Aggregation-based information retrieval system for geospatial data catalogs

Article *in* International Journal of Geographical Information Science · May 2017

DOI: 10.1080/13658816.2017.1319949

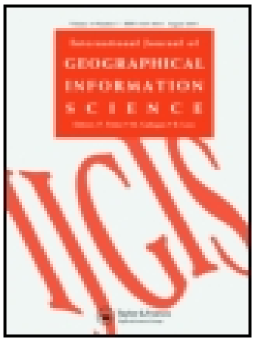
5 authors, including:

Javier Lacasta
University of Zaragoza
58 PUBLICATIONS **421** CITATIONS
SEE PROFILE

Francisco J. Lopez-Pellicer
University of Zaragoza
71 PUBLICATIONS **553** CITATIONS
SEE PROFILE

Borja Espejo García
Agricultural University of Athens
22 PUBLICATIONS **691** CITATIONS
SEE PROFILE

Javier Nogueras-Iso
University of Zaragoza
172 PUBLICATIONS **1,453** CITATIONS
SEE PROFILE



International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Aggregation-based information retrieval system for geospatial data catalogs

Javier Lacasta, F. Javier Lopez-Pellicer, Borja Espejo-García, Javier Nogueras-Iso & F. Javier Zarazaga-Soria

To cite this article: Javier Lacasta, F. Javier Lopez-Pellicer, Borja Espejo-García, Javier Nogueras-Iso & F. Javier Zarazaga-Soria (2017): Aggregation-based information retrieval system for geospatial data catalogs, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2017.1319949

To link to this article: http://dx.doi.org/10.1080/13658816.2017.1319949

Published online: 02 May 2017.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2017
https://doi.org/10.1080/13658816.2017.1319949

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

Aggregation-based information retrieval system for geospatial data catalogs

Javier Lacasta, F. Javier Lopez-Pellicer, Borja Espejo-García, Javier Nogueras-Iso and F. Javier Zarazaga-Soria

Aragon Institute of Engineering Research (I3A), Universidad de Zaragoza, Zaragoza, Spain

ABSTRACT
Geospatial data catalogs enable users to discover and access geographical information. Prevailing solutions are document oriented and fragment the spatial continuum of the geospatial data into independent and disconnected resources described through metadata. Due to this, the complete answer for a query may be scattered across multiple resources, making its discovery and access more difficult. This paper proposes an improved information retrieval process for geospatial data catalogs that aggregates the search results by identifying the implicit spatial/thematic relations between the metadata records of the resources. These aggregations are constructed in such a way that they match better the user query than each resource individually.

ARTICLE HISTORY
Received 24 August 2016
Accepted 12 April 2017

KEYWORDS
Geospatial data catalog; information retrieval; catalog service for the web; spatial data infrastructure

1. Introduction

Geographical information is commonly used by organizations, institutions and common citizens for daily work and leisure activities. In the last years, the number, variety and goals of geographical data creators and users have increased, thanks to the progressive cost reduction of the technologies needed for acquiring, processing, analyzing, accessing, presenting and transferring geographical information (Anderson and Gaston 2013, Paneque-Gálvez *et al*. 2014). Part of this cost reduction is the result of more than two decades of work by public and private initiatives to promote the generation and use of geographical information through spatial data infrastructures (SDI).

The geospatial data catalogs are responsible for facilitating the location and access to spatial resources in SDIs (OGC 2007a). The creation of international standards has facilitated its adoption. Some examples are the International Organization for Standardization metadata standards for geographical data (ISO/TC 211 2014, 2016), the Open Geospatial Consortium (OGC) standards for discovering geographical information on the web (OGC 2007b) and the implementing rules of the European INSPIRE Directive that ensure the interoperability of European SDIs (Nogueras-Iso *et al*. 2009).

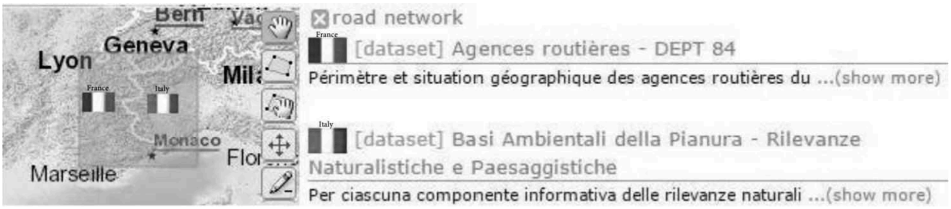
Technologically, geospatial data catalogs are similar to digital libraries that provide access to textual documents, images or any other kind of resource described with a metadata record (Smith 1996). The most basic geospatial data catalogs provide a text/

CONTACT Javier Lacasta jlacasta@unizar.es

keyword-based search system and a location-based search component to filter and sort the resources by their spatial features (Göbel and Klein 2002). This kind of query is usually named 'concept at location' query in the literature (Hübner *et al*. 2004). In the geographical context, the concepts in a 'concept at location' query are the themes of the resources. The text-based search usually provides free-text queries on the metadata records, and the selection of terms from controlled vocabularies. The location-based search component usually allows constraining the user query to an area defined by coordinates or by geographic identifiers. The answer always consists of a metadata record list with the resources that partially fulfill the query restrictions, sorted by some similarity criteria.

These approaches are valid in many situations but they have limitations. From a spatial perspective, geographical information forms a continuum that covers the Earth surface. However, the data creators divide it into independent resources that cover different spatial and thematic extents. This division is usually done to fulfill the producer goals and they ignore the nature of the stored information. For example, governments develop geographical resources that cover their country surface, but there are several geographical features such as river basins that are shared between countries. In the case of rivers, Wolf *et al*. (1999) estimated that 45.3% of the land surface corresponds with river basins covering more than one country. In this context, the organization of the information into datasets spatially delimited by the boundaries of countries and other kinds of administrative regions enters into conflict with users that need continuous data (e.g. a pan European provider of road maps, or a flood risk manager covering different autonomous regions in a country). Moreover, data belonging to related topics in the same area may be scattered across different datasets. That is, this producer-oriented approach may enter into conflict with users who need continuous data.

The pan-European INSPIRE geoportal[1] is an example of a system with the previous limitations. This geospatial data catalog was created for providing support for the implementation of the INSPIRE Directive. It is a state-focused geospatial data catalog where each member state describes its geospatial resources. This focus may lead to incomplete answers for some kind of queries. For example, Figure 1 shows the answer to a query about 'road networks' in an area between Italy and France. This kind of query will always produce incomplete results because there are not resources in the catalog covering both countries. In this example, the first result describes exclusively the road networks in France and the second one describes only those in Italy. This data partition makes ranking an unhelpful feature. Each result only provides a part of the required information, and the user is forced to review all of them to compose a set of suitable



☒road network

France

[dataset] Agences routières - DEPT 84

Périmètre et situation géographique des agences routières du ...(show more)

Italy

[dataset] Basi Ambientali della Pianura - Rilevanze Naturalistiche e Paesaggistiche

Per ciascuna componente informativa delle rilevanze naturali ...(show more)

Figure 1. Example of query answer from the INSPIRE geospatial data catalog.

road resources that fulfill his needs. This review task is challenging because the lack of feedback about how the search parameters define the search results makes difficult the comparison and interpretation of results (Göbel and Klein 2002). Moreover, in the same way as there is spatial fragmentation, there is also thematic fragmentation. As each resource may contain only a small set of the themes of the information available about an area, in multi-theme queries, the results will also be thematically fragmented. Nowadays, the geospatial data catalogs of public institutions (such as the INSPIRE geoportal) are the most technologically advanced, but all of them have similar problems in terms of results interpretation.

The main contribution of this paper is an improved information retrieval (IR) process for geospatial data catalogs that aggregates the search results by identifying the implicit spatial/thematic relations between the metadata records of the resources. These aggregations are constructed in such a way that they match better the user query than each resource individually. The returned aggregations are composed of metadata records that describe resources that complement each other and fill the spatial gaps that each individual resource has for each queried theme. This paper is focused on analyzing the suitability of the aggregation of the metadata records provided as query results in the geospatial data catalog context. To analyze the result composition issues, other IR issues such as terminological heterogeneity or the use of imprecise spatial references in user queries are left aside. The system performance is evaluated comparing the behavior of the proposed IR system with another one that is similar to those used in prevalent geospatial data catalogs.

The rest of the paper is organized as follows. Section 2 reviews other works related to IR systems for geospatial data catalogs. Section 3 explains the IR issues that we analyze in this paper. Section 4 describes the proposed IR system, which is evaluated through a series of experiments described in Section 5. Section 6 discusses the obtained results. Finally, the paper ends with some conclusions and outlook on future work.

2. State of the art

There are many works that have proposed IR improvements through better similarity measures for result ranking, and through the increase of the metadata and query description quality. This section reviews a selection of these works.

Related to the definition of a spatial similarity measure between resources for ranking purposes, on the web context, Watters and Amoudi (2003) propose as a ranking factor for queries the distance between the place where the user is located and the place where the web server with the relevant data is located. They translate URLs into the spatial coordinates of the place where the web domain is situated and they use the linear distance of these coordinates to the user spatial location as a ranking factor of the results. Asadi *et al*. (2005) analyze the different types of textual queries that involve spatial information and describe how to adjust their ranking formulas. They review direct queries about facts, local queries that restrict the relevant results to those describing an area and location-based queries where the objective is to locate specific entities in an area (e.g. train stations).

Focused on geospatial data catalogs, Larson and Frontiera (2004) describe a statistically based ranking formula for geometry-based spatial queries. They make a review of previous spatial-based ranking formulas and propose a statistical measure that includes a corrective factor to deal with the problems caused by the imprecise definition of bounding boxes in border areas such as coasts. Lanfear (2006) suggests another ranking method for spatial features in geospatial data catalogs that takes into account the overlap between the query area and the resource, and the dimensions of the area outside the overlap. More recently, Renteria-Agualimpia *et al*. (2016) detect incoherencies in metadata collections by comparing the explicit geographical extension defined by coordinates and an implicit one defined by geographic identifiers found in metadata records. Their use of the Hausdorff distance to detect how similar are two geometries can be directly used in the IR context to determine a spatial similarity measure of a resource with the user query.

In addition to the previous works, there are ranking proposals that focus on the integration of different relevance measures. Göbel and Klein (2002) propose a linear ranking formula that in addition to the similarity with the spatial feature of the query (both coordinate and gazetteer based), it includes the degree of thematic coincidence and temporal overlap. Martins *et al*. (2005) compare different approaches to generate a combined ranking value from individual spatial and thematic distances. This comparison includes the use of a linear combination, the product, the maximum similarity and a step-linear function. Finally, Megler and Maier (2011) present a ranking method for integrating spatial and temporal query features based on the mean between the spatial and temporal distances to the center of the selected period or selected area.

Another approach frequently used to obtain better IR systems for geospatial data catalogs has been to improve the resource and the query descriptions. Lieberman (2006) describes an SDI architecture where online resources are able to self-describe themselves. This solution requires a semantic facade on top of OGC standard services that describes their content through the use of ontologies. Lutz *et al*. (2009) propose instead a semantic catalog where geospatial resources are described using roles and concepts from a domain ontology. Somehow related, Janowicz *et al*. (2010) propose a transparent semantic layer for SDI. They annotate the resource descriptions using ontologies and they relate these ontologies through a reasoning service implemented as a profile of an OGC catalog and a processing service, respectively. Finally, Florczyk *et al*. (2010) add semantics into an SDI catalog with a linked data administrative geography ontology that is used for data integration and referencing geographic themes.

Related to the query description improvement, other works have focused on the identification of textual patterns describing locations or location-based references (e.g. *north of X*). Works such as Sallaberry (2013), Ferrés and Rodríguez (2015) and Kim *et al*. (2017) show that the identification of textual patterns describing locations can greatly improve the quality of the results when spatial description in metadata records is textual. However, in geospatial data catalogs, these solutions are less relevant because, by design, the metadata records of spatial resources specify their spatial limits as coordinates.

Our process focuses on automatically providing improved aggregated search results for geospatial data catalogs by using raw metadata. There have been other proposals that perform aggregations of resources at the data level but they require either human intervention or an extra layer of complexity such as adding domain knowledge in the form of ontologies. For example, Hübner *et al*. (2004) describe an ontology-based reasoning system that integrates heterogeneous geographical information in 'concept at location in time' queries. The user employs provided ontologies to define a query and the system returns a list of resources sorted by relevance. Then, the system facilitates its visualization and integration. Similarly, Lutz and Klien (2006) propose a retrieval system in which features published at Web Feature Service (WFS)s are described in terms of a shared domain ontology. This system offers a user interface that allows formulating queries using such ontology. A different approach can be found in Latre *et al*. (2009). This work describes a retrieval system that identifies non-explicit relations between hydrologic feature types published at WFSs and uses this knowledge to expand results. Finally, Zhu *et al*. (2015) describe a user-focused spatial data analysis service that unifies the access to heterogeneous data by creating linked layers after parsing user requests.

3. Spatial and thematic issues in geospatial data catalogs

This section reviews the IR systems used in a representative set of geospatial data catalogs and describes their features and issues. We have analyzed the pan-European INSPIRE catalog, and the national catalogs in USA[2] (GeoPlatform), Spain[3] (IDEE), United Kingdom[4] (Data.Gov) and Canada[5] (GeoDiscovery). Below, we present an analysis of their user query interfaces and how they answer when the queries include spatial and thematic constraints. Then, we describe the issues that the process described in this paper tries to correct.

3.1. *Prevalent approaches*

Table 1 shows the type of search and ranking provided by each analyzed system. All the analyzed systems provide free-text search and some kind of controlled topic list or faceted solution. Additionally, their advanced search components focus on specific metadata fields, such as the resource type or data format. Among them, only the Spanish and Canadian systems offer temporal search (periods of time). Regarding the spatial search, the queries in the UK catalog cannot include textual and spatial features simultaneously. The remaining systems provide a bounding box-based spatial search component that can be combined with other query elements.

In order to determine how the search process is performed, we have analyzed the result of queries using controlled fields, queries using free-text fields, queries using a spatial bounding box and queries with the three restrictions. The query terms have been selected so they return multiple resources (e.g. 'road network' in INSPIRE geoportal) and we have counted the occurrences of the query terms in each of the obtained metadata records and the percentage of spatial area in query that they cover.

Table 1. Search and ranking features.

Catalog	Country	Type of search	Type of ranking
INSPIRE	EU	Spatial, free text, term cloud, topics	Relevance
GeoPlatform	USA	Spatial, free text, facets	Relevance, popularity
IDEE	Spain	Free text, topics, date	Rating, relevance, popularity
Data.gov.uk	United Kingdom	Spatial or free text, facets	Relevance, popularity
GeoDiscovery	Canada	Spatial, free text, date, topics	Relevance

Through this analysis, we have found that the systems do ‘AND’ style queries when the queries involve two or more types of constraints (spatial, controlled or free text). That is, the responses only contain records that match all the constraints. The results can be sorted according to a relevance rank (some of them also include popularity and user rating) or alphabetically by different fields (e.g. title). Regarding the relevance rank, the number of occurrences of the query term in any part of a metadata record is used as a ranking factor (the metadata records with more query term occurrences are first). However, when the query involves controlled fields, only the existence is taken into account as ranking factor. We also tried to identify if any of the systems uses ontologies, or any other kind of formal model, for query expansion or refinement. However, since all the results in the tested systems contain the used query terms, it seems that queries have not been expanded with additional terms such as synonyms, hypernyms or hyponyms. In the systems supporting spatial restrictions, the more the query area and the geographical extent of a resource overlap, the better its rank is.

We have been unable to identify the exact ranking formula used for combining the spatial and textual rankings in the systems that support ‘concept at location’ style queries (INSPIRE, GeoPlatform and GeoDiscovery). However, we have detected that, in these systems, a spatially closer resource is ranked first even if it has far fewer occurrences of the textual query terms. This indicates that the ranking weight given to the spatial similarity is higher than the used one for the textual similarity.

The functionalities offered by these systems seem suitable in many situations but they are problematic when performing queries about multiple themes in an area crossing multiple countries or regions. In these cases, the results obtained are similar to those described in Figure 1. That is, the results are partially relevant and none can be considered a complete answer.

3.2. *Data fragmentation issues*

The search features identified are very common and they can be found in digital libraries outside the spatial field (e.g. Europeana[6]). It is important to note that the reviewed systems manage the resources as independent entities. However, in the geospatial context, the resources are related by the themes they cover and by spatial proximity (all this is indicated in their metadata records). If these relations are not taken into account, when a user query does not fit the artificial divisions (spatial and thematic) of the data continuum performed by the data creators, the catalogs will return incomplete results. Depending on the user query and the spatial and thematic data fragmentation, the answer may suffer from under-coverage, over-coverage and partial coverage of the results with respect to the query. Below, we are going to characterize each of these issues.

The first issue is the under-coverage of the results with respect to the query. At spatial level, this happens when the results include resources that only slightly intersect with the query bounding box. At thematic level, it happens when the result includes resources about a small subset of the query themes. This is a problem because resources that only slightly fulfill the query may be considered very relevant results. As an example of the spatial under-coverage, Figure 2(A1) shows the bounding box of a query focused on the ‘Castilla la Mancha’ region in Spain (continuous line) and a resource focused on ‘Valencia’ region that only slightly intersects with the query

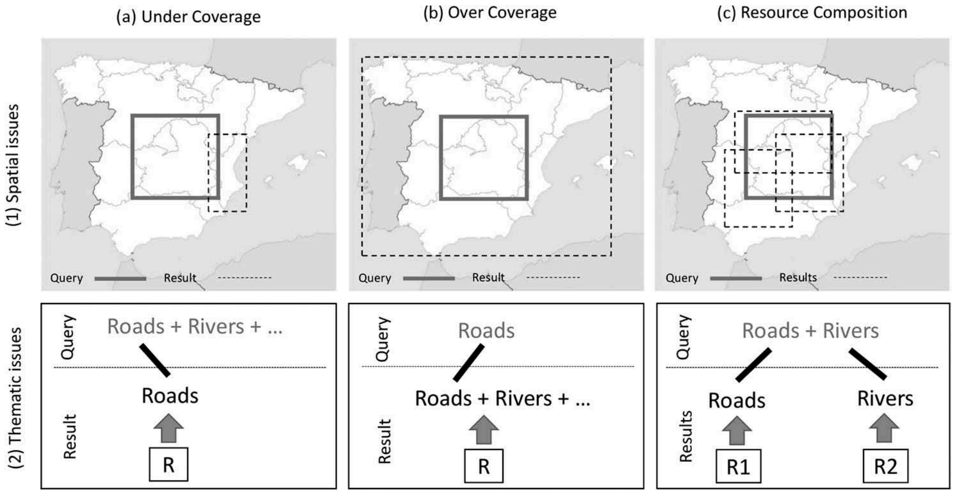


Figure 2. Spatial and thematic issues in the IR system of geospatial data catalogs. (a) under coverage, (b) over coverage, (c) resource composition.

(discontinuous line). Figure 2(A2) shows a thematic under-coverage example. It contains a query about many subjects ('roads', 'rivers' and others) and a result (R) detailing only 'roads'. In both cases, the amount of information provided with respect to the requested query is small. Thus, although they fulfill the query, they have little relevance.

The second issue is the over-coverage of the results with respect to the query. At spatial level, it happens when the results include resources that cover an area much bigger than the requested one. At thematic level, it occurs when a result contains information about many more themes than the requested ones. Over-coverage is a problem in the sense that the amount of irrelevant information in the results makes difficult to identify the desired content. Any other result more adjusted to the query is probably a better option for the user. The spatial over-coverage example is shown in Figure 2(B1). It depicts the same 'Castilla la Mancha' query, but in this case, the result covers the desired region and the rest of the Iberian Peninsula. Figure 2(B2) shows a thematic over-coverage example with a query about 'roads' and a result (R) containing information about 'roads', 'rivers' and many other themes. In both cases, the result contains not only relevant information but also a disproportionate amount of irrelevant data.

The last issue happens when there are many partial results to a query. All of them are partially relevant, but none of them completely fulfill the query specification. Existent systems sort these results according to a spatial/theme similarity criterion (usually some spatial/theme overlap variant). However, when results are presented in this way, it is difficult to distinguish which areas/themes of the query are described by each resource and how they complement each other. An example of the spatial aspect of the partial coverage is shown in Figure 2(C1). It shows the same 'Castilla la Mancha' query with multiple partial relevant results. Figure 2(C2) continues with the previous 'roads' and 'rivers' query but providing as answer a resource about 'roads' and a different one about 'rivers' to show the thematic partial coverage. In both cases, none of the results are a complete answer to the query.

These spatial and thematic issues can happen in any combination in a 'concept at location' style query, especially when resources only cover part of the spatial area and part of the requested themes. In this case, as previously indicated, the results generated by the reviewed systems are not able to completely fulfill the query restrictions. Next section proposes an IR system able to deal with these issues by aggregating the metadata records in the result list into collections of compatible records that, as a set, are a better answer to the query. The construction of these aggregations helps to solve the composition issue and mitigates under and over-coverage problems.

4. Generating thematically and spatially aggregated results

In a classic IR system, when performing the intersection between the metadata of a resource and the query to determine if it is relevant, only a subset of the themes may intersect, and only a part of the query area may be covered. This means that each retrieved resource only provides a partial result. However, combining it with others, a more complete result could be obtained. For example, in a query about 'highways' in 'Spain,' we can find a resource about highways in the south of Spain. This result is incomplete but if combined with other one covering the north, we can compose a good result. The same happens with respect to the themes. For example, in a query about 'highways' and 'motorways' in Spain, a resource about Spanish motorways may be the perfect complement to another one depicting the Spanish highways.

Figure 3 shows the main steps of the IR process created to generate aggregations of metadata records as results of 'concept at location' queries. The query analysis step is a simple decomposition process where the query is processed to separate spatial (bounding box) and thematic (keywords) requirements. The next step is to obtain the metadata records describing resources that are partially relevant to the query. This is done using a spatial and an inverted textual index. Only those that intersect spatially with the query area and contain at least one of the query keywords are returned. Then, the obtained metadata records are sorted according to their relevance degree. Finally, the IR process aggregates the records in suitable groups. This section focuses on describing how the ranking and aggregation process is performed.

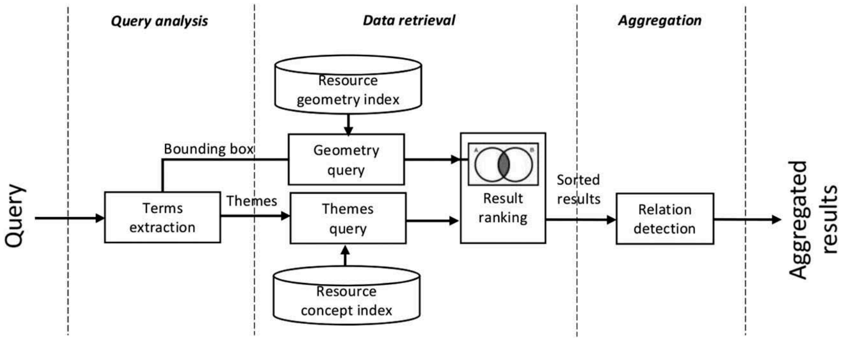


Figure 3. Process for aggregation of query results.

Equation (1) shows the similarity formula used for result ranking. It represents the similarity with a value between 0 and 1, 0 being 'irrelevant' and 1 'perfect match'. In the formula, we use the following symbols: $dH(G_{\mathrm{Q}}, G_{\mathrm{R}})$ represents the Hausdorff distance between the query geometry (G_{Q}) and a metadata record geometry (G_{R}); *Max DH* is the biggest Hausdorff distance of all the partially relevant resources with respect to the query; $\mathrm{size}\left(T_{\mathrm{Q}} \cap^{T_{\mathrm{R}}}\right)$ indicates the number of themes in common between the metadata record (T_{R}) and the query (T_{Q}); and $\mathrm{size}(T_{\mathrm{Q}})$ indicates the number of themes in the query.

$$\mathrm{Similarity}(\mathrm{Q},\mathrm{R}) = \frac{\mathrm{Max}\ DH - dH\ (G_{\mathrm{Q}}, G_{\mathrm{R}})}{\mathrm{Max}\ DH} \times \frac{\mathrm{size}(T_{\mathrm{Q}} \cap T_{\mathrm{R}})}{\mathrm{size}(T_{\mathrm{Q}})} \tag{1}$$

The Hausdorff distance is the greatest of all the distances between any point in a geometry and the closest point in another geometry. Since the Hausdorff distance between geometry A and B may be different from the Hausdorff distance between geometry B and A, the maximum is used. The Hausdorff distance of overlapping geometries of similar size is smaller than the equivalent one between overlapping geometries with very different dimensions. Therefore, it is very appropriate for ordering resources of different administrative levels (e.g. region vs. country size). Additionally, the Hausdorff distance can be used with complex geometries. Thus, replacing the metadata bounding boxes with approximate geometries would directly increase the quality of results without having to modify the IR system. This is also valid for resources with multiple disjoint geometries (e.g. Iberian peninsula and Canary islands) that can be represented as a single multi-polygon geometry for distance measure purpose. A problem of the Hausdorff distance is that it can give small distances to nonoverlapping resources (if they are similar in size and they are spatially close). However, this issue is not relevant for our system because the ranking is only applied on resources that overlap.

Once the ranking of the results has been performed, the last step aggregates the results that are spatially and thematically compatible. We consider that two metadata records (and therefore the resources they describe) are spatially compatible if the combined area for all their themes is significantly closer to the query area than each record individually. Regarding the thematic compatibility, we only consider compatible those that are thematically disjoint (they do not have any common query themes), and those that share one query theme and also half of the rest of the keywords. This avoids aggregating records of resources that describe a theme in very different ways. For example, if we are making a query about 'river basins', we may find a resource that focuses on the 'water flow' and another one describing the 'geology' of the basin. In this case, they are too different to be in the same aggregation. When the records do not share a query theme, they can always be aggregated because they are fulfilling disjoint restrictions in the query.

Algorithm 1 describes the method used to aggregate the list of ranked results obtained with the previous steps. For each metadata record obtained as result, the process searches other results that complement it. The aggregation process is performed only if there is a relevant part of the query area that is not covered in all the themes. The *coverageFactor* indicates the percentage of area in the query (summing up all themes) that can be left uncovered. The smaller it is, the more complete the results are. However, it is better to not completely cover the query bounding box with results to deal with imprecisions in the definition of the query or the resources. For example, it is counterproductive to complement a resource with a 99% of query coverage just to cover a small gap in a border. The value selected in the experiment section is a compromise between a complete coverage of the query and the management of deficiencies in the query formulation and the resources.

Algorithm 1 Spatio-thematic aggregation of results.

```
  function AGGREGATIONSTEP($results, query$)
    $aggregationList \leftarrow \emptyset$
    for $result \in results$ do
      $aggregation \leftarrow result$
      $reducedQuery \leftarrow query - aggregation$
      $resultExtended \leftarrow true$
      while $area(reducedQuery) > coverageFactor \times area(query)$ & $resultExtended$ do
        $resultExtended \leftarrow false$
        $possibleAggregated \leftarrow getBestResult(results, reducedQuery, aggregation)$
        if $possibleAggreated \neq \emptyset$ then
          $reducedQuery \leftarrow reducedQuery - possibleAggregated$
          $aggregation \leftarrow aggregation \cup possibleAggregated$
          $resultExtended \leftarrow true$
        end if
    end while
    $aggregationList \leftarrow aggregationList \cup aggregation$
   end for
   return $duplicateRemoval(aggregationList)$
 end function
```

The search of results that complement a given one is done by the function *getBestResult* depicted in Algorithm 2. This process is repeated until no more suitable resources for the aggregation are found. The identification of suitable complementary results is done by removing those that are spatially and thematically incompatible with the current aggregation (*thematicFilter* and *spatialFilter* functions). Then, the rest are sorted according to the dimension of the uncovered part of the query, and the closest one is selected.

Algorithm 2 Function to obtain a new element for an aggregation.

```
  function GETBESTRESULT($results, reducedQuery, aggregation$)
    $filteredResults \leftarrow spatialFilter(results, reducedQuery, infoFactor);$
    $filteredResults \leftarrow thematicFilter(filteredResults, reducedQuery, aggregation);$
    $sortedResult \leftarrow rankResults(filteredResults, reducedQuery)$
    if $sortedResult \neq \emptyset$ then
```

```
        return $sortedResult[0]$
    else
        return $\emptyset$
    end if
function spatialFilter($results, reducedQuery, infoFactor$)
    $filteredResult \leftarrow \emptyset$;
    for $result \in results$ do
        if $area(intersection(result, reducedQuery)) > area(reducedQuery) * infoFactor$ then
            $filteredResult \leftarrow filteredResult \cup result$;
        end if
    end for
    return $filteredResult$
end function
function thematicFilter($results, reducedQuery, aggregation$)
    $filteredResult \leftarrow \emptyset$;
    for $result \in results$ do
        $themesInQuery = themes(result) \cap themes(reducedQuery)$;
        $themesInAggr = themesInQuery \cap themes(aggregation)$;
        $commonThemes = themes(result) \cap themes(aggregation)$
        if $result \notin aggregation \wedge ((themesInQuery \neq \emptyset \wedge themesInAggr == \emptyset) \vee (themesInAggr \neq \emptyset \wedge size(commonThemes) \geq size(themes(aggregation)) \div 2)$ then
            $filteredResult \leftarrow filteredResult \cup result$;
        end if
    end for
    return $filteredResult$
end function
```

In the process to identify the best possible record to add to an aggregation depicted in Algorithm 2, the spatial and thematic filters avoid adding resources that only improve the results in a negligible amount and the creation of thematically heterogeneous aggregations. The spatial filter removes the metadata records of resources that do not cover a relevant part of the query area uncovered by the aggregation. This behavior is adjusted by the *infoFactor* parameter that represents the amount of new information a resource has to provide to be included in the aggregation. A low value generates more complete aggregations, but some of their components may provide very little new information. A high value creates aggregations with more relevant elements, but it may leave important parts of the query uncovered. The thematic filter removes those results already in the aggregation and those that share a theme with the query and the aggregation but do not have in common at least half of the keywords. This is done because results with fewer keywords in common are likely to be too different between them to be integrated, even if they share a queried theme. After applying both filters, the remaining results are ranked according to the similarity formula shown in Equation (2), and the most similar one is selected as a new element in the aggregation.

$$\mathrm{Similarity}(\mathrm{Uq},\mathrm{R}) = \frac{\sum_{T\in\mathrm{Uq}}\left(\mathrm{Max}\,DH_T - dH\left(G_{T_{\mathrm{Uq}}}, G_{T_{\mathrm{R}}}\right)\right)/\mathrm{Max}\,DH_T}{\mathrm{Size}\left(T_{\mathrm{Uq}}\right)} \tag{2}$$

Equation (2) is a generalization of Equation (1). Since we try to find the metadata record of the resource that is the most similar to the area of the themes that is not covered by the current members of an aggregation, the geometry of each theme in the query is different. For example, in a query about 'highways' and 'motorways' in Spain, we may have constructed an aggregation with a resource with the highways in the south of Spain, and another one covering the motorways in the east. In this context, the extension that is needed to cover with additional resources is different for the theme 'motorways' and the theme 'highways'. In the equation, we calculate the similarity of a metadata record (R) that is candidate for the aggregation with respect to the area of the query themes not covered by the aggregation (Uq). It is calculated as the sum of the spatial similarity of each theme of the query with some spatial extension uncovered with respect to the metadata record extension for these themes, divided by the number of query themes that have a spatial part uncovered ($\mathrm{Size}\left(T_{\mathrm{Uq}}\right)$). The spatial similarity for each theme is obtained in a way analogous to Equation (1). In Equation (2), the following symbols are used: $\mathrm{Max}\,DH_T$ represents the maximum Hausdorff distance between the theme extension of all the candidates and the uncovered extension of the query for this theme, and $dH\left(G_{T_{\mathrm{Uq}}}, G_{T_{\mathrm{R}}}\right)$ is the Hausdorff distance of the theme of the metadata record that is being analyzed ($G_{T_{\mathrm{R}}}$) with respect to the uncovered part of the query for the theme ($G_{T_{\mathrm{Uq}}}$).

This process may generate redundant aggregations with the same elements in different order (e.g. it can aggregate the 1st result with the 10th one and then aggregate the 10th result with the 1st one) and aggregations that are a superset of another one (in this case, some elements in the superset are not relevant). The last step removes these redundancies.

5. Experiments

This section compares the performance of the proposed IR process (*Aggregated IR System*) with a basic IR system (*Basic IR System*) similar to the ones used in the geospatial data catalogs described in Section 3.1.

Our *Aggregated IR System* has been tuned to try to create aggregations with at least a 90% of query coverage ($\mathit{coverageFactor} = 0.1$) and to add elements to the aggregation even if they only provide a small improvement of the result ($\mathit{infoFactor} = 0.1$). The *Basic IR System* used for comparison applies a similarity measure that behaves similarly to the geospatial data catalogs described in Section 2 (see Equation (3)). This measure performs a combination between the spatial intersection of the query and the metadata record of the resource, and the theme intersection. The spatial similarity is obtained as the area of the intersection between the query (A_{Q}) and the record (A_{R}), divided by the maximum area of intersection between all the resources and the query. For the thematic similarity, we have directly used the Jaccard coefficient, which is calculated as the number of themes in common between the query (T_{Q}) and the record (T_{R}) divided by the total number of themes. Finally, this similarity values are weighted with α and β factors to be able to adjust the weight of the spatial aspect (α) of the query with respect to the thematic ones (β). To avoid giving any advantage to our proposal, the experiments with the *Basic IR System* have been performed multiple times with different α and β values, and the best obtained results have been the ones used in the comparison.

$$\mathrm{Similarity}(\mathrm{Q},\mathrm{R}) = \alpha\left(\frac{\cap(A_{\mathrm{Q}}, A_{\mathrm{R}})}{\mathrm{maxAreaIntersection}}\right) + \beta(\mathrm{Jaccard}(T_{\mathrm{Q}}, T_{\mathrm{R}})) \quad (3)$$

5.1. *Evaluation methodology*

For this experiment, we have used the metadata records provided through the Geoportal of the Spanish National Spatial Data Infrastructure[7] (IDEE) in 2015. This collection contains 97,867 records describing geographical resources created by different Spanish governmental institutions. This includes themes so different such as topology, environment, mineral resources, industry and infrastructures, among other themes.

The performance of an IR system to solve the issues described in Section 3 cannot be simply described in terms of classical precision/recall measures. These measures are often based on a binary classification results as relevant and nonrelevant as a whole (Baeza-Yates and Ribeiro-Neto 2011). In our system, only the metadata records of resources that contain part of the selected area and some of the query themes are returned. Therefore, all the results contain at least a bit of relevant information. The problem here is to measure the degree of relevance for result ordering.

The proposed system is focused on improving the results of ‘concept at location’ queries. The objective of this type of queries is to return first the results that have an exact match with the query restriction, second those that cover the selected area but include much additional information (over-coverage), third those that have only a partial coverage, and finally results that only slightly fulfill the query restrictions (under-coverage).

To evaluate the ranking of the two systems, we have used the discounted cumulative gain (DCG) measure shown in Equation (4) (Baeza-Yates and Ribeiro-Neto 2011). This measure calculates the gain of adding each document to the result set based on its position in the result list. To obtain this measure, it is needed to describe the gain that each result adds to the result list (G_i). For this task, we have used the gain criteria described in Table 2. In these criteria, higher values indicate that the result is more adjusted to the spatial or thematic restrictions in the query. The lower ones indicate that there is less similarity. The spatial and thematic content of each result of the analyzed queries has been classified according to these criteria. The final gain of each result is calculated as the mean of the spatial and the thematic gains.

Table 2. Criteria values used to determine the quality of a result.

Gain value	Meaning	Description
3	Exact match	The spatial or thematic features of the result are approximately equal to the query
	Over-coverage	The spatial or thematic features of the result approximately cover the query but they are much more extensive
	Partial coverage	The spatial or thematic features of the result just cover a part of the query
	Under-coverage	The spatial or thematic features of the result just slightly cover a part of the query

$$\mathrm{DCG}[i] = \begin{cases} i = 1, G_1 \\ i \neq 1, G_i/\log_2(i) - \mathrm{DCG}_{i-1} \end{cases} \tag{4}$$

5.2. *Description of the experiments and results obtained*

The main advantage of our system with respect to a basic one is that it is able to identify subsets of low gain results and transform them into higher gain aggregations. For example, it can combine several results with partial coverage of the query to obtain an exact match. To evaluate how the system performs, we have selected four themes commonly used in fields such as hydrology, ecology, infrastructure planning, industry or agriculture [some recent examples on the interest in these areas can be found in Graser *et al*. (2015) and Pereira *et al*. (2015)] that have a high presence in the collection and four spatial areas that contain information about these themes. The themes are 'elevation' (model), 'road network', 'soil use' and 'hydrography'. Using these themes and areas, we have generated all the possible queries that include one or two of the themes and one spatial area. This makes a total of 40 different queries (combinations without repetitions of 2 themes selected from the 4 original ones plus the 'empty' theme, and the 4 different areas, i.e. $\binom{5}{2} * 4$). For these queries, we have obtained the DCG for the 10 first positions of their result lists, and we have calculated the mean DCG at each position. This mean takes into account that some queries return less than 10 results. Figure 4 compares the mean DGC of the two systems at each position of their result list (number of query result). It can be observed how the aggregated system has always a higher mean gain. This means that the obtained resources and the way they are positioned in the result list are better in the aggregated system than in the basic one.

To explicitly show how the system behaves with respect to the undesired effects described in Section 3, we analyze in detail the results of a small set of the selected queries. These queries show how our system behaves in two main scenarios: when there are results that perfectly match the spatial and/or thematic query restriction, and when there are not close matches. Table 3 summarizes the selected queries (Q1–4). The table shows the query bounding box (min and max longitude, latitude), a toponymical reference (Location) of the Spanish region containing the bounding box (for illustrative

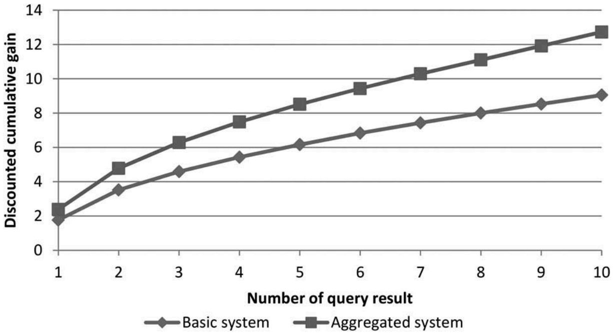


Figure 4. Mean DGC comparison.

Table 3. Queries selected for the evaluation.

Number	Bounding box (min; max)	Location	Themes
Q1	−6.02, 37.37; −5.93, 37.41	Andalucía	Elevation
Q2	−6, 37.35; −5.9, 37.41	Andalucía	Elevation
Q3	−8.38, 42.25, −7.50, 43.08	Galicia	Road network, soil use
Q4	−7.8, 42.21; −5.9, 43.29	Galicia, Castilla y León	Road network, soil use

purposes) and the themes in the query (Themes). In the case of the first query (Q1), there are results that perfectly match the query restrictions. The detailed analysis of the result ordering is illustrative of the difference in behavior between our system and the basic solution. The rest of the queries display result sets that do not contain any result that perfectly match the query. Specifically, Q2 focuses on the spatial features. It requests information about a theme (‘elevation’) in a region (a part of ‘Andalucía’) where there are not resources that fit well with the selected area for the selected theme. However, there are resources that have over-coverage and others with partial coverage. Q3 focuses on the thematic features. It includes multiple themes (‘road networks’ and ‘soil use’) and it selects a region (*Galicia*) that contains resources that match the spatial aspects of the query but only partially match the thematic aspects. Q4 describes the more general case where none of the resources match well the query area (a part of *Galicia* and *Castilla y León*) and the selected themes (‘road networks’ and ‘soil use’ again).

Table 4 shows a summary of the performance of each system, measured as the spatial and thematic similarity of the results with respect to the query specification. It includes the number of results obtained from each query (*Num Res*), the mean spatial coverage of the results with respect to the spatial restriction in the query (*Mean SpCov*), the mean thematic overlap between the results and the thematic restrictions in the query (*Mean ThCov*) and the size of the biggest aggregation obtained (*Max AggSize*). The spatial and thematic mean coverage visualize the degree of fulfillment of the user needs, while the size of the biggest aggregation indicates the number of individual results that are needed to fulfill the query in the worst case. Table 5 details the first three results of each query. It includes the title of the result in each position (result order), the percentage of spatial (SpCov) and thematic (ThCov) coverage of each result with respect to the query and the gain value (Gain) obtained according to the criteria indicated in Table 2. The results that aggregate several resources to compose a better result are marked with (A). In the figure, it can be observed the difference between the results of the basic system, where most of them have over-coverage and partial coverage, and the ones obtained in the aggregated system, which generates aggregations closer to the query constraints.

Table 4. Comparison of system results.

	Basic IR			Aggregated IR			
	Num Res	Mean SpCov (%)	Mean ThCov (%)	Num Res	Mean SpCov (%)	Mean ThCov (%)	Max AggSize
Q1	12	51.37	100	6	99.98	100	1
Q2	9	55.55	100	6	100	100	4
Q3	25	100	50	24	100	100	2
Q4	31	81.15	50	29	97.5	100	4

Table 5. Detailed comparison of system results.

Order		Basic IR				Aggregated IR			
		Result title	SpCov (%)	ThCov (%)	Gain	Result title	SpCov (%)	ThCov (%)	Gain
Q1	1	Orography of Andalucía	100*	100	2.5	Andalucía EDM 98433	99	100	3.0
	2	Contour lines	100*	100	2.5	Orography of Andalucía	100*	100	2.5
	3	Digital Terrain Model	100*	100	2.5	Contour lines	100*	100	2.5
Q2	1	Orography of Andalucía	100*	100	2.5	(A) Andalucía EDM 98433/34/43/44	100	100	3.0
	2	Contour lines	100*	100	2.5	Orography of Andalucía	100*	100	2.5
	3	Digital Terrain Model	100*	100	2.5	Contour lines	100*	100	2.5
Q3	1	Topographic Base of Galicia	100	50	2.0	(A) Topographic Base of Galicia/Map of Coverages and Soil Uses	100	100	3.0
	2	Map of Coverages and Soil Uses	100	50	2.0	(A) Map of Coverages and Soil Uses/Basic Cartography of Galicia	100	100	3.0
	3	Basic Cartography of Galicia	100	50	2.0	(A) Topographic Base of Galicia/Soil Uses, Polygons	100	100	3.0
Q4	1	CORINE Land Cover 1990	100*	50	1.5	(A) Topographic Base of Galicia/Transport Network CyL/Map of Coverages and Soil Uses/Land Cover CyL	97	100	3.0
	2	CORINE Land Cover 2000	100*	50	1.5	(A) Basic Cartography of Galicia/Transport Network CyL/Map of Coverages and Soil Uses/Land Cover CyL	97	100	3.0
	3	CORINE Land Cover changes 1990–2000	100*	50	1.5	(A) Soil Uses. Polygons/Topographic Base of Galicia/Transport Network CyL/Land Cover CyL	96	100	3.0

Spatial coverage percentages marked with a star (*) indicate a big spatial over-coverage (they are resources at country level for queries about a small region).

Q1 is representative of the situation when a query has a perfect match with the collection. It has been selected because there is a resource in the collection that matches at 99% the query bounding box (tile 98433 of the 'Andalucía Elevation Digital Model'). Additionally, there are five resources relevant for the query theme but that cover all Andalucía/Spain (they have spatial over-coverage). Finally, there are other six thematically relevant resources with spatial under-coverage. In the basic system, the most relevant resource is provided as the sixth result and the previous places are occupied by the resources with spatial over-coverage that completely cover the query area. The results with under-coverage are placed last. In the proposed system, no aggregation is generated for this query, but the ordering is improved since the best result is placed first and those with over-coverage are sorted according to the spatial similarity with the query. Additionally, the resources with spatial under-coverage are not returned because the aggregation process identifies that they are only reliable if complemented with another one that is reliable by itself.

Q2 analyzes the behavior of the systems when there are resources completely covering the thematic restrictions but not the spatial ones. For Q2, the collection contains nine relevant resources, five with spatial over-coverage (the same in Q1) and four with partial spatial coverage. They are four tiles of the 'Andalucía Elevation Digital Model' ('98433', '98434', '98443' and '98444'). In the basic system, the five first results are those with spatial over-coverage, and the last four ones are those that have partial coverage. In the proposed system, the four resources with partial spatial coverage are aggregated into a single result that perfectly matches the query. This aggregation is provided first in the result list. The rest of the results are sorted according to their spatial distance with respect to the query.

Q3 analyzes a scenario with resources that cover the spatial aspects of the query but only partially the thematic ones. For this query, there are 25 resources focused on *Galicia* and Spain about 'road networks' and 'soil use', but none containing both. In the basic IR system, the obtained results are not distinguishable since they all completely cover the query area and contain one query theme. In the proposed system, 24 compatible aggregations that fulfill the user needs are found. These aggregations add to each result (i.e. focused on one theme) the spatially closest result of the other theme. For example, the first result is the aggregation consisting of the two first results of the basic system, i.e. 'Topographic Base of Galicia' (providing road networks) and 'Map of Coverages and Soil Uses' (providing soil use).

Finally, Q4 focuses on the most general case: a query that has no clear candidate for the thematic and spatial query restrictions. It uses the same themes as the third query, but it covers an area that includes part of *Galicia* and their neighbor region of *Castilla y León*. In the collection, there are 31 relevant resources, and all of them have partial coverage of the spatial or the thematic query restrictions. In the basic system, as in the third query, the results only cover a single theme and they are sorted according to the degree of intersection with the query bounding box. This places the results with spatial over-coverage upper in the result list. For example, the first five results are different versions of the CORINE land cover project about 'Soil Uses' in Spain. The result 17 is the first one about 'Soil Uses' focused on a region close to the query bounding box ('Land cover of Castilla y León'), and the result 19 is the first about 'road networks' ('Topographic Base of Galicia'). The proposed system

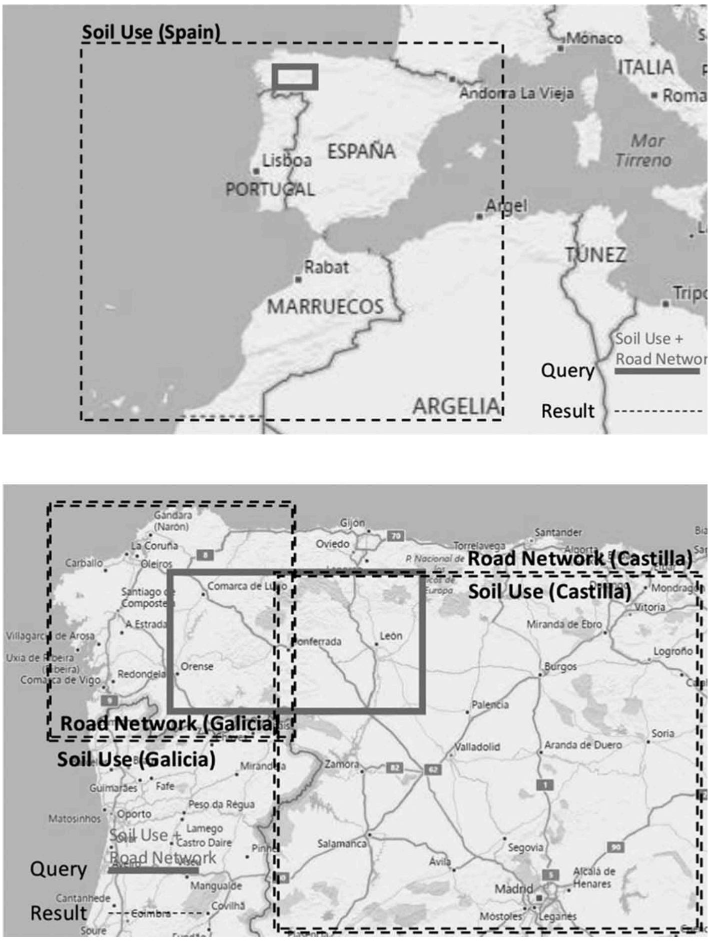


Figure 5. Graphical comparison of the first result of Q4 in both systems.

aggregates resources focused on *Galicia* and *Castilla y León* to form results that almost perfectly fit the user needs. Figure 5 compares graphically the first result obtained in both systems (queries are shown in gray, results in black). The basic system provides an unfocused result about a single query theme (Soil Use) and the query area is only a small fragment of the area covered by the resource. Regarding the proposed system, it returns an aggregation containing four results that provide an almost complete answer to the user query.

As a final comparison between the two systems, Figure 6 shows the DCG of the four selected queries (discounted cumulative gain) at each position of their result lists (number of query results) for the basic and the aggregated IR systems. It can be observed how the proposed system behaves better for all the query types, being especially advantageous in the most general case (Q4). In this case, none of the collection resources perfectly match the query but there are several partial matches. Therefore, the aggregation process can show its maximum potential.

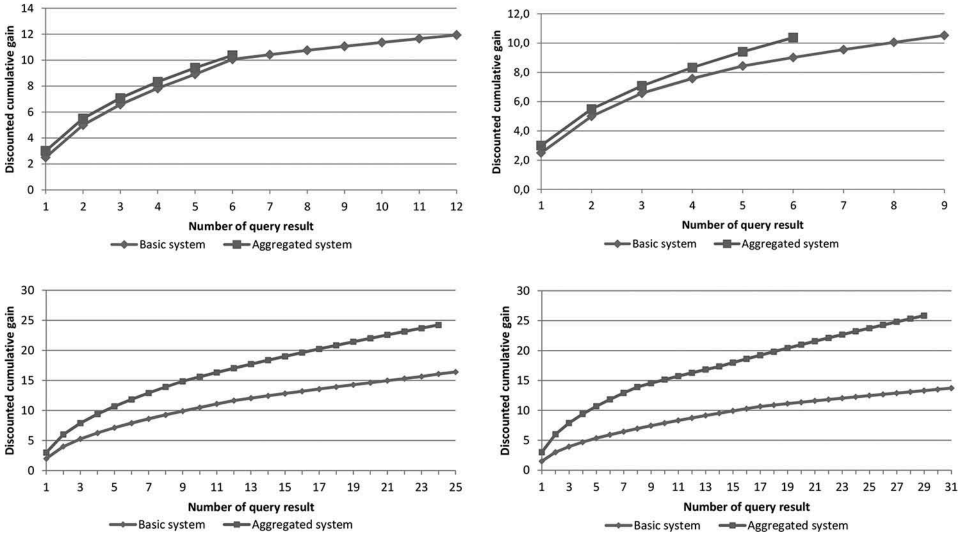


Figure 6. Detailed DCG comparison of the analyzed queries.

6. Discussion

The management of the spatial information as a continuous set is needed for many tasks such as analyzing the morphology of a river or identifying routes in a road network. However, when dealing with geospatial data catalogs, usually continuous information can be found divided into individual resources that do not provide the implicit spatial/thematic connection between them. This problem is not architectural, it is related to how data producers manage information. The technologies used for IR are, in most cases, general purpose solutions not adapted to the nature of spatial data.

The IR system proposed in this paper identifies the spatial and thematic relations in a collection of metadata records of geospatial resources to produce results closer to the query restrictions. To identify the spatial closeness of the records with respect to a query, it uses criteria similar to the one indicated in Lanfear (2006), but using the Hausdorff distance as the spatial similarity measure. The thematic similarity is the ratio between the common themes in the record and the query. Finally, our system integrates the spatial and thematic similarity with a ranking formula like the one detailed in Martins *et al*. (2005). What makes our system very different from them is the addition of a processing layer that identifies the spatial and thematic relations between the results to generate collections of metadata records as query results. The system uses these relations to combine the results into coherent aggregations that are closer to the user query constraints than the individual resources.

In this aspect, the paper has similarities with Lieberman (2006) or Lutz *et al*. (2009) in the sense that we use a layer on top of a basic catalog to provide improved results. The difference is that they use ontologies to generate individual results and we focus on the use of raw metadata to produce aggregations of results. They deal with resources as individuals while our system goes a step beyond that by considering that a result can be a composition of several metadata records.

The aggregation of the catalog metadata records can be seen as a data integration task that helps to improve the quality of a catalog IR process. From this perspective, our work is related to proposals such as Hübner *et al*. (2004) and Lutz and Klien (2006) but working at metadata level instead of at data level.

Our aggregation proposal shows that taking into account the collection context in geospatial data catalogs is a way to provide more complete results. However, we have observed a limitation caused by the difficulty of generating consistent aggregations. In general, the themes related to geographical information are a quite homogeneous terminology set where the aggregations are really meaningful (landforms, infrastructures, cadaster). However, even if two resources share the same thematic, they may not be completely compatible if they provide too different information. For example, a resource containing the geometry of parcels and their type of crop is not very compatible with an aerial thermal image used for crop analysis. The problem has been mitigated using all the keywords of the resources as integration context. However, its effectiveness depends on the content of a single metadata field. A more sophisticated solution would require taking into account additional metadata elements to avoid noise in the aggregations. Additional elements to take into account as factors for data integration would be the data information models, formats, scales, or resolutions.

With respect to the processing time, the step for computing the aggregations does not significantly delay the search process because the spatial operations required to perform the aggregation are restricted to the resources in the result list. Including the time to access the spatial and textual indices, all the queries have returned their results in less than 1 s.

7. Conclusions

This paper has identified and analyzed three issues (under coverage, over coverage, partial coverage) from 'concept at location' style queries in geospatial data catalogs. These issues are caused by the lack of adaptation of prevalent IR engines used in these geospatial data catalogs to the specific nature of geoinformation. As a solution, we have proposed an IR method that yields aggregations of search results that match better 'concept at location' query restrictions.

The proposed IR method takes all the metadata records of resources that partially fulfill a query (intersect the bounding box or the themes) and finds the spatial and thematic relations between them. Next, it uses these relations to generate sets of metadata records that are a better answer to the query than each one individually. To evaluate its performance in archetypical 'concept at location' queries, we have compared the performance of our proposal with respect to an IR system similar to those used in prevalent geospatial data catalogs. The results have shown that this approach may complement the traditional plain list of results of geospatial data catalogs.

Additionally, our proposed aggregation-based functionality could be easily offered in any geospatial data catalog by extending the catalog service for the web (CSW) interface provided by OGC consortium (OGC 2007a). In the CSW, the GetRecords operation is responsible for locating resources according to the user query specified restrictions. CSW standard establishes three levels of detail in query results (*Brief, Summary* and *Full*). To provide interoperability between systems, *Brief* and *Summary* results structure is restricted to the Dublin Core (DCMI 2007) based schema defined by OGC. In the case of *Full* results, the standard allows the definition of profiles that extend the service functionality. Through these profiles, the result structure could be redefined to provide aggregations as a new type of resource that can be returned by the CSW.

Since the structure of the *Brief* and *Summary* levels of detail is restricted, it is not possible to completely describe the aggregations: just the type of returned resource would indicate that is a collection (e.g. using *dct:Collection* as *dc:type* value), and the identifiers of the elements conforming it would be referenced (using *dc:relation* field). However, the *Full* basic description could be extended as needed to indicate the themes and area of the query added by each resource in the aggregation. This approach would make the aggregation-based system compatible with any existent CSW client. In all the three views, a client could obtain the description of resources with standard metadata fields, and only in the *Full* view, it would need to be specifically adapted to process the details of the aggregation composition.

Future work will explore the use of other metadata elements to solve problems related to scale and information content. Clustering the resources that only differ in representation fields, such as the scale, before the indexing process would reduce the heterogeneity of the results, showing the alternative scales and the type of content available in each cluster. Including temporal information can be also used to extend the proposed method for dealing with 'concept at location' at time queries. Finally, once we have aggregated metadata results, it could be possible to produce virtual resources that give access to the associated resources in an integrated way, even if this information was originally scattered across multiple resources.

Notes

1. http://inspire-geoportal.ec.europa.eu/
2. https://www.geoplatform.gov/
3. http://idee.es/
4. https://data.gov.uk
5. https://geodiscover.alberta.ca/geoportal/
6. http://www.europeana.eu/portal/
7. http://www.idee.es/csw-inspire-idee/srv/spa/catalog.search#/home

Acknowledgments

We are grateful to the collaboration of the Spanish National Geographic Institute to provide us with data for the experiments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work of Borja Espejo-Garcia has been partially supported by Aragon Government through the grant number [C38/2015].

ORCID

Javier Lacasta http://orcid.org/0000-0003-3071-5819
F. Javier Lopez-Pellicer http://orcid.org/0000-0001-6491-7430
Javier Nogueras-Iso http://orcid.org/0000-0002-1279-0367
F. Javier Zarazaga-Soria http://orcid.org/0000-0002-6557-2494

References

Anderson, K. and Gaston, K.J., 2013. Lightweight unmanned aerial vehicles will revolutionize spatial ecology. *Frontiers in Ecology and the Environment*, 11, 138–146. doi:10.1890/120150

Asadi, S., *et al*., 2005. Searching the World Wide Web for local services and facilities: a review on the patterns of location-based queries. *In*: W. Fan, Z. Wu and J. Yang, eds. *Advances in web-age information management, lecture notes in computer science*. Cham, Switzerland: Springer, Vol. 3739, 91–101.

Baeza-Yates, R. and Ribeiro-Neto, B., 2011. *Modern information retrieval. The concepts and technologies behind search*. Reading, MA: Addison-Wesley.

DCMI, 2007. *DCMI abstract model*. Dublin Core Metadata Initiative, Technical report. Seoul, Korea: National Library of Korea.

Ferrés, D. and Rodríguez, H., 2015, Evaluating geographical knowledge re-ranking, Linguistic processing and query expansion techniques for geographical information retrieval. *In*: C. Iliopoulos, S. Puglisi and E. Yilmaz, eds. *Proceedings of the 22nd International Symposium, SPIRE 2015*, 9309 of *Lecture Notes in Computer Science*, September, London, UK. New York, USA: Springer-Verlag, 311–323.

Florczyk, A.J., *et al*., 2010. Applying semantic linkage in the geospatial web. *Lecture Notes in Geoinformation and Cartography (LNG&C). Geospatial Thinking*, 201–220.

Göbel, S. and Klein, P., 2002. Ranking mechanisms in metadata information systems for geospatial data. *In*: *Proceedings of the Earth Observation & Geo-Spatial Web and Internet Workshop*, Ispra, Itally. Munich, Germany: Fraunhofer, 13–13.

Graser, A., Asamer, J., and Ponweiser, W., 2015. The elevation factor: Digital elevation model quality and sampling impacts on electric vehicle energy estimation errors. *In*: *Proceedings of the International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)* [online], 3–5 June 2015. Budapest, Hungary: Domokos Esztergár-Kiss, 81–86.

Hübner, S., *et al*., 2004. Ontology-based search for interactive digital maps. *IEEE Intelligent Systems*, 19, 80–86. doi:10.1109/MIS.2004.15

ISO/TC 211, 2014. *ISO 19115-1:2014. Geographic information – Metadata – Part 1: fundamentals*. Geneva, Switzerland: International Organization for Standardization.

ISO/TC 211, 2016. *ISO 19135-3:2016. Metadata – Part 3: XML implementation of fundamentals*. Geneva, Switzerland: International Organization for Standardization.

Janowicz, K., *et al*., 2010. Semantic enablement for spatial data infrastructures. *Transactions in GIS*, 14, 111–129. doi:10.1111/j.1467-9671.2010.01186.x

Kim, J., Vasardani, M., and Winter, S., 2017. Similarity matching for integrating spatial information extracted from place descriptions. *International Journal of Geographical Information Science*, 31, 56–80. doi:10.1080/13658816.2016.1188930

Lanfear, K.J., 2006. *A spatial overlay ranking method for a geospatial search of text objects*. Reston, VA: USGS, Technical report.

Larson, R. and Frontiera, P., 2004. Ranking and representation for geographic information retrieval. *In*: R. Purves and C. Jones, eds. *Proceedings of the SIGIR* 2004 *Workshop on Geographic Information Retrieval* [online], 25–29 July. Sheffield. Available from: http://www.geo.uzh.ch/~rsp/gir/abstracts/

Latre, M., *et al*., 2009. An approach to facilitate the integration of hydrological data by means of ontologies and multilingual thesauri. *In*: M. Sester, L. Bernard and V. Paelke, eds. *Advances in GIScience. Lecture notes in geoinformation and cartography (LNG&C)*, 02–05 June, Hannover, Germany. Berlin, Germany: Springer-Verlag, 155–171.

Lieberman, J., 2006. *Geospatial semantic web interoperability experiment report*. Open Geospatial Consortium, Technical report. Abingdon, UK: Taylor & Francis.

Lutz, M., *et al*., 2009. Overcoming semantic heterogeneity in spatial data infrastructures. *Computers & Geosciences*, 35, 739–752. doi:10.1016/j.cageo.2007.09.017

Lutz, M. and Klien, E., 2006. Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science*, 20, 233–260. doi:10.1080/13658810500287107

Martins, B., Silva, M.J., and Andrade, L., 2005. Indexing and ranking in Geo-IR systems. *In*: C. Jones and R. Purve, eds. *Proceedings of the Workshop on Geographic information retrieval*, 04 November, Bremen, Germany. New York, USA: ACM, 31–34.

Megler, V.M. and Maier, D., 2011. Finding haystacks with needles: ranked search for data using geospatial and temporal characteristics. *In*: J.B. Cushing, J. French and S. Bowers, eds. *Scientific and statistical database management, no. 6809 of lecture notes in computer science*. Berlin, Germany: Springer Verlag, 55–72.

Nogueras-Iso, J., *et al*., 2009. Development and deployment of a services catalog in compliance with the INSPIRE metadata implementing rules. *In*: B. Van Loenen, J.W.J. Besemer and J.A. Zevenbergen, eds. *SDI convergence: research, emerging trends, and critical assessment*. Amersfoort, Netherlands: The Netherlands Geodetic Commission (NGC), 21–34.

OGC, 2007a. *OpenGIS catalogue service implementation specification*. Wayland, MA: Open Geospatial Consoritum, Technical report Version 2.02.

OGC, 2007b. *OpenGIS catalogue services specification 2.0.2 - ISO metadata application profile*. Wayland, MA: Open Geospatial Consoritum, Technical report Version 2.02.

Paneque-Gálvez, J., *et al*., 2014. Small drones for community-based forest monitoring: An assessment of their feasibility and potential in tropical areas. *Forests*, 5, 1481–1507. doi:10.3390/f5061481

Pereira, P., *et al*., 2015. The impact of road and railway embankments on runoff and soil erosion in eastern Spain. *Hydrology and Earth System Sciences Discussions*, 12, 12947–12985. doi:10.5194/hessd-12-12947-2015

Renteria-Agualimpia, W., *et al*., 2016. Improving the geospatial consistency of digital libraries metadata. *Journal of Information Science*, 42, 507–523. doi:10.1177/0165551515597364

Sallaberry, C., 2013. *Geographical information retrieval in textual corpora*. Hoboken, NJ: Wiley.

Smith, T.R., 1996. A digital library for geographically referenced materials. *Computer*, 29, 54–60. doi:10.1109/2.493457

Watters, C. and Amoudi, G., 2003. GeoSearcher: Location-based ranking of searchengine results. *Journal of the American Society for information Science and Technology*, 54, 140–151. doi:10.1002/(ISSN)1532-2890

Wolf, A.T., *et al*., 1999. International river basins of the world. *International Journal of Water Resources Development*, 15, 387–427. doi:10.1080/07900629948682

Zhu, X., *et al*., 2015. Integrating spatial data linkage and analysis services in a geoportal for China urban research. *Transactions in GIS*, 19, 107–128. doi:10.1111/tgis.2015.19.issue-1