JIS

Article

Improving the geospatial consistency of digital libraries metadata

Journal of Information Science
1–18

DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

Walter Renteria-Agualimpia
Department of Computer Science and Systems Engineering, University of Zaragoza, Spain

Francisco J. López-Pellicer
Department of Computer Science and Systems Engineering, University of Zaragoza, Spain

Javier Lacasta
Department of Computer Science and Systems Engineering, University of Zaragoza, Spain

F. Javier Zarazaga-Soria
Department of Computer Science and Systems Engineering, University of Zaragoza, Spain

Pedro M. Muro-Medrano
Department of Computer Science and Systems Engineering, University of Zaragoza, Spain

Abstract
Consistency is an essential aspect of the quality of metadata. Inconsistent metadata records are harmful: given a themed query, the set of retrieved metadata records would contain descriptions of unrelated or irrelevant resources, and would even do not contain some resources considered obvious. This is even worse in when the description of the location is inconsistent. Inconsistent spatial descriptions may yield invisible or hidden geographical resources that cannot be retrieved by means of spatially themed queries. Therefore, ensuring spatial consistency should be a primary goal when reusing, sharing and developing georeferenced digital collections. We present a methodology able to detect geospatial inconsistencies in metadata collections based on the combination of spatial ranking, reverse geocoding, geographic knowledge organization systems, and information retrieval techniques. This methodology has been applied to a collection of metadata records describing maps and atlases belonging to the Library of Congress. The proposed approach was able to identify automatically inconsistent metadata records (870 out of 10,575) and propose fixes to most of them (91.5%) These results support that the proposed methodology could assess the impact of spatial inconsistency in the retrievability and visibility of metadata records and improve their spatial consistency.

Keywords
Metadata Quality; Digital Library; Consistency; Geospatial Clustering; Spatial Ranking.

1. Introduction

Geospatial information, i.e., information that references to a place, is a core component of Digital Libraries (DL). It helps to reveal unknown spatial patterns, increases the recall of information retrieval systems, and enhances real world experiences of users, since most events can be visualized, explained, and understood in geographic terms [1]. Libraries have traditionally included geospatial information, and Geographic Information Retrieval (GIR) systems have been developed to perform spatial queries on metadata [2, 3]. The University of California library, the Library of Congress (LoC) and the National Archives of the United Kingdom are good examples of the relevance of geospatial Information in Digital Libraries and National Archives. For example, Petras [4] analysed around 5 million records from the University of California library catalogue and found that approximately 35% of the records contain data in MARC21

Corresponding author:
Walter Renteria-Agualimpia, Computing and Systems Engineering Dept., University of Zaragoza, C/ María de Luna 1 CP 50018 Zaragoza (Spain)
walrterra@unizar.es

fields related to geospatial information. The Geography and Map Division[1] 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 [5].

The analysis, interpretation, efficient accessibility and re-use of the information depend on its meta-information, i.e. the metadata. Metadata provide basic information for archiving, discovering, describing data and data integration. As is mentioned by Schindler and Diepenbroek [6], geospatial metadata at a minimum have to answer the questions: Who has measured, observed, or calculated what, where, when, and how? A number of metadata standards define the corresponding content structures for collecting metadata. The most important ones in the geographic context are ISO 19115 [7], FGDC [8], DIF [9], and Dublin Core [10]. These content standards allow users to identify data sets not only by bibliographic information, such as authors, title, date, publisher, etc.; they also make available spatial coverage, parameters used, and data quality.

The creation and maintenance of a digital library requires a significant effort [11] that can be easily wasted if the used metadata are inconsistent. As [12] indicates, metadata consistency should be the primary consideration in the development of digital collections. In the context of geographic meta-information, the consistency is even more essential [13, 14]. It is more useful when an increasing amount of the content is also available digitally because it provides geo-based ways for browsing and searching data [15]. Hill [16] points out how inconsistent spatial description of geographic resources can easily generate discrepant results, inadequate weights for ranking search results, 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 queries with both constraints. 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 within Germany, such resource will not be retrieved through keyword queries. Likewise, if its 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 shows real examples of spatial inconsistencies found in metadata records published by the LoC related to Germany. The bounding boxes were extracted from metadata records using the procedure described in Section 3.2. The problems are not restricted to a date or to a topic. From west to east, we can highlight in Figure 1 a metadata record about a map of the administrative and political divisions of Prussia in 1807 that locates Germany in the middle of the Atlantic Ocean (A), three metadata records about the Federal Republic of Germany (military maps, travel maps, regional atlas) dated between 1978-1982 that locate Germany over Ireland (B), and two metadata records that locate the Confederation of the Rhine (1810) and the northwest of Germany (1816) over Ukraine (C). The URL of each metadata record can be found in the Table 1***.

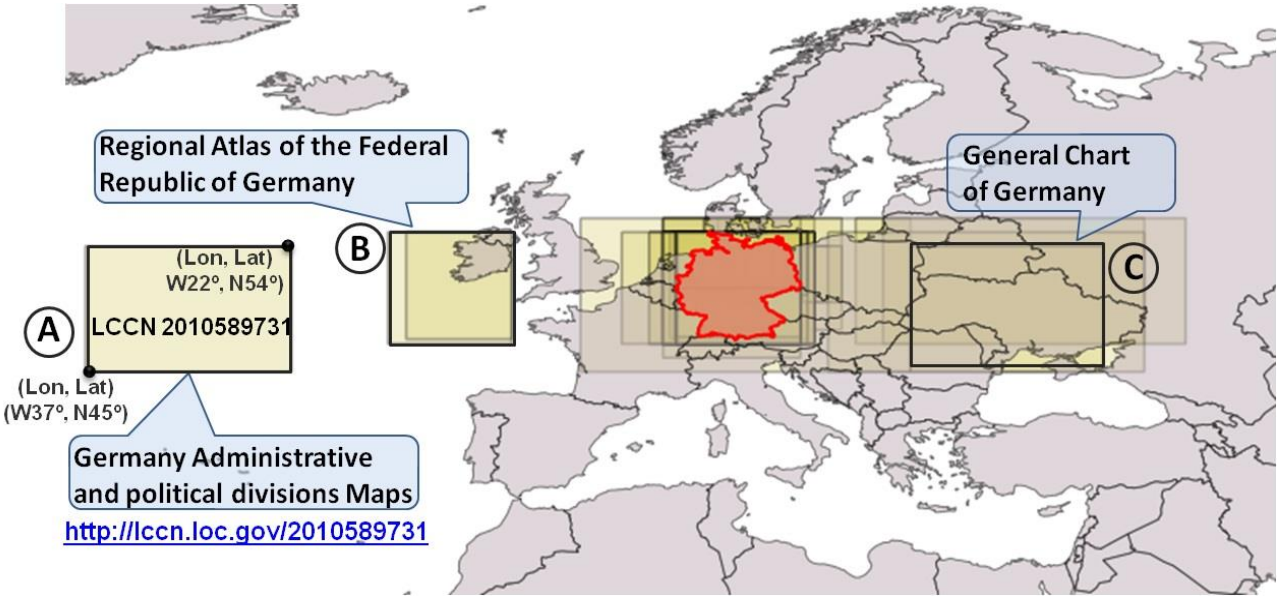

Figure 1. Examples of geospatial inconsistencies on LoC.

Table 1. Details of some geospatial inconsistencies in the LOC about Germany.

LCCN	Place Name	Description	Case	URL
2010589731	Germany	Interims chart of Deutschland	A	http://lccn.loc.gov/2010589731
325901201x	Germany	Deutsche Travel Map 1:250 000	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 Federal Republic Deutschland	B	http://lccn.loc.gov/82215282
2011585205	Germany	General Chart of Deutschland	C	http://lccn.loc.gov/2011585205
2011585214	Germany	General chart of Northwest Deutschland	C	http://lccn.loc.gov/2011585214

LCCN is the Library of Congress Control Number.

The problem of metadata consistency has drawn research interest [17-21]. In order to detect spatial inconsistencies, [22] proposed the hypothesis that geospatial clusters could reveal an implicit consensus among documentation experts for identifying some geographic areas. This hypothesis uses geospatial clustering and Knowledge Organization Systems (KOS) to compare Indirect Spatial References (ISR) from unstructured content with Direct Spatial References (DSR) from structured content of metadata. The 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.

This work presents and formalizes a semi-automatic methodology for digital curation processes, particularly, processes involved in preservation tasks of digital repositories of cartographic materials. To do so, this work extends the methodology proposed in [22] by adding a double validation process that improves inconsistency detection results. Additionally, the methodology has been updated to be able to process online spatial metadata collections.

This work presents the application of the extended methodology in a real use case that analyses more than 42,000 MARC21 metadata records describing spatial resources retrieved from the LoC. Results provide a quantitative view of problems related to resource discovery, invisibility and retrieval of such metadata records, and therefore interoperability consequences that may affect digital library distributed systems.

This paper is organized as follows. Section 2 discusses related work on exploiting location in the context of digital libraries. Section 3 introduces the methodology of inconsistencies detection. Section 4 describes an experimental and quantitative study and presents the results. Section 5 discusses the main results of the study. Finally, section 6 concludes the paper and outlines our ideas for future work.

2. Related work

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, an example these works citing previous geospatial Digital Library projects is shown in the Figure 2. The most relevant digital library projects experimenting with georeferenced data and metadata are focused on three main areas: information visualization, geographic information retrieval and information validation.

Some works focusing on information visualization are the Geo-Referenced Information Network, the Electronic Cultural Atlas Initiative (ECAI) [23], the Old Maps Online [24] and the Alexandria Digital Library (ADL) [25], probably one of the most widely cited research projects that made use of georeferencing in the context of digital libraries. They are focused on representing, exploring and browsing digital collections on a map, and some of them offer additional search functionalities. The area of geographic information retrieval deals with the disambiguation of place names based on internal and external evidence from the text content of metadata. Internal evidence includes the use of generic geographic labels, or linguistic environment. External evidence includes knowledge organization systems, gazetteers, biographical information, and general linguistic knowledge [25-27]. Some works in this area are the Spatially-Aware Information Retrieval on the Internet (SPIRIT) [27], the Geographic Awareness Tool (GAT) [15], the MapRank [29] and the Old Maps Online.

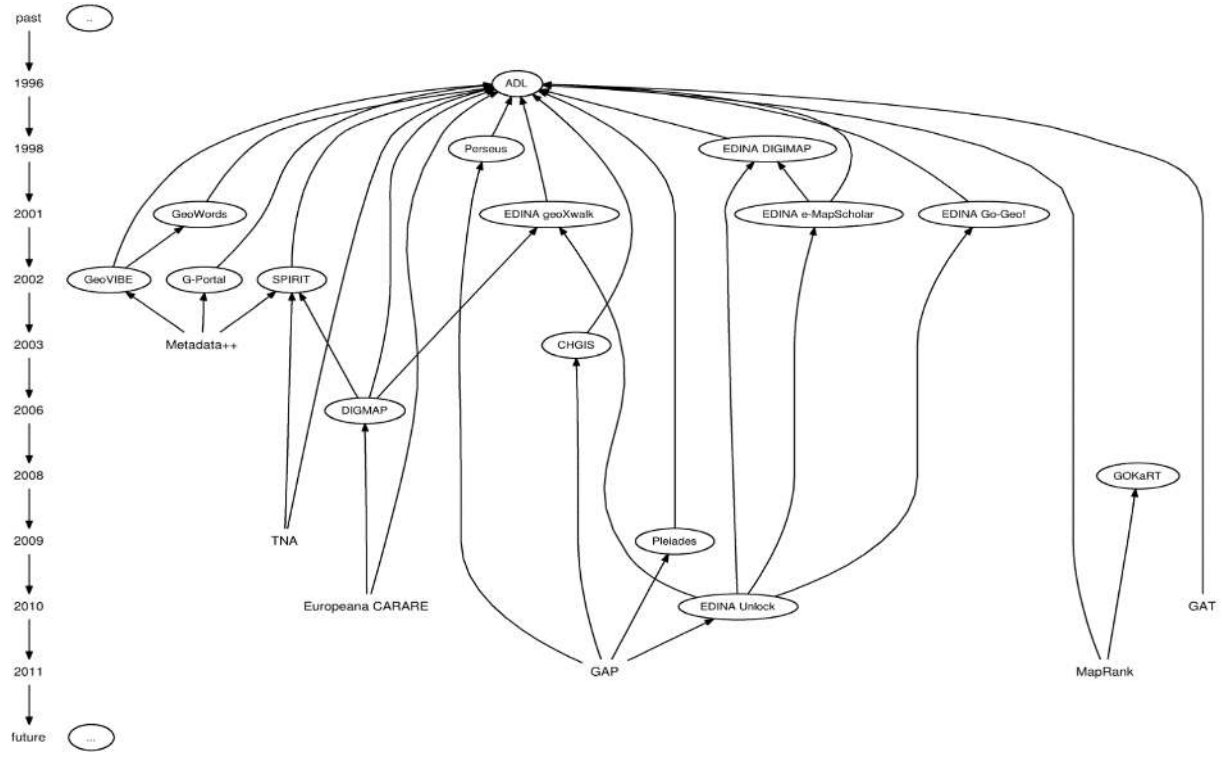

Figure 2. Some of the most popular Digital Libraries projects working with georeferenced data and metadata.

With respect to information validation, these kinds of works are focused on data and metadata quality; we centre our attention on this last kind of works. Metadata quality is a semantically slippery term. Park [30] suggested 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 [31-36]. 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 geospatial information included in the metadata and the georeferenced 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.

There are some works with a quantitative approach. Tolosana-Calasanz et al. [37] developed a quantitative method and realised a statistical analysis for assessing the quality of geographic metadata. The authors first formulated a list of geographic quality criteria by consulting domain experts. The identified criteria indicated quality preferences. The authors also noticed the need to ensure the completeness of the spatial fields in order to guarantee a minimum level of quality. Ma et al. [38] presented a study about the quality assessment of metadata on the Internet Public Library. This work is based on a combination of human evaluation (qualitative) and automatic evaluation (quantitative). The qualitative method gave an indication of the quality of information by rating accuracy, completeness, consistency and functionality. These works are different from the approach presented here because their quantitative methods only measure the completeness of metadata in the collection, however our approach is focused on evaluating the spatial consistency quantitatively, that is, we use spatial best matches for finding and measuring inconsistencies.

Regarding to the clustering focus of our approach, one of the most relevant related works is Hays and Efros [39]. Their work also uses the idea of implicit consensus of spatial co-occurring resources for estimate the location of an image. One of the main differences with our work lies in the final use of the consensus: Hays and Efros use the consensus for estimating a geographic location, the work presented here uses the consensus for detecting geographic inconsistencies. Works such as [29,30,40] estimate the geographic location of resources (text, images, etc.). They use classifiers and knowledge from social datasets to disambiguate references to locations using generally textual context, but their approach does not take into account information from a wider spatial context, they do not take into account any information from co-occurring metadata such as spatial descriptive consensus provided by neighbours (e.g. cluster). Here we show the utility of incorporating spatial knowledge of co-occurring metadata description in inconsistency detection systems. Many of the cited works develop approaches of geolocation based mainly on text. They also suggest that an interesting extension of their works is to rely upon the natural clustering of related documents. This is the focus of the research presented here. We take advantage of spatial co-occurrences found in metadata.

3. Methodology.

This section presents our extension of [23] proposal to detect geospatial inconsistencies in DL metadata. A general outline of the process is shown in Figure 3. Our methodology uses the principles proposed in [42]. Its main insight is the use of KOS combined with geospatial ranking functions for finding the most relevant toponyms associated with a footprint and then compare these with place names described in the metadata. We integrate this idea with the concepts of two-dimensional spatial clustering to refine the detection of spatial inconsistencies in other fields such as DL. The resulting extended methodology has six main steps: Harvesting, Geo-Extraction, Reverse Geocoding, Spatial Clustering, Metadata Validation, and Report Generation. These steps are described in detail in the following subsections.

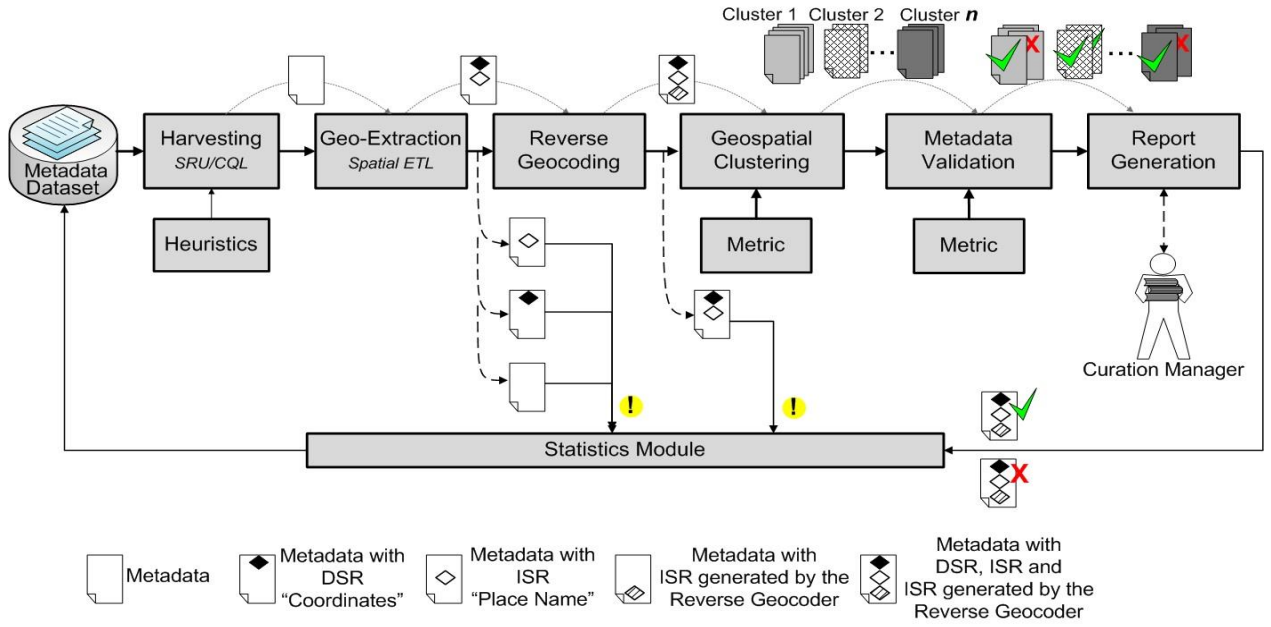

Figure 3. Methodology to detect geospatial inconsistencies in metadata.

3.1. Harvesting

The first step is the harvesting of metadata records that may contain geospatial information in form of geographic coordinates. This step is optional if we have access to some kind of metadata dump. The process of metadata harvesting consists in collecting metadata descriptions stored in digital repositories using protocols such as OAI-PMH [42] or Search/Retrieve via URL (SRU)[2] In our case we are using the SRU protocol. The SRU protocol uses three types of operations: *explain*, *scan* and *searchRetrieve*. Our methodology uses the last one. The *searchRetrieve* operation allows submitting a query using the high-level Contextual Query Language (CQL) and retrieving the list of items that match the query [43]. 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 the resource, DSR specifically. This heuristic is based on the recommended procedures of the Map Cataloging Manual[3] of the LoC and allows us to harvest those metadata records that have been created according to the Map Cataloging Manual. For example, the query "W12*" in CQL can match a coordinate referencing a point whose longitude is between W120 to W129. If we want to retrieve metadata that may contain such information, we should formulate the following query:

*http://z3950.loc.gov:7090/voyager?version=1.1&operation=searchRetrieve&maximumRecords=10&startRecord=1&recordSchema=mods&recordPacking=xml&query=W12**

The response always report us the total number of records that matches such query. The *maximumRecords* parameter sets an upper limit to the number of records returned. Some systems may ignore the value of such parameter if it is over an internal constant. Thus, in order to retrieve all matching records, we repeat the query modifying the value of the *startRecord*, which provides a means to page through large numbers of results records, until retrieving all the matching records. Using this heuristic, the harvesting module retrieved all MARCXML[4] and MODS[5] metadata records matching with this type of query (see examples in Figure 4) in a range from 180º east to 180º West and from 90º North to 90º South by generating the appropriate query patterns. Later, the system verified if retrieved metadata records effectively contain geographic coordinates. The output of this process is the set of metadata records that contain an explicit DSR following well-known cataloguing rules that are retrievable through the SRU endpoint.

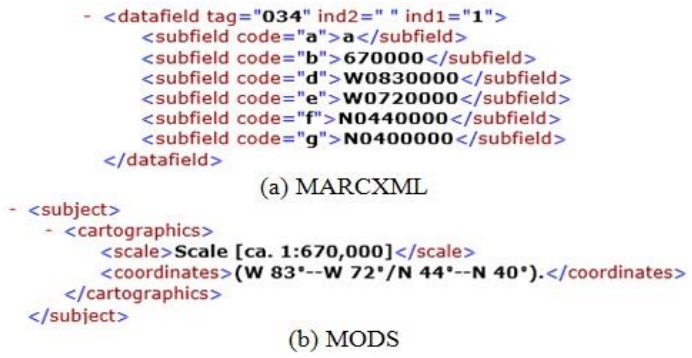

Figure 4. Example of coordinates in MARCXML and MODS formats.

3.2. Geo-Extraction

The geo-extraction step applies to harvested metadata records, or, if already available, a metadata dump. This step is a geospatial Extraction, Transformation and Load process (ETL) [44]. This module extracts and homogenizes Direct Spatial References (DSR) encoded in MARC21 metadata records. This module also extracts the Indirect Spatial References (ISR) from textual place name fields. In MARC21 metadata, a DSR has the form of a bounding box and can be found in the field "034 - Coded Cartographic Mathematical Data". A bounding box is a pair of latitude/longitude pairs that defines the northern, southern, east and west extremes of a geographic region. ISR 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 title field. The output of this process is a stream of metadata records annotated with the extracted DSR (explicit bounding box) and ISR (place name). Metadata without DSRs, ISRs or both are accounted as incomplete metadata in the statistics module. Incomplete metadata records are not taken into account for further processing because the purpose of this workflow is the identification of inconsistencies between DSR and ISR in metadata records.

A particular observation in the context of this step is that, 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. The geo-extraction step focuses on the bounding box field, it was the most frequent DSR field in the dataset analysed. One potential drawback to this approach is that erroneous interpretations for the coordinates given in the bounding box associated to a particular resource may be due to problems in accounting with geospatial referencing systems. To deal these issues, we examined manually the detected inconsistencies.

3.3. Reverse geocoding

This step is a conversion process from a reference systems based on coordinates (i.e., a bounding box) into a reference systems based on geographic identifiers. The goal is to find the best ISR (place name) for the geographic region covered by the DSR (explicit bounding boxes). For this task, we use the reverse geocoder described in [22]. This reverse geocoder uses the Hausdorff distance [45] to measure the geospatial similarity between the geometrical shape of a DSR and the geographic extent of entities belonging to a geospatial KOS. This 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 particular we used the geodetic distance as the internal metric. The mathematical expression of the Hausdorff distance is shown in Eq. (1):

$$\mathrm{dis}_H(\mathrm{X},\mathrm{Y}) = \left\{ \underset{x \in X}{Sup} \ \underset{y \in Y}{Inf}\, d(x,y), \underset{y \in Y}{Sup} \ \underset{x \in X}{Inf}\, d(x,y) \right\} \quad (1)$$

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.

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 [46]. This module annotates each processed metadata record with the list of entities that best describe its Direct Spatial Reference. The geographical KOS used consists of several public models, databases and KOS. Its main sources are available online: GADM[6] (its current version delimits 556,049 administrative areas (or 218,238 if you count only the lowest level for each country), U.S. Census Bureau[7], Natural Earth Data[8], and the National Oceanic and Atmospheric Administration's[9].

3.4. Geospatial clustering

We define *geospatial metadata cluster* as a group of metadata records whose spatial references co-occur in the same area and have similar geographical extent. This step assumes that a cluster of such characteristics may reveal an implicit consensus among library experts about the spatial references that are more likely to be used to describe textually a geographic location in the area covered by such cluster. This idea will serve to validate spatial descriptions in the metadata record and detect potential inconsistencies. This step uses the density-based DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) [47] for computing spatial clusters using as input values DSR (explicit bounding boxes) 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 an efficient algorithm for big collection of data [48]. As Wang et al. [49] summarizes 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:

$$N_{Eps}(\mathrm{p}) = \{q \in D | dis(p,q) \leq Eps\} \quad (2)$$

If $\left\|N_{Eps}(p)\right\| \geq MinPts$, 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 labelled 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. We are using an adaptation of the DBSCAN algorithm that uses the Hausdorff distance as distance measurement instead of the Euclidean distance [50, 51] to works with bounding boxes. Many library metadata records contain a geographical extent which is a two-dimensional footprint. The use of the Hausdorff distance instead of the Euclidean distance allows computing clusters from two-dimensional data directly (bounding boxes, multi-Polygons or complex geometries). In our approach, we normalize 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 DSR (explicit bounding boxes). The similarity threshold value is 0.5. The normalization function is similar to the function described in [52].

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*. As a basic consideration, a cluster is group of at least two elements, for this reason, the MinPts parameter is set 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 the values *Eps*=0.2 and *MinPts*=2.0. 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), but they are georeferencing spatial entities of different levels/size (example, a city, a province and a state centred in the same point but with different extents coverage). The recommended technique for the parameter selection is described in [47], 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. 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. The authors also ensure that choosing 4 as the default value of *MinPts* produces the best results in two-dimensional clustering. In our experiments lower values, usually 2, obtained better results for the spatial ranking.

3.5. Metadata validation

This step computes first for each cluster two sets of ISR (places names). The first set is the union of the place names generated by the reverse geocoding module for each metadata record belonging to the cluster. The second set is the union of the explicit place names in the metadata description belonging to a cluster. Next, this step performs in each metadata record belonging to a cluster a dual validation process. This validation process verifies if exist geospatial inconsistencies between the original ISR and the ISR generated by an external process (e.g. reverse geocoding or clustering); the first validation process validates the ISR with respect to the geographical KOS, 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 (VSM) [53]. They measure the similarity between the spatial description of a metadata record and two vectors of place names associated with the cluster given a metadata record 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 metadata record. 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 metadata record. In both cases, a metadata record 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 DSR analysed. Although we use VSM for calculating the similarity, it is possible to use other metrics for measuring the similarity between them [54]. In our work, the similarity between two vectors is assessed by the next expression:

$$cos(x) = \frac{g_i \cdot t_i}{\|g_i\| \cdot \|t_i\|} \tag{3}$$

Where t_i is the vector of original place names of a metadata record 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).

$$g_i = \{g_{1,i}, g_{2,i}, \dots, g_{n,i}\} \tag{4}$$

$$t_i = \{t_{1,i}, t_{2,i}, \dots, t_{m,i}\} \tag{5}$$

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

$$\{"g_1": 1, "g_2": 2, \dots "g_n": n, "t_1": n+1, "t_2": n+2, \dots "g_m": n+m\} \tag{6}$$

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:

$$g_i' = \{1,2,0,4, \dots, k-3, 0, k-1, k\} \tag{7}$$

$$t_i' = \{0,2,3,0, \dots, k-3, k-2, 0, k\} \tag{8}$$

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). A high value of consistency for a metadata indicates that the metadata 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 metadata records inconsistent with co-occurring

metadata records in the cluster. In this step we analysed different string comparison methods, and we selected the simplest method: a simple matching string to compare the searched place name with two fields, the official place name and an alternative name (when it exist) provided by the KOS.

3.6. Report generation

This last step reports the consistency of each metadata with respect to its own geospatial information and with respect to its neighbours. Metadata records identified as consistent could be annotated as having a high quality value, and linked to the place name from the geographical KOS used by the reverse geocoder. Metadata records identified as inconsistent could be annotated with an alert value to advertise the need to review them. This information can be useful for curation managers [55]. All information reported is included in a general report produced by the Statistics Module. This module counts the number of uncompleted metadata and reports the kind of inconsistency found in individual and collective validation. This report is used for the analysis of results. An example is shown in Table 2.

Table 2. Inconsistent metadata record report.

Acronym	Symbol	Field	Value
ISR	◇	Place Name	Ohio
		Title	Ohio, major land resource areas
DSR	◆	MBB	(W 85°, N 42°; W 80°, N 38°)
		LCCN	92681234
		Type	Map
		Consistency	0,005
ISR	◈	Suggested Place Name	North Dakota

4. 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 section 3.1. 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 DSR (bounding box) locates a state, a county, a city, a forest or a watershed. The distribution is shown in Figure 5.

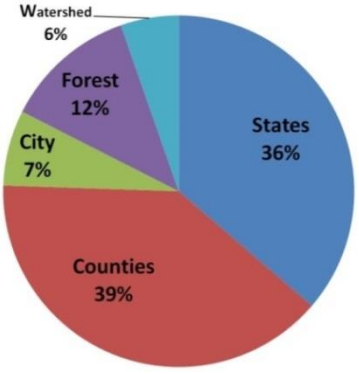

Figure 5. Distribution of the type of extent coverage of the metadata used in the experiment.

The validation processes of the methodology have helped to detect three kinds of inconsistencies: (1) *syntactic inconsistency*, (2) *geospatial semantic inconsistency or geosemantic inconsistency*, and (3) *contextual inconsistency*.

(1) *Syntactic 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[10] [56-58]. For example, the range of latitude and longitude coordinates need to be checked: the absolute value of latitude must be between 90º North and 90º South, and the absolute value of longitude must be between 180º East and 180º West. In some cases, a simple query such as: "*Are the values of latitude coordinate always between -90º and 90º?*" can reveal a logical geospatial inconsistency. Our methodology reveals distorted (extra-long) DSR (bounding boxes) shown in the Figures 6 and 7 without the need of checking the coordinate values. These kinds of errors can be easily solved at the source by a careful conversion of the MARC21 coordinates. In many cases the results shows that they were encoded in the description data out of range.

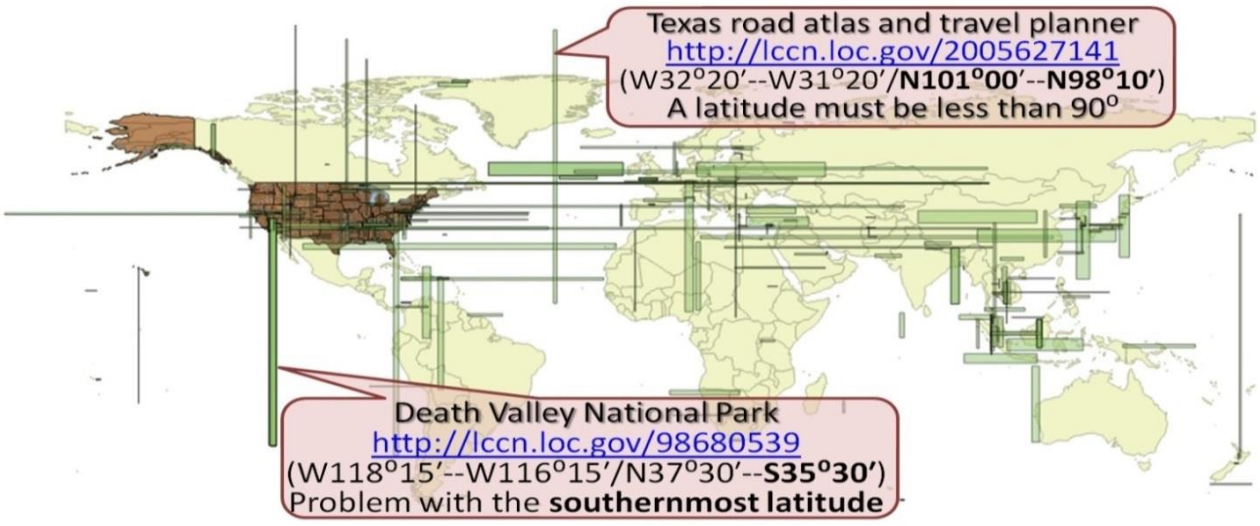

Figure 6. Global vision of the spatial logical inconsistencies in the LOC metadata records.

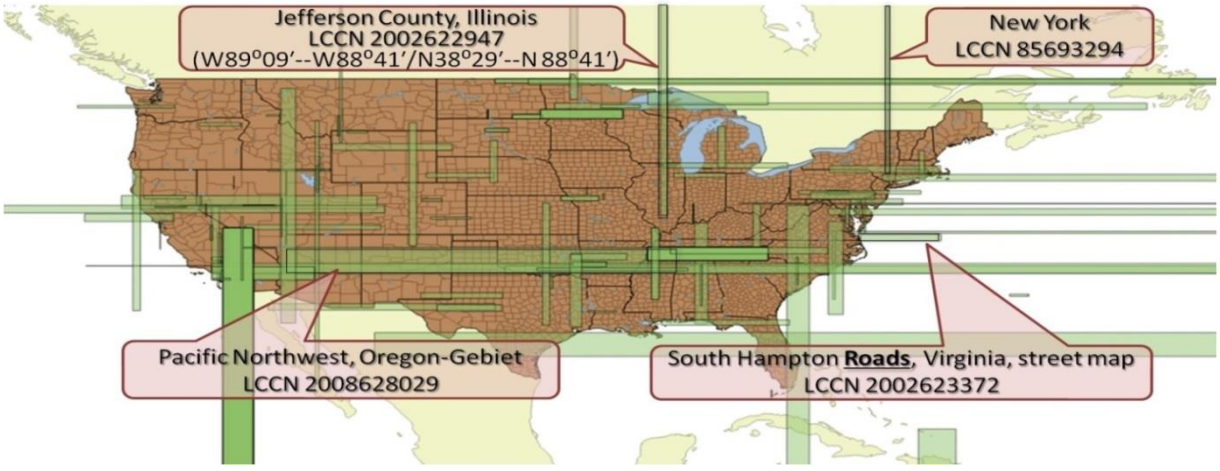

Figure 7. Spatial logical inconsistencies in the LOC metadata records focused on USA.

(2) *Geosemantic inconsistency*. This kind of inconsistency is originated in the conceptual incoherency between the DSR (e.g. bounding boxes) and the ISR (e.g. place names) according a specific KOS. There are three cases: micro-macro, macro-micro, and inconsistent. The micro-macro case happens when the ISR of the metadata record has a micro scope (e.g. forest, etc.) but its DSR has a macro scope (e.g. state). The macro-micro case is its inverse, where the ISR (e.g. county) has a macro scope but the DSR (e.g. city) covers a small area. Finally, inconsistent are the trivial cases or complete disagreement between DSR and ISR. Inconsistent cases can be found automatically by using the reverse geocoder with the help of the Hausdorff distance and a threshold of 0.5. Figures 8 (a) and (b) show examples of inconsistent cases.

(a) Case of Cuba (b) Case of Caledonia County

Figure 8. Example of place name and footprint with geospatial inconsistence.

(3) *Contextual inconsistency*. This kind of inconsistency is caused by a disagreement between DSR (e.g. bounding boxes) and ISR (e.g. place names) of similar metadata records that describe the same area. For example, Figures 9 and 10 show a disagreement between a metadata record and the consensus of their neighbourhood. In the first case, the methodology identifies a disagreement between the metadata describing (Ohio State http://lccn.loc.gov/92681234) and the consensus of their neighbourhood (North Dakota State). This case is an obvious example of spatial inconsistency that could be detected without clustering. However, in the second case, the methodology identifies two subtypes of contextual inconsistencies more complexes. Although they overlap, they are contextual inconsistencies (spatial mismatches specifically) because in their spatial context their place names are unusual. Experts usually catalogue the same spatial area with other place name. This kind of inconsistencies could be seen as a geospatial synecdoche (taking a part for the whole and vice versa). Figures 10 (a) and (b) shows these cases. 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 11. The contextual 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 inconsistency is applied on clusters and it could use KOS optionally.

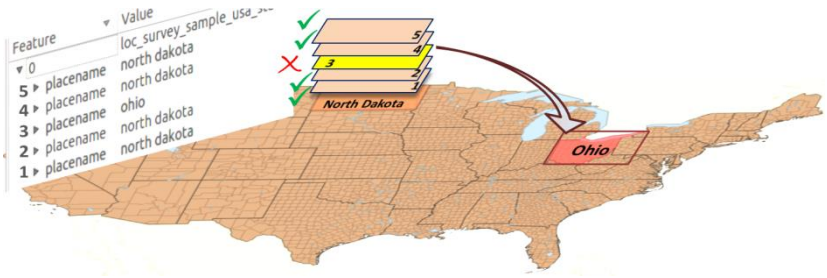

Figure 9. Results showing a disagreement between a metadata describing (Ohio State http://lccn.loc.gov/92681234) and the consensus of their neighbourhood (North Dakota).

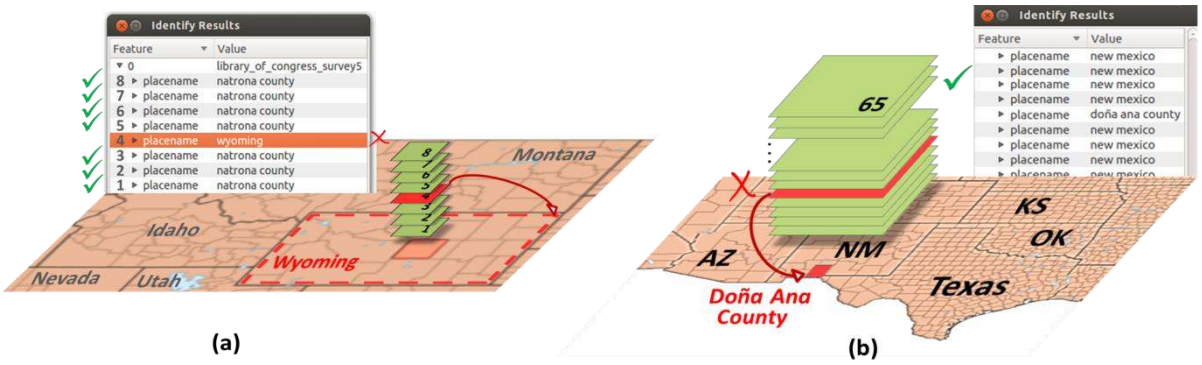

Figure 10. (a) Disagreement between a metadata (Wyoming http://lccn.loc.gov/2011593232) and the consensus of their neighbourhood (Natrona County). (b) Disagreement between (Doña Ana http://lccn.loc.gov/93682208) and New Mexico

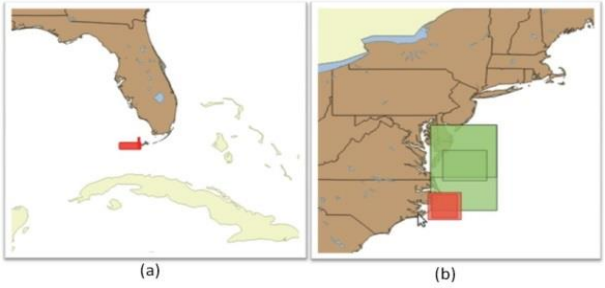

Figure 11. Example of lack of information about the area in the geographical KOS, (a) Lower Keys fishing map in Florida and (b) Hatteras offshore fishing chart in North Carolina.

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 12. In this case, 8 out of 14 elements in the cluster are inconsistent, thus the cluster is inconsistent. Table 3 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.

Table 3. Example of contextual inconsistency caused by systematic error probably.

Current location of the DSR	Real location according to the ISR	URL
Lake County, Cook County, Minnesota State.	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

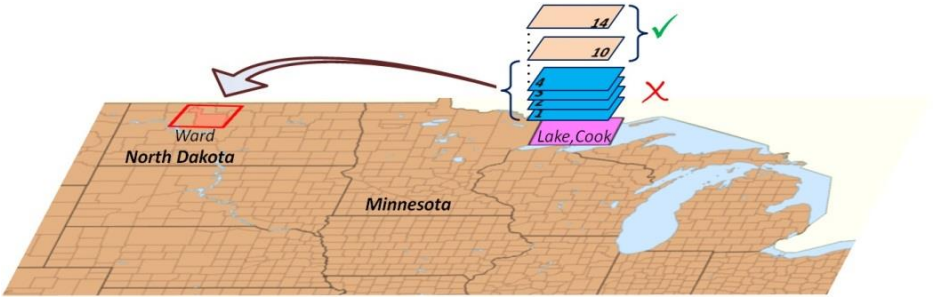

Figure 12. Results showing an inconsistency due to the disagreement among geospatial co-occurring Direct Spatial References belonging to a cluster.

The results are summarized in Table 4. We have found geospatial inconsistencies in 870 out of 10,575 metadata records. Our methodology identified 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 neighbourhood. The administrative types (states and counties) present fewer inconsistencies than types with 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, it is 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 4. Example of contextual inconsistency caused by systematic error probably.

Type	N°	Geospatial inconsistency			Total
		Logical	Semantic	Contextual	
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 (21.3%)	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. Discussion

There are four issues that deserve to be discussed with respect to the methodology and its results: heuristics for validating spatial descriptions, outlier detection and inconsistencies, the dimension in the spatial representation, and metadata reuse.

The methodology proposed uses a heuristic for validating spatial descriptions based on comparing sets of place names. Alternatively, a heuristic 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. In addition, the selection of an appropriate geographic KOS is crucial for a good reverse geocoding no wonder the heuristic applied. We need to take into account requirements such as having descriptions with different levels of details (geographical extents of different sizes) and topic variety.

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 to the dimension in the spatial representation, we have identified many cases where the one-dimensional representation generates problems. All these problems are due to bounding boxes that cannot be considered reasonably as similar, 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 centred 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 provide separability for co-occurring spatial objects in one dimension, but georeferencing spatial entities of different levels (example, a city, a province and a state centred in the same point but with different extents coverage). Figure 13 illustrates this situation; (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.

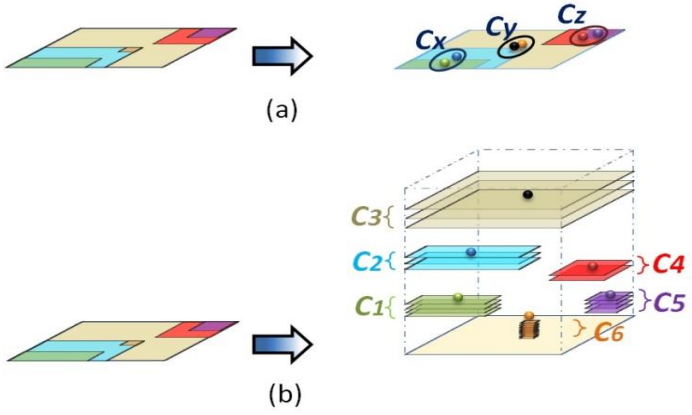

Figure 13. Differences between (a) one-dimensional and (b) two-dimensional clustering.

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” [59]. 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

geospatial consistency validation in order to ensure the data is retrieved and the quality of the entire library processes. We believe that before applying interoperability and sharing processes in digital libraries, we need to revise the geospatial consistency between the fields of spatial references (DSR and ISR) used in tasks such as retrieval, exploration and visualization of geospatial 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.

6. Conclusions and future work

This paper has presented an extension of a methodology based on two-dimensional geospatial clustering, geographic knowledge organization systems, spatial ranking and information retrieval techniques that checks the geospatial consistency of a metadata collection. 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 advantage in terms of inaccuracy detection, but also a gain of the use of 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.

Based on the results, even if we accept the importance of metadata, we need to worry about its consistency. For instance, metadata sharing and reuse in digital libraries should include richer geospatial consistency validation: to ensure the data retrievability, to improve the quality of the entire library processes, and also, to improve the user experience. According to the results, we think that it is necessary to automatically review the geospatial consistency of metadata records. The omission of this review can lead to problems of retrievability and invisibility of the referenced resources. Moreover, this review should be extended for detecting other kinds of geospatial inconsistencies, such as topological, temporal and thematic focused on the context of two-dimensional footprints. We plan to develop an automatic process for fixing geospatial inconsistencies to assess digital libraries in cataloguing tasks. This is the focus of our future research along with exploring the impact of geospatial inconsistencies in other domains.

Notes

1. http://www.loc.gov/rr/geogmap/
2. http://www.loc.gov/standards/sru/sru-1-1.html
3. http://www.itsmarc.com/crs/crs.htm#mergedProjects/mapcat/mapcat/
4. http://www.loc.gov/standards/marcxml
5. http://www.loc.gov/standards/mods
6. http://www.gadm.org
7. http://www.census.gov
8. http://www.naturalearthdata.com/downloads
9. http://www.nws.noaa.gov/geodata/catalog/national
10. http://www.iso.org/iso/catalogue_detail.htm?csnumber=26018

Acknowledgements

This work has been partially supported by the Spanish Government (project TIN2012-37826-C02-01), the Government of Aragon (project INNOVA-A1-038-13), the National Geographic Institute (IGN) of Spain and GeoSpatiumLab S.L. The work of Walter Renteria-Agualimpia has been partially supported by a grant (ref. B181/11) from the Aragon Government.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

[1] Samulenok L. and Rubin V., Geographically Aware Information Access with Geoparsing, Geocoding, and Georeferencing, In In *Proceedings of the 40th Annual Conference of the Canadian Association for Information Science*, 2012.

[2] Buckland M, Chen A, Gey F, Larson R, Mostern R and Petras V. Geographic search: catalogs, gazetteers, and maps. *College and Research Libraries* 2007; 68(5): 376-387.

[3] Zong W, Wu D, Sun A, Lim E and Goh D. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, 2005, pp. 354-362.

[4] Petras V. Statistical analysis of geographic and language clues in the MARC record, Technical Report; 2004. Available from: http://metadata.sims.berkeley.edu/papers/Marcplaces.pdf
[5] Clough P, Tang J, Hall M and Warner A. Linking archival data to location: a case study at the UK National Archives. *Aslib Proceedings* 2011; 63(2/3): 127-147.
[6] Schindler U and Diepenbroek M. Generic XML-based framework for metadata portals, *Computers & Geosciences* 2008; 34(12): 1947-1955.
[7] ISO/TC 211: ISO 19115: Geographic information — Metadata; 2003.
[8] FGDC: Content standard for digital geospatial metadata. Federal Geographic Data Committee; 1998. Available from: http://www.fgdc.gov/metadata/csdgm/.
[9] Barton G. Directory Interchange Format: a metadata tool for the NOAA Earth System Data Directory, *The role of metadata in managing large environmental science datasets*. Pacific Northwest Laboratory, Richland, Washington, USA, 1995, pp. 19-23.
[10] Duval E, Hodgins W, Sutton S and Weibel S. Metadata principles and practicalities. *D-lib Magazine*, 2002: 8 (4). Available from: http://www.dlib.org/dlib/april02/weibel/04weibel.html
[11] Huang C, Chuang T, Deng D and Lee H. Building GML-native web-based geographic information systems. *Computers & Geosciences*, 2009; 35(9): 1802-1816.
[12] Toy-Smith V. UALC best practices metadata guidelines: a consortial approach, *Journal of Library Metadata* 2010; 10(1): 1-12.
[13] Lutz M. Ontology-based discovery and composition of geographic information services. 2005.
[14] Piasecki M, Bermudez L, Beran B, Islam S, Choi Y, Liang X and Jeong S. Hydrologic Metadata, in *Hydrologic Information System Status Report*, 2005.
[15] Powell J, Mane K, Collins L, Mark L, Martinez B and McMahon T. The Geographic Awareness Tool: techniques for geo-encoding digital library content, *Library Hi Tech News* 2010; 27(9/10): 5-9.
[16] Hill L. Georeferencing: The Geographic Associations of Information. The MIT Press, 2006.
[17] Servigne S, Ubeda T, Puricelli A and Laurini R. A methodology for spatial consistency improvement of geographic databases. *GeoInformatica* 2000; 4(1): 7-34.
[18] Rodríguez A. Inconsistency issues in spatial databases in Inconsistency tolerance. Springer, 2005.
[19] Devillers R and Jeansoulin R. Fundamentals of spatial data quality. ISTE London, 2006.
[20] Hillmann D. Metadata quality: From evaluation to augmentation, *Cataloging and Classification Quarterly* 2008; 46(1): 65-80.
[21] Brisaboa N, Luaces M, Rodríguez M and Seco D. An inconsistency measure of spatial data sets with respect to topological constraints. *International Journal of Geographical Information Science* 2014; 28(1): 56-82.
[22] Renteria-Agualimpia W, Lopez-Pellicer FJ, Florczyk A, López de Larrinzar J, Lacasta J, Muro-Medrano PR and Zarazaga-Soria FJ. Detectando anomalías en los metadatos de cartotecas, *Scire: representación y organización del conocimiento* 2013; 19(1): 23-29.
[23] Buchel O and Hill L. Treatment of Georeferencing in Knowledge Organization Systems: North American Contributions to Integrated Georeferencing, Knowledge organization. In *Proceedings from North American Symposium on Knowledge Organization* 2009; 2: 72-78.
[24] Southall H and Pridal P. Old Maps Online: Enabling global access to historical mapping. *e-Perimetron* 2012; 7(2): 73-81.
[25] Goodchild MF. Alexandria Digital Library. 1996. Available from: http://www.geog.ucsb.edu/~good/papers/251.pdf.
[26] Crane G. The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. *D-Lib Magazine* 1998. Available from: http://www.dlib.org/dlib/january98/01crane.html.
[27] Kanagavalli V and Raja K. 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.
[28] Jones CB. Purves R, Ruas A, Sanderson M, Sester M, Kreveld M and Weibel R. 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* 2002: 387-388.
[29] Oehrli M, Přidal P, Zollinger S, and Siber R: MapRank: Geographical search for cartographic materials in libraries, *D-lib Magazine*, 2011; 17(9/10)
[30] Park J. Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging and Classification Quarterly* 2009; 47(3-4): 213-228.
[31] Moen W, Stewart E and McClure C. Assessing metadata quality: Findings and methodological considerations from an evaluation of the US Government information locator service (GILS). In *IEEE International Forum on Research and Technology Advances in Digital Libraries*, 1998, pp. 246-255.
[32] Bruce T and Hillmann D. The continuum of metadata quality: defining, expressing, exploiting, 2004.
[33] Zeng M, Subrahmanyam B and Shreve G. Metadata quality study for the national science digital library (NSDL) metadata repository in Digital libraries: International collaboration and cross-fertilization.Springer, 2005.
[34] Shreeves S, Knutson E, Stvilia B, Palmer C, Twidale M and Cole T, Is "quality" metadata "shareable" metadata? The implications of local metadata practices for federated collections. In *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries*, 2005, pp 223-237.
[35] Harrilal B. Quality Assessment of Metadata in Open Archives, 2011.

[36] Shen R, Gonçalves M and Fox E. Key Issues Regarding Digital Libraries: Evaluation and Integration, *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2013
[37] Tolosana-Calasanz R, Álvarez J, Lacasta J, Nogueras-Iso J, Muro-Medrano PR and Zarazaga-Soria FJ, On the problem of identifying the quality of geographic metadata. In *Research and Advanced Technology for Digital Libraries*. Springer, 2006.
[38] Ma S, Lu C, Lin X and Galloway M. Evaluating the metadata quality of the IPL. In *Proceedings of the American Society for Information Science and Technology*, 2009.
[39] Hays J and Efros A. IM2GPS: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[40] Freire N, Borbinha J, Calado P, and Martins B. A metadata geoparsing system for place name recognition and resolution in metadata records. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, 2001, pp. 339-348.
[41] Renteria-Agualimpia W, Lopez-Pellicer FJ, Lacasta J, Zarazaga-Soria FJ and Muro-Medrano PR. Identifying hidden geospatial resources in catalogues. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, 2013.
[42] Barrueco J and Coll IS. Open Archives Initiative. Protocol for Metadata Harvesting (OAI-PMH): descripción, funciones y aplicaciones de un protocolo. *El profesional de la información*, 2003;12(2):99-106.
[43] Denenberg R. SRU (Search/Retrieve via URL), The Library of Congress, Washington, DC, 2007. Accessible from: http://www.loc.gov/standards/sru/.
[44] Bédard Y, Merrett T and Han J.,Fundamentals of spatial data warehousing for geographic knowledge discovery. *Geographic data mining and knowledge discovery*, 2001, pp. 53-73.
[45] Rockafellar R, Wets R and Wets M. Variational analysis. Springer, 1998.
[46] Janée G. Spatial Similarity Functions. 2003. Available from: http://www.alexandria.ucsb.edu/~gjanee/archive/2003/similarity.html.
[47] Ester M, Kriegel H, Sander J and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise, KDD, 1996, pp. 226-231.
[48] Sander J, Ester M, Kriegel H and Xu X. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications, *Data mining and knowledge discovery*, 1998, pp. 169—194.
[49] Wang J, Wang X and Liang S. GeoClustering: A Web Service for Geospatial Clustering. *Advances in Web-based GIS, Mapping Services and Applications*, 2011, pp. 37-54.
[50] Joshi D, Samal A and Soh L. Density-based clustering of polygons. In *IEEE Symposium on Computational Intelligence and Data Mining*, 2009, pp. 171-178.
[51] Wang S, Chen C, Rinsurongkawong V, Akdag F and Eick C. A polygon-based methodology for mining related spatial datasets. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics*, 2010.
[52] Renteria-Agualimpia W. and Levashkin S. Multi-criteria geographic information retrieval model based on geospatial semantic integration. *GeoSpatial Semantics*, 2011, pp. 166-181.
[53] Salton G and McGill M. Introduction to modern information retrieval, 1983.
[54] Mihalcea R, Corley C and Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity, AAAI, 2006, pp. 775-780.
[55] Janée G. Digital Curation in Encyclopedia of Database Systems. Springer, 2009.
[56] Gong P and Mu L. Error detection through consistency checking, *Geographic Information Sciences*, 200, pp. 188-193.
[57] Wang F. Handling Data Consistency through Spatial Data Integrity Rules in Constraint Decision Tables. 2008.
[58] Xie Z, Tian G, Wu L and Xia L. A framework for correcting geographical boundary inconsistency. In *Proceedings of 18th International Conference on Geoinformatics*, 2010.
[59] FGDC: Ten most common metadata errors. 1998. Available from: http://www.fgdc.gov/metadata/documents/top10metadataerrors.pdf.