services*. PhD thesis, Institute for Geoinformatics, University of Munster, Germany, 2005.
- S. Ma, C. Lu, X. Lin, and M. Galloway. Evaluating the metadata quality of the ipl. *Proceedings of the American Society for Information Science and Technology*, 46(1): 1–17, 2009.

- B. Martins and P. Calado. Learning to rank for geographic information retrieval. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, page 21. ACM, 2010.
- B. Martins, M. J. Silva, S. Freitas, and A. P. Afonso. Handling locations in search engine queries. In *3rd ACM Workshop on Geographic Information Retrieval, GIR 2006, Seattle, WA, USA*, volume 6, pages 1–6, 2006. URL <http://www.geo.uzh.ch/~rsp/gir06/papers/individual/martins.pdf>. Last visited on 12.11.2014.
- B. Martins, J. Borbinha, G. Pedrosa, J. Gil, and N. Freire. Geographically-aware information retrieval for collections of digitized historical maps. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, pages 39–42. ACM, 2007.
- S. Mäs. *On the consistency of spatial semantic integrity constraints*. PhD thesis, Bundeswehr University Munich, 2009.
- MDWG. Content standard for digital geospatial metadata (revised june 1998). Standard FGDC-STD-001-1998, Metadata Ad Hoc Working Group (MDWG), Federal Geographic Data Committee, Washington, D.C., 1998. URL <http://www.fgdc.gov/metadata/csdgm/>.
- R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *American Association for Artificial Intelligence AAAI*, volume 6, pages 775–780, 2006. URL <http://www.aaai.org/Papers/AAAI/2006/AAAI06-123.pdf>. Last visited on 12.11.2014.
- A. Miles and S. Bechhofer. Skos simple knowledge organization system reference. *W3C recommendation*, 18:W3C, 2009. URL <http://www.w3.org/2009/08/skos-reference/skos.html>.
- E. Miller. An introduction to the resource description framework. *Bulletin of the American Society for Information Science and Technology*, 25(1):15–19, 1998.
- H. Moellering and R. Hogan. *Spatial Database Transfer Standards 2: Characteristics for Assessing Standards and Full Descriptions of the National and International Standards in the World: The ICA Commission on Standards for the Transfer of Spatial Data*. International Cartographic Association. Elsevier Science, 1997. ISBN 9780080541525. URL <http://books.google.es/books?id=kcDxdD3gTkC>.
- W. E. Moen, E. L. Stewart, and C. R. McClure. Assessing metadata quality: Findings and methodological considerations from an evaluation of the us government information locator service (gils). In *Research and Technology Advances in Digital*

- Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on*, pages 246–255. IEEE, 1998.
- L. Moncla, W. Renteria-Agualimpia, J. Nogueras-Iso, and M. Gaio. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, Texas, USA, November 4-7*. ACM, 2014.
- F. Naumann. *Quality-driven query answering for integrated information systems*, volume 2261. Springer, 2002.
- F. Naumann and C. Rolker. Assessment methods for information quality criteria. In *IQ-00, 5th International Conference on Information Quality*, number 138 in Informatik-Berichte, pages 148–162. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik, 2000. URL <http://edoc.hu-berlin.de/docviews/abstract.php?id=25314>. Last visited on 12.11.2014.
- D. Nebert, A. Whiteside, and P. Vretanos. Open GIS Catalogue Services Specification. OpenGIS Publicly Available Standard OGC-07-006r1, Open GIS Consortium Inc., February 2007. Version 2.0.2.
- D. D. Nebert. *Developing Spatial Data Infrastructures: The SDI Cookbook*. Global Spatial Data Infrastructure, 2004.
- J. Nogueras-Iso, F. J. Zarazaga-Soria, J. Lacasta, R. Béjar, and P. R. Muro-Medrano. Metadata standard interoperability: application in the geographic information domain. *Computers, environment and urban systems*, 28(6):611–634, 2004.
- J. Nogueras-Iso, F. J. Zarazaga-Soria, and P. R. Muro-Medrano. *Geographic information metadata for spatial data infrastructures*. Springer, 2005.
- M. Oehrli, P. Pridal, S. Zollinger, and R. Siber. Maprank: Geographical search for cartographic materials in libraries. *D-lib Magazine*, 17(9/10), 2011.
- P. V. Oort. *Spatial data quality: from description to application*. Publications on Geodesy 60. Netherlands Geodetic Commission, Delf, 2005.
- F. M. R. Pardo, L. R. Pardo, D. Buscaldi, and P. Rosso. Gir pharma: a geographic information retrieval approach to locate pharmacies on duty. In *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*, page 33. ACM, 2010.

- J.-R. Park. Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47(3-4):213–228, 2009.
- V. Petras. Statistical analysis of geographic and language clues in the marc record. Technical report, Technical Report, 2004. URL <http://metadata.sims.berkeley.edu/papers/Marcplaces.pdf>. Last visited on 12.04.2013.
- M. Piasecki, L. Bermudez, B. Beran, S. Islam, Y.-R. Choi, X. Liang, and S. Jeong. Hydrologic metadata. *Hydrologic Information System Status Report*, 88, 2010.
- J. Powell, K. Mane, L. M. Collins, M. L. Martinez, and T. McMahon. The geographic awareness tool: techniques for geo-encoding digital library content. *Library Hi Tech News*, 27(9/10):5–9, 2010.
- T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web (TWeb)*, 3(1):1, 2009.
- T. C. Redman. *Data quality: the field guide*. Digital Press, 2001.
- R. Register, K. Cohn, L. Hawkins, H. Henderson, R. Reynolds, S. C. Shadle, W. Hoffman, S. Rajan, and P. W. Yue. Metadata in a digital age: new models of creation, discovery, and use. *The Serials Librarian*, 56(1-4):7–24, 2009.
- W. Renteria-Agualimpia. Precision-controlled retrieval of qualitative information from data repositories (in spanish) recuperación controlada de información cualitativa desde repositorios de datos. Master’s thesis, Instituto Politécnico Nacional. Centro de Investigación en Computación, México D.F., 2009. URL <http://tesis.ipn.mx/bitstream/handle/123456789/8626/100.pdf?sequence=1>.
- W. Renteria-Agualimpia and S. Levashkin. Multi-criteria geographic information retrieval model based on geospatial semantic integration. In *GeoSpatial Semantics*, pages 166–181. Springer, 2011.
- W. Renteria-Agualimpia, F. J. López-Pellicer, P. R. Muro-Medrano, J. Nogueras-Iso, and F. J. Zarazaga-Soria. Exploring the advances in semantic search engines. In *Distributed Computing and Artificial Intelligence*, pages 613–620. Springer, 2010.
- W. Renteria-Agualimpia, F. J. López-Pellicer, A. J. Florczyk, J. López de Larrinzar, J. Lacasta, P. R. Muro-Medrano, and F. J. Zarazaga-Soria. Detectando anomalías en los metadatos de cartotecas. *Scire: representación y organización del conocimiento*, 19(1):23–29, 2013a.

- W. Renteria-Agualimpia, F. J. López-Pellicer, J. Lacasta, P. R. Muro-Medrano, and F. J. Zarazaga-Soria. Aproximación geosemántica para detectar inconsistencias en los metadatos de servicios Web geoespaciales. *GeoFocus: International Review of Geographical Information Science and Technology*, 13(1):154–176, 2013b.
- W. Renteria-Agualimpia, F. J. López-Pellicer, J. Lacasta, F. J. Zarazaga-Soria, and P. R. Muro-Medrano. Identifying hidden geospatial resources in catalogues. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, page 32. ACM, 2013c.
- W. Renteria-Agualimpia, F. J. López-Pellicer, J. Lacasta, P. R. Muro-Medrano, and F. J. Zarazaga-Soria. Identifying geospatial inconsistency of Web services metadata using spatial ranking. *Earth Science Informatics*, pages 1–11, 2014. ISSN 1865-0473. doi: 10.1007/s12145-014-0172-4.
- J. Renz. *Qualitative spatial reasoning with topological information*. Springer-Verlag, Heidelberg, 2002.
- R. J. Robertson. Metadata quality: implications for library and information science professionals. *Library Review*, 54(5):295–300, 2005.
- R. T. Rockafellar, R. J.-B. Wets, and M. Wets. *Variational analysis*, volume 317. Springer, 1998.
- A. Rodríguez. Inconsistency issues in spatial databases. In *Inconsistency tolerance*, pages 237–269. Springer, 2005.
- G. Salton and M. J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, New York, 1983.
- L. Samulenk and V. L. Rubin. Geographically aware information access with geoparsing, geocoding, and georeferencing. In *Proceedings of the 40th Annual Conference of the Canadian Association for Information Science*, School of Library and Information Studies, Wilfrid Laurier University and the University of Waterloo, Waterloo, Canada., May 31 - June 2 2012. ACM.
- J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data mining and knowledge discovery*, 2(2):169–194, 1998.
- M. Sanderson, S. Ramage, and L. Van Linden. Sdi communities: Data quality and knowledge sharing. In *proceeding of the 11th GSDI conference, Rotterdam* <http://www.gsdiconf/gsdiconf/gsdiconf11/papers/pdf/283.pdf>, 2009.

- U. Schindler and M. Diepenbroek. Generic xml-based framework for metadata portals. *Computers & Geosciences*, 34(12):1947–1955, 2008.
- S. Servigne, T. Ubeda, A. Puricelli, and R. Laurini. A methodology for spatial consistency improvement of geographic databases. *GeoInformatica*, 4(1):7–34, 2000.
- R. Shen, M. A. Gonçalves, and E. A. Fox. Key issues regarding digital libraries: Evaluation and integration. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 5(2), 2013.
- S. L. Shreeves, E. M. Knutson, B. Stvilia, C. L. Palmer, M. B. Twidale, and T. W. Cole. Is “quality” metadata “shareable” metadata? the implications of local metadata practices for federated collections. In *Twelfth National Conference of the Association of College and Research Libraries, April 7-10 2005, Minneapolis, MN. Chicago, IL: Association of College and Research Libraries*, pages 223–237. Association of College and Research Libraries, 2005. URL <http://hdl.handle.net/2142/145>. Last visited on 12.11.2014.
- M. J. Silva, B. Martins, M. Chaves, A. P. Afonso, and N. Cardoso. Adding geographic scopes to Web resources. *Computers, Environment and Urban Systems*, 30(4):378–399, 2006.
- A. K. Sinha. *Societal challenges and geoinformatics*, volume 482. Geological Society of America, 2011.
- T. F. Søndergaard, J. Andersen, and B. Hjørland. Documents and the communication of scientific and scholarly information: revising and updating the UNISIST model. *Journal of documentation*, 59(3):278–320, 2003.
- H. Southall and P. Pridal. Old maps online: Enabling global access to historical mapping. *e-Perimetron*, 7(2):73–81, 2012.
- B. Stvilia, L. Gasser, M. B. Twidale, S. L. Shreeves, and T. W. Cole. Metadata quality for federated collections. In *Proceedings of the International Conference on Information Quality-ICIQ 2004*, pages 111–125, Cambridge, MA: MITIQ, November 2004.
- B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smith. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12):1720–1733, 2007.
- R. S. Taylor. *Value-added processes in information systems*. Greenwood Publishing Group, 1986.

- W. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46:234–240, 1970.
- R. Tolosana-Calasanz, J. A. Álvarez-Robles, J. Lacasta, J. Nogueras-Iso, P. R. Muro-Medrano, and F. J. Zarazaga-Soria. On the problem of identifying the quality of geographic metadata. In *Research and Advanced Technology for Digital Libraries*, pages 232–243. Springer, 2006.
- M. Torres, S. Levachkine, M. Moreno, R. Quintero, and G. Guzmán. Retrieving geospatial information into a Web-mapping application using geospatial ontologies. In *Agent and Multi-Agent Systems: Technologies and Applications*, pages 267–277. Springer, 2007.
- V. Toy-Smith. Ualc best practices metadata guidelines: A consortial approach. *Journal of Library Metadata*, 10(1):1–12, 2010.
- J. Triglav, D. Petrovič, and B. Stopar. Spatio-temporal evaluation matrices for geospatial data. *International Journal of Applied Earth Observation and Geoinformation*, 13(1):100–109, 2011.
- O. Valkeapää, O. Alm, and E. Hyvönen. Efficient content creation on the semantic Web using metadata schemas with domain ontology services (system description). In *The Semantic Web: Research and Applications*, pages 819–828. Springer, 2007.
- H. Veregin. Data quality parameters. *Geographical information systems*, 1:177–189, 1999.
- D. R. Walker, I. A. Newman, D. J. Medyckyj-Scott, and C. L. Ruggles. A system for identifying datasets for GIS users. *International Journal of Geographical Information Science*, 6(6):511–527, 1992.
- F. Wang. *Handling Data Consistency through Spatial Data Integrity Rules in Constraint Decision Tables*. PhD thesis, Bundeswehr University Munich, 2008.
- J. Wang, X. Wang, and S. Liang. Geocustering: A Web service for geospatial clustering. In *Advances in Web-based GIS, Mapping Services and Applications*, pages 37–54, 2011.
- L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li. Detecting dominant locations from search queries. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 424–431. ACM, 2005.

- R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, pages 5–33, 1996.
- S. Wang, C.-S. Chen, V. Rinsurongkawong, F. Akdag, and C. F. Eick. A polygon-based methodology for mining related spatial datasets. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics*, pages 1–8. ACM, 2010.
- A. Whiteside and J. Greenwood. OGCWeb Services Common Standard. No. OGC 06-121r9 v 2.0., Open Geospatial Consortium, 2010.
- Z. Xie, G. Tian, L. Wu, and L. Xia. A framework for correcting geographical boundary inconsistency. In *Geoinformatics, 2010 18th International Conference on*, pages 1–5. IEEE, 2010.
- O. L. Zavalina. Exploring the richness of collection-level subject metadata in three large-scale digital libraries. *International Journal of Metadata, Semantics and Ontologies*, 7(3):209–221, 2012.
- M. L. Zeng and J. Qin. *Metadata*. Facet publishing. London, UK., 2008.
- M. L. Zeng, B. Subrahmanyam, and G. M. Shreve. Metadata quality study for the national science digital library (NSDL) metadata repository. In *Digital libraries: International collaboration and cross-fertilization*, pages 339–340. Springer, 2005.
- W. Zong, D. Wu, A. Sun, E.-P. Lim, and D. H.-L. Goh. On assigning place names to geography related Web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 354–362. ACM, 2005.