International Journal of Geo-Information

MDPI

Article

Approaches for the Clustering of Geographic Metadata and the Automatic Detection of Quasi-Spatial Dataset Series

Javier Lacasta *, Francisco Javier Lopez-Pellicer, Javier Zarazaga-Soria, Rubén Béjar and Javier Nogueras-Iso

Aragón Institute of Engineering Research (I3A), Universidad de Zaragoza, 50018 Zaragoza, Spain; fjlopez@unizar.es (F.J.L.-P.); javy@unizar.es (J.Z.-S.); rbejar@unizar.es (R.B.); jnog@unizar.es (J.N.-I.)
* Correspondence: jlacasta@unizar.es

Abstract: The discrete representation of resources in geospatial catalogues affects their information retrieval performance. The performance could be improved by using automatically generated clusters of related resources, which we name quasi-spatial dataset series. This work evaluates whether a clustering process can create quasi-spatial dataset series using only textual information from metadata elements. We assess the combination of different kinds of text cleaning approaches, word and sentence-embeddings representations (Word2Vec, GloVe, FastText, ELMo, Sentence BERT, and Universal Sentence Encoder), and clustering techniques (K-Means, DBSCAN, OPTICS, and agglomerative clustering) for the task. The results demonstrate that combining word-embeddings representations with an agglomerative-based clustering creates better quasi-spatial dataset series than the other approaches. In addition, we have found that the ELMo representation with agglomerative clustering produces good results without any preprocessing step for text cleaning.

Keywords: geospatial catalogues; metadata; information retrieval; clustering; word embeddings

check for updates

Citation: Lacasta, J.; Lopez-Pellicer, F.J.; Zarazaga-Soria, J.; Béjar, R.; Nogueras-Iso, J. Approaches for the Clustering of Geographic Metadata and the Automatic Detection of Quasi-Spatial Dataset Series. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 87. https://doi.org/10.3390/ijgi11020087

Academic Editor: Wolfgang Kainz

Received: 29 November 2021
Accepted: 25 January 2022
Published: 26 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Geospatial catalogues are discovery and access systems that use metadata as the target for querying geospatial resources [1]. They are typically either directly downloadable datasets (identifiable collections of data) or services for visualising and accessing these datasets. Metadata indicates the purpose, quality, timeliness, location, subjects, and relationships enabling the discovery, evaluation, and application of geospatial resources within and beyond the objectives of the originating data provider [2].

The objective of any catalogue storage system is to make the contained resources findable, accessible, interoperable, and reusable, which is commonly known as the FAIR principle [3]. With respect to findability, the prevalent approach for searching geospatial data in geospatial catalogues is the "concept at location in time" query [4]. That is, users expect that geospatial catalogues will return information based on their conceptual, spatial, and temporal relevance with respect to a query. This approach is natural, but it is known to be ineffective in the real world without improving the different geospatial catalogue components with intelligent metadata curation methods or the use of advanced search engines [5]. There are multiple works in the literature proposing search improvements in the fulfilment of FAIR principles through adding semantics and ontologies to the metadata, using new ranking algorithms, or boosting data storage [6–9].

However, none of these proposals addresses the mismatch between the continuous nature of geospatial information and the discrete nature of data production. When a query is submitted to a geospatial catalogue about a concept covering a wide spatial extent, likely none of the retrieved resources will cover the entire extent (most of the results will only cover small parts of this extent). For example, an analyst could search for data about the hydrological behaviour in a given mountain range in a catalogue. However, river basin

datasets usually cover a single basin, since each basin is separated topographically from adjacent ones by a mountain range forming a drainage divide. Therefore, a query for rivers (concept) covering the mountain range (location) will return a collection of datasets describing all the river basins where the mountain range acts as drainage divide among many others containing the concepts of river and mountain range. If there is no single resource containing all the requested information, the analyst is forced to review all search results to locate those that contain relevant information and merge their content.

We think that one of the main sources of this problem is the lack of alignment between user needs and data producer objectives. Data producers create resources based on their areas of responsibility. Users query catalogues based on their own areas of interest. These areas of interest may be defined by themes covering a spatial continuum from the user's point of view, which does not necessarily match with the areas of responsibility of data producers. Therefore, geospatial catalogues may often return resources that partially cover the query area without contextual information useful to discover sets of results that, seen as a collection, cover the whole query area. We refer to this problem as data fragmentation in the results. This problem can be solved by identifying the set of resources that conceptually belong to the same thematic layer. From the provider side, a solution for this challenge is the use of spatial dataset series, which are collections of spatial datasets that share similar properties of theme, scale, or purpose [10]. However, from the point of view of users, they are not enough since the relevant data of an area could be scattered across different spatial dataset series from different data providers. Therefore, a more general solution is required to identify thematic relationships in the resources of a geospatial catalogue, create virtual spatial dataset series from them, and return them as part of related query result sets.

Objectives and Contributions

Our proposal to deal with data fragmentation is to change the way in which results are presented in geospatial catalogues. The described information retrieval (IR) problem is not specifically related to the selection of an IR algorithm, since all resources can be considered as partially relevant, but about how the information fragments (individual resources) are presented to the users. Instead of a list of individual results covering parts of the requested area, we think that it is better if these results are grouped by spatial compatibility with respect to the spatial extent specified in the user query. That is, compatible resources that provide jointly a better answer to the user query should be shown as a set.

To generate these collections, we propose to cluster the metadata records to identify sets of similar resources from different providers that describe the same theme. Due to their heterogeneous creation, these resources may have a different format, resolution, or data granularity. However, as they have the same theme and their union covers wider areas than each resource individually, these sets can be perceived by users as a valid response to their search. We can name these sets as quasi-spatial dataset series because they can be described as virtual collections of spatial datasets that share some features attributed to data series. That is, these sets are collections of spatial datasets with a close product specification. These series aggregate, by similarity, the resources that are compatible, preventing the user from needing to do this analysis. These collections will probably contain resources with different resolution, overlapping areas, or different temporal extent, but, from the point of view of the user, they aggregate resources that can be seen as a whole in a similar way to a dataset series. In the case of dataset series, their homogeneity makes their integration direct, while the proposed quasi-spatial dataset series would require harmonisation of their content. We do not perform data integration in this work, but it would be the next natural step of the proposal presented in this paper. In this way, the user could obtain the available information in the defined quasi-spatial dataset series homogeneously.

This work evaluates whether state-of-the-art clustering processes can efficiently aggregate spatial resources into quasi-spatial dataset series, using only textual information from elements in their associated metadata records. To identify which clustering process is best suited for this task, we compare the performance obtained using different data clean-

ing, feature representation models, and clustering algorithms. The evaluation has been performed with a collection of 630 metadata records obtained from a catalogue published at IDEE (the National Spatial Data Infrastructure of Spain), a leading national spatial data infrastructure in Europe. These records, compliant with the ISO 19115 geographic metadata standard [2], contain descriptive textual information on a range of themes such as cadastre, environment, and infrastructures.

The contributions of this paper are focused on two areas: the study of the IR problem of current geospatial catalogues and the comparison of different clustering alternatives that can reduce this problem. It addresses the following research questions:

RQ1 What causes the ineffectiveness of current geospatial catalogue IR systems, and how can they be improved? To answer this question, we analyse the current state of geospatial catalogues and describe the IR problems related with the dissonance between the continuous nature of geospatial information and the digital library-based structure of these metadata catalogues. As a solution to reduce these IR issues, we propose the generation of collections of related resources, which we call quasi-spatial dataset series, defined to improve the display of query results.

RQ2 Can current clustering techniques generate good quality quasi-spatial dataset series? Here, we have established as baseline a collection of metadata records with manually tagged quasi-spatial dataset series. Then, we have performed experiments with multiple clustering process configurations to determine if they could automatically identify the collections. We perform different kinds of cleaning of the source data and compare the results using the classic TF-IDF feature representation with respect to modern embeddings (Word2Vec, GloVe, FastText, ELMo, Sentence BERT, and Universal Sentence Encoder). As clustering algorithms, we have compared K-Means, DBSCAN, OPTICS, and Agglomerative clustering.

RQ3 Which clustering processes are the most suitable for this task? The different processes performed are compared with respect to the manually tagged collections using V-measure and Adjusted-Mutual-Information. Apart from identifying the best performing configurations, we also search for general solutions (those that do not preprocess the input text in any way) to determine if they are good enough to be used by a catalogue.

The paper is organised as follows: Section 2 describes the state-of-the-art in clustering techniques to generate the desired quasi-spatial dataset series. Section 3 introduces the problems that produce data fragmentation in geospatial catalogues. Section 4 explains the characteristics of the clustering processes used in the experiments. Then, Section 5 presents the dataset and the experimental setup used for the experiments, and Section 6 compares the results obtained with respect to the selected reference collection. The paper finishes with a discussion about the results, conclusions, and an outlook on future work.

2. Related Work

In the geospatial catalogue context, there have been numerous works trying to improve search processes in different ways. Larson and Frontiera [11] make a comparison of several ranking algorithms for georeferenced objects including simple, topological, and extent overlap, then propose a probabilistic spatial ranking based on logistic regression that uses the area of the overlap as the main similarity factor. Zhan et al. [12] propose a semantic description model for geographic information able to deal with heterogeneity problems in descriptions by using ontologies. The proposal is focused on allowing the user to express the meaning of their queries so the results obtained are improved. Zhang et al. [13] show an approach to extract geospatial data from multiple sources, model it as RDF to eliminate heterogeneity, and link it using a semantic matching algorithm. de Andrade et al. [14] remark the limitations that make it difficult to find geospatial data in current geospatial catalogues. Some of the identified problems are the use of a single record to describe the feature types in a service, the lack of semantics in the descriptions, and the lack of a

suitable ranking to organise the results in a query. They propose a framework with ranking metrics to improve spatial, semantic, temporal, and multidimensional queries. Li et al. [15] describe an information retrieval process for geospatial catalogues that uses semantic latent analysis to improve effectiveness of the search engines. This enables the discovery of the semantics between word patterns that allows the identification of relevant resources not directly containing the query terms. Fugazza et al. [16] and Fugazza et al. [17] propose a methodology to add semantic features to metadata that allows metadata delegation and facilitates the identification of relations and simplifies their evolution management. Miao et al. [18] show how to improve the effectiveness of geospatial data discovery using a measurement model of spatio-temporal similarity. Finally, Li et al. [19] describe a deep-learning solution to improve search ranking of geospatial data using logs of previous user interactions in the catalogue. They model the relevance of data according to user interaction and use a deep learning ranking model to determine the order of the results for the queries. They propose a similarity measure that uses the maximum semantic distance between any pair of nodes in the ontology used for matching and the weighted distance from the lowest common ancestor node to the root node.

In the field of digital libraries, clustering has been frequently used to generate sets of related resources that facilitate search and browsing. Aggarwal and Zhai [20] make a detailed compilation of the classic clustering techniques for organisation, browsing, summarisation, and classification of documents. Metadata records can be treated as short documents in which the description plays the role of document content, and, therefore, can be clustered according to their similarity.

A basic aspect of clustering is feature representation. Document Frequency, Latent Semantic Indexing, and Non Negative Matrix Factorisation are classic solutions for this task [20]. Word embeddings is a recent word representation also suitable for clustering [21,22]. It maps words to a multidimensional vector space model so that semantically similar/related words tend to be close in that space. There are multiple neural network architectures able to generate these embeddings. Word2Vec [23], GloVe [24], or FastText [25], ELMo [26], BERT [27], or GPT-3 [28] are among the most popular ones. They have evolved from context independent architectures to context dependent ones that produce better results for understanding the semantics of the words. Since word embeddings are word representations, to represent text sentences, Arora et al. [29] propose different means of the word embeddings of words in a sentence. Sentence embeddings is the evolution of word embeddings to encode complete sentences into vector representations. They have the advantage of obtaining a sentence representation directly without having to consider each word independently. The most popular architectures are Doc2Vec [30], Sentence BERT [31], InferSent [32], and the Universal Sentence Encoder [33]. An example of the use of embeddings in clustering is Kusner et al. [34], who use the minimum distance between their document embeddings as a distance metric. Similarly, Zhang et al. [35] describe the generation of classification taxonomies from documents using word embeddings of the document content. They define an embedding module that learns discriminative embeddings at the different levels of the taxonomy. Hu et al. [36] analyse the evolution of topics in scientific papers through their representation as Word2Vec embeddings and measuring their spatial autocorrelation in the embeddings space. They measure how the popularity of some keywords affects the surrounding ones. Diaz et al. [37] propose the embedding of spatio-temporal textual data in a representation that allows identifying patterns associated with time and location of human activities described as text. Their model allows suggesting periods or locations linked to a sentence and vice versa. Li et al. [38] perform text clustering using Sentence BERT as encoding of the text sentences, a weighting layer to increase the relevance of sentences as a function of the named entities contained, and K-means as the clustering algorithm. Arenas-Márquez et al. [39] describes the use of a convolutional neural network to identify topics of interest in a collection of TripAdvisor messages using Word2Vec embeddings of the words in the documents as input. They compare this approach with respect to Latent Dirichlet Allocation encoding of the texts and Word2Vec mean.

Multiple clustering algorithms can use these feature representations. The most used algorithms are distance-based solutions such as K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), or Ordering Points To Identify the Clustering Structure (OPTICS), and probabilistic ones like Probabilistic Latent Semantic Indexing (PLSI) [20]. The work of Zola et al. [40] is a good example of how some of these clustering techniques are currently used in the spatial data context to identify patterns in text collections. They estimate Twitter user location based on their tweets using Google Trends frequencies of tweet nouns and clustering to identify the most probable location. Newman et al. [41] show how statistical topic models classify thematically a collection of metadata records and provide faceted search. Lacasta et al. [42] describe a clustering process for metadata that uses the hierarchical structure of the concepts contained in Knowledge Organisation Systems (KOS) to improve the clustering results. Thomas and Khan [43] propose a clustering process for documents that uses the metadata information associated with each document to improve the quality of the clusters. Rajan et al. [44] depict a clustering process to aggregate patent descriptions into similar groups to facilitate the search process in patent databases. Rakib et al. [45] propose an iterative classification method that improves the clustering of short texts. This is done by detecting outliers during the clustering process and changing the clusters to which they are assigned. They apply this improvement to different K-means and hierarchical agglomerative clustering variants to determine the applicability to multiple clustering algorithms. Cai et al. [46] propose the Adaptive DBSCAN clustering algorithm, a DBSCAN variant to deal with issues related to linear connections between objective clusters and parametrisation complexity. It uses a data splitter and merger coordinated in local and global clustering steps. This allows dynamically discovering clusters from local to global. Lou et al. [47] analyse the evolution of research methods in the Chinese information science community. Multiple features, such as publication time, researcher age, novelty, or paper diversity, are taken into account for the analysis. To identify similarities by period, theme, or researcher, they cluster the works in the analysed collection using Euclidean distance similarity, Partitioning Around Medoids, and K-means. Misztal-Radecka and Indurkhya [48] describes a Bias-Aware Hierarchical Clustering algorithm to improve recommendation systems by identifying clusters of users with unsuitable recommendations. It is a variation of K-means where splitting depends on high biases instead of minimum variance. They compare this solution with respect to other K-means variants, agglomerative clustering, Hierarchical BDSCAN, and Local Outlier Factor between others.

The work presented in this paper is similar to the previously described works that use clustering techniques to identify similarities in digital library collections. However, in our case, we search clusters in the metadata descriptions that can be classified as quasi-spatial dataset series, which puts limits on the way the clustering process is performed. To analyse the suitability of the different techniques and models, we compare a set of classical and modern clustering techniques. This includes different data-cleaning processes, feature representation models, and parametrisation of clustering algorithms.

3. Geospatial Catalogues and the Spatial Data Continuum

Geospatial catalogues are repositories of spatial resources defined by multiple providers and described through metadata. Data providers focus on specific areas because of legal obligations, economic limitations, and changing objectives over time.

These catalogues are technologically similar to digital libraries as they manage their content as any other discrete digital resource (e.g., a photo, a book, or a video). However, the spatial dimension makes geospatial catalogue content to be a patchwork of regions over the earth's surface about heterogeneous themes. Ahmad and Ali [49] show a comprehensive compilation of services with 153 active catalogues providing spatial data all around the globe that follow this approach. Among them, some relevant examples are the pan-European INSPIRE catalogue (https://inspire-geoportal.ec.europa.eu/, accessed on 26 November 2021) and the national catalogues of the USA (GeoPlatform) (https://www.

geoplatform.gov/, accessed on 26 November 2021), Spain (IDEE) (https://www.idee.es/es, accessed on 26 November 2021), United Kingdom (Data.Gov) (https://data.gov.uk, accessed on 26 November 2021), and Canada (GeoDiscovery) (https://geodiscover.alberta.ca/geoporta, accessed on 26 November 2021).

These catalogues provide a simple solution to publish resources, but the way they present results limits their usability. Geospatial information forms a continuum around Earth characterised by the spatial location (point, line, or polygon), and the theme, which is a conceptual abstraction of the nature/purpose of the represented data. Even discrete geospatial types, such as tree locations, rivers, or streets, are part of a bigger set covering all the earth's surface (e.g., all the trees, rivers, or streets on the earth). Any division of this continuum is artificial and makes data management more complex, since the continuum has to be reconstructed to obtain the information distributed in multiple fragments. This indirectly downgrades the performance of any search system using this approach as partial data results are presented as complete results. This makes results of "concept at location in time" queries incomplete because, in most cases, the area queried by users does not fit the arbitrary partition of the spatial data. It is as if each spatial resource were a "book page" whose author, page title, date of creation, or publisher contained in the metadata could help to decide which "book page" fits better the user needs, even though the needed information may be found along all the "book". Spatial data fragmentation increases existing challenges regarding metadata generation, update, and improvement [50] and makes it difficult to maintain complete, up-to-date, and useful metadata. It causes heterogeneity and lack of harmonisation in descriptions, even in versions of the same resource, which is one of the causes of their poor performance [5]. Finally, since the provided results are only partially relevant, they are difficult to present in a suitable way for the users. A sequential list of results is confusing when the provided results are only fragments of the data in a given theme.

Figure 1 illustrates some of the problems of spatial fragmentation in a simplified way. It shows the coverage of LIDAR resources in the south of Spain from different providers (Provincial Councils of Málaga, Cádiz, and Huelva). They contain equivalent content, but there is no connection between them. In the current geospatial catalogues, a query covering all the south of Spain will return a list containing the three results (among others) because they partially cover the user needs. Then, the user will have to review manually the entire result list to identify the items that cover his needs. This may seem simple, but if there are hundreds of resources with similar issues, finding those that are related can be time-consuming. For example, a query about Laser Imaging (LIDAR) in the Spanish catalogue provides 305 results about themes such as land cover, forestry, water, or coast information, among others. They are presented without any order that could simplify the identification of those that are implicitly related.

Figure 1. Coverage of resources with LIDAR points information in the south of Spain.

The problem here is that the spatial needs of the user do not fit with the classic digital library-based organisation. Due to the spatial data fragmentation, a resource covering the user-required data may not even exist. In this context, to provide good search capabilities and to improve user satisfaction, we should evolve geospatial catalogues from IR systems for data producers dealing with self-contained metadata records with spatial properties to IR systems for data consumers dealing with continuous content layers. Hennig and Belgui [51] and Masó et al. [52] already highlighted the need to make user-centric SDIs instead of focusing on products or processes. Specifically, they describe the need to improve metadata descriptions in geospatial catalogues to focus on the user needs and to avoid the disconnection between data and metadata descriptions.

Applications such as Google Maps (https://www.google.com/, accessed on 26 November 2021) or Open Street Map (https://www.openstreetmap.org/, accessed on 26 November 2021) show that continuous layers of spatial information improve the user experience in some scenarios. They provide seamless layers of information for a few data types, such as cartography, roads, and commercial business, so that users can directly select/visualise/copy information in any part of the globe. This simplifies the search process, and, independently of the area requested, all the information is in a single resource with the same format and quality.

Defining such continuous layers is not currently viable. In addition to the huge cost of cleaning, harmonisation, and integration of existing data, the manual management and update of resources created by multiple providers with different interests would be extremely difficult. An alternative to this manual work would be the development of a process for the automatic identification of thematically-compatible resources so that they can be presented as a set in the result lists. The identifiable sets of thematically-related resources would not be continuous layers, but it could be the closest possible representation with the available metadata. In that form, it would be simpler for the user to get all the resources needed for answering his or her query. This idea comes from the concept of spatial dataset series. When a provider creates a spatial dataset series from a set of uniform and similar resources, the user can manage them as a single resource. The kind of sets we want to identify can be named as quasi-spatial dataset series because, as previously indicated, they are collections of spatial datasets with a close product specification that can help IR systems to provide results that are more compact.

The improved IR process for geospatial catalogues using these quasi-spatial dataset series is shown in Figure 2. All algorithms and methods used in the search process of current geospatial catalogues return a ranked list of results. Our proposal is to identify the relations between the datasets (the quasi-spatial dataset series) and use them to cluster the result list in the query post-processing step. The task consists of grouping those resources in the result list that are part of the same quasi-spatial dataset series and placing them in the best ranked positions of the result set. Table 1 shows how this organisational change improves the result list. Following the previous LIDAR example, the table shows a selected subset of the 305 results of a query with the LIDAR term submitted to the Spanish geospatial catalogue according to our definition of quasi-spatial dataset series. The results have been simplified for illustrative purposes to remark how a clustered list of results shows relations that would be hidden if the list were not organised. A few products occur several times in the result list for the same type of data in different areas, such as LIDAR points for administrative divisions, Photogrammetric-LIDAR for river basins, or digital elevation models of rivers and coasts. The shown groups have similarity in their titles, but, in many cases, this is not enough, as the description may show that their content is too different (e.g., Cloud of points of Cerro Muriano fire and Guadalete-Barbate river basin), or their titles may differ even if their description is similar. It is also important to note that the shown clusters are only partially compatible. They are from different years, and, if observed deeply, they may have different formats, resolution, or other incompatible technical aspects. However, from the user perspective, knowing easily which types of resources are available

is a relevant improvement, as the integration part can be done by him on the final selected subset that fulfils his needs.

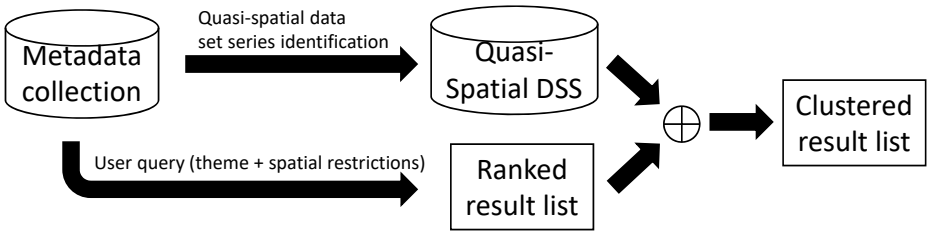

Figure 2. IR process using quasi-spatial dataset series.

Table 1. Example of lists of clustered results (quasi-spatial dataset series).

Cluster	List of Clustered Results
1	LIDAR 2016. 0.5 points per square meter Alicante province LIDAR 2015. 0.5 points per square meter Valencia province LIDAR 2009. 0.5 points per square meter Valencia autonomous region...
2	Cloud of points of Spanish PNOA-LIDAR
3	Cloud of LIDAR points of Cerro Muriano forest fire (Córdoba), 2007
4	LIDAR second Coverage (2015-Today) LIDAR first Coverage (2008–2015)...
5	Photogrammetric-LIDAR data from Guadalete-Barbate river basin (Cádiz), 2008 Photogrammetric-LIDAR data from Guadalhorce-Guadiaro (Cádiz), 2008...
6	Cloud of LIDAR points Guadalete-Barbate river basin, 2008...
7	Digital Elevation Model Guadalhorce-Guadiaro river basin (Cádiz), 2008 Digital Elevation Model Granada coast, 2006 Digital Elevation Model Oriental and Occidental Málaga coast, 2007...
8	Terrain Elevation Model from LIDAR. Resolution 2 meters. 2017...

The identification of these quasi-spatial dataset series is not an easy task because available resources are not evenly distributed and have different characteristics. The descriptions in their metadata contain technical domain terminology, such as scales, resolutions, or formats; textual place names that complement numeric spatial bounding boxes; and various information about the multiple themes of the described data (e.g., agriculture, environment, pollution, or cadastre).

The literature has some works in this field. For instance, Lacasta et al. [53] describe an IR process for geospatial data catalogues that focuses on solving this fragmentation problem by identifying the implicit spatial/thematic relations among query results. Their process focuses on finding resources spatially and thematically compatible with the user query and identifying their theme and spatial overlap. Result sets constructed this way fulfil user queries better than each resource individually (they cover a bigger part of the required area for the required keywords). However, the need to construct dynamically the aggregated result sets from each performed query and the complexity of selecting thematically compatible results complicate its application. Previously, Latre et al. [54] proposed a process for the integration of hydrological data by merging the ontologies that represent their models. This process allows providing a unified view of fragmented data collections at the cost of creating complex ontologies that describe the data.

We think that clustering is a suitable approach for data aggregation tasks such as the one proposed in this paper. However, geospatial data have features that make this process difficult. Firstly, the resources must be aggregated by thematic similarity and not by other dimensions such as location, format, or resolution between others. Additionally, the number of sets and their dimensions may be heterogeneous, and many of the resources may

not have any thematic relation with the rest (they are independent). Solving these issues is possible, but it requires processes adapted to the analysed data that may not be generalised to other collections. Due to these considerations, our objective has been to identify not only the best clustering solution but also the best between those without data preprocessing.

4. Evaluation Framework

To identify quasi-spatial dataset series in a geospatial catalogue, we cluster the textual information in the contained metadata records. Figure 3 shows the details of this process. It performs the cleaning of the selected metadata properties, transforms them into features, and clusters them into quasi-spatial dataset series. For each step, we have compared different solutions used in the literature. The developed pipeline contains the classic processes to remove undesired elements that affect the results. However, since cleaning steps are specific to the processed data, we have also tested process configurations without any cleaning.

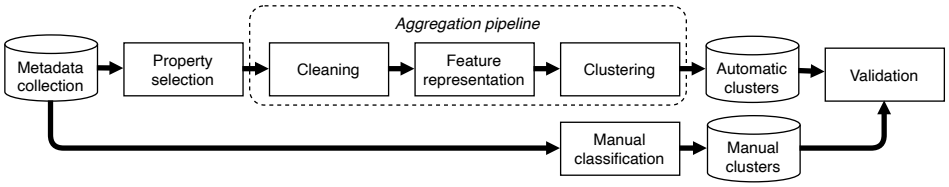

Figure 3. Clustering pipeline.

4.1. Property Selection

The initial step is the selection of the metadata properties to process.

We have decided to focus on the properties that act as title and abstract because they are the main metadata elements filled with free text in metadata records. The use of keyword elements, although they refer to concepts, is not a viable alternative because they contain just one or two words, whose TF-IDF or word-embeddings representation refer to general concepts and would probably generate big and heterogeneous clusters (there may be thousands of datasets classified as "land cover"). Given a metadata schema, the title property holds the distinguishing name of the resource and may convey a minimal summary of its contents, whereas the abstract property describes the contents of the resource in a more detailed way. In general, the more information available for a clustering technique, the better it can identify the similarity of resources. However, we do not want a clustering of maximum similarity, as we want to avoid clustering by location, scale, or other aspects described in the metadata records that are not the theme. In this context, adding more textual content may lead to incorrect aggregations (i.e., clusters of data about the same place but different theme). To assess the potential impact of this possibility, we have evaluated three scenarios: using only the title property as input of the clustering, using only the abstract property as input, and using both properties as input.

4.2. Cleaning

The selected text is tokenised into words (tokens), and tokens that may negatively affect the clustering results are removed. For this task, we have evaluated a set of basic normalisation and cleaning processes from the literature to increase the uniformity of tokens [55]. Specifically, we have tested all possible combinations of the following cleaning processes: conversion to lowercase, removal of stop words, removal of place names, removal of text within parentheses, and reduction of word forms to stems. The removal of stop words and place names is performed thanks to the use of word lists. Regular expressions are used to remove text within parentheses. Finally, we use the Snowball algorithm for stemming [56].

4.3. Feature Representation

The next step transforms the cleaned tokens of the metadata record into features that will be the input of the clustering algorithms. As feature representation, we have compared word embeddings, sentence embeddings, and the classic TF-IDF matrix representation as baseline.

TF-IDF feature representation is a document-term matrix where each position (d, t) contains the frequency of a term t in a metadata record d multiplied by the inverse document frequency of the term t in the collection D. From the different TF-IDF variants, we use the one shown in Equation (1). It measures the relevance of a term contained in a document. The term frequency takes into account the number of occurrences of the term in the document, and the inverse document frequency depicts how rare and informative the term is in the collection to reduce the TF-IDF value of common terms. The term frequency of a term t in a metadata record d is the number of occurrences of the term $tf_{t,d}$ divided by the total number of terms in the metadata record ($size(d)$). The inverse term frequency of a term t in the collection is the logarithm of the number of metadata records in the collection (N) divided by the number of metadata records containing t in the collection (df_t).

$$tf.idf_{t,d} = \frac{tf_{t,d}}{size(d)} * log\frac{N}{df_t} \tag{1}$$

Word embeddings represent words as a multidimensional vector space model in such a way that semantically similar/related words are represented as close points in that space. There are multiple implementations of word embeddings depending on the used neural network architecture and training data. We cannot use these word embeddings directly as we need to compare the similarity of complete sentences to determine if they are about the same theme. Therefore, we transform them into a sentence representation through summarisation. For this transformation, we have compared the use of the word-embeddings mean and word-embeddings weighted mean in each sentence as indicated in Arora et al. [29]. Word-embeddings mean uses the mean of the different embeddings of each document as document representation. Word-embeddings weighted mean uses TF-IDF to adjust the weight of each embedding. The sentence representation of a metadata record (d) using a word-embeddings mean ($\vec{se}(d)$) is computed as the sum of the word-embeddings representation of each different term in the metadata record ($\vec{we}(t_i)$) divided by the number of different terms in the metadata record ($size(distinct(t_i \in d))$) (see Equation (2)). The weighted mean sentence representation objective ($s\vec{we}(d)$) is to correct the difference in frequency of the words in the collection, so common terms weigh less in the mean than uncommon ones. It is calculated as in the previous equation but with multiplying the word-embeddings representation of each different term in the metadata records by the TF-IDF of such term in the collection (see Equation (3)). We have also tested pure sentence embeddings as feature representation. These systems directly represent sentences as a multidimensional vector space model, avoiding the need of summarisation.

$$\vec{se}(d) = \frac{\sum_{distinct(t_i \in d)} \vec{we}_{(t_i)}}{size(distinct(t_i \in d))} \tag{2}$$

$$s\vec{we}(d) = \frac{\sum_{distinct(t_i \in d)} (\vec{we}(t_i) * tf.idf_{t_i,d})}{size(distinct(t_i \in d))} \tag{3}$$

Specifically, we have tested the following embeddings generated with Spanish text collections (the language of our experiment data). As word embeddings, we have used: Word2Vec [57], GLoVe [24], FastText [25] generated with the text collection proposed by Cardellino [57], and ELMo multilingual embeddings [58]. As sentence embeddings, we have used: Sentence BERT [31] and Universal Sentence Encoder [33].

4.4. Clustering

Given the federated nature of geospatial catalogues, they include data from national to local governments and organisations. Upper-level governments and organisations publish datasets covering large areas under their jurisdiction. These, in turn, are often subdivided into smaller units that also publish data covering areas under their jurisdiction. In many cases, but not always, datasets from smaller units can be aggregated to form a quasi-spatial dataset series. That is, they can be grouped in clusters. For example, it is possible to create an address gazetteer by hand for a specific purpose by aggregating addresses published by local governments. This is not the case of upper-level datasets, as they are whole sets. That is, they are one-element clusters. For example, national address gazetteers are datasets created by upper-level entities that bring together address information from all authorities from a country.

This characteristic is an issue for many of the classic clustering algorithms, as they usually are not able to identify one-element clusters. We have compared K-means [59], DBSCAN [60], OPTICS [61], and agglomerative clustering. K-means, DBSCAN, and OPTICS are some of the most-used clustering techniques in the literature, but they have some drawbacks with one-element clusters. K-means generates one-element clusters, but it requires selecting the desired number of clusters manually, and finding it in each collection requires much experimentation. DBSCAN can generate them depending on the configuration, but their hyper-parameters are difficult to adjust. OPTICS is quite stable, but it is unable to generate one-element clusters. It assigns isolated elements to other clusters or marks them as spurious data. Finally, we have tested an agglomerative clustering process that directly allows the generation of one-element clusters. This process is a simplification of an agglomerative clustering algorithm [62] that stops constructing the cluster tree when the similarity between all the elements of different clusters is less than a given threshold. It calculates the similarity between two metadata records using the cosine distance and aggregates the pair with the highest similarity value, that is, the dot product between the vector representations of the metadata records divided by the product of their norms. The process is repeated until the highest similarity found is less than the selected value.

In these clustering algorithms, the hyper-parameters related to the minimum cluster size have been selected to be the minimum possible to facilitate the identification of metadata records with no relations or small clusters. The values of the remaining parameters have been obtained through a value sweep done with the experiment data. The distance between samples in DBSCAN has been set to 1.05 for Euclidean distance and 0.09 for Cosine distance since other values increased the number of heterogeneous clusters or the division of uniform ones. Similarly, in the agglomerative solution, the selected similarity value to identify if two resources are in the same cluster has been set to 0.98. Bigger values divided too many uniform sets and lower ones created additional heterogeneous ones. The parametrisation has been done to compare the best configuration of the different algorithms, but our experiments have shown that the differences with the default configuration of the algorithms are not big, and they could have been directly used at a small performance cost.

4.5. Validation of Results

We have compared the results of each approach with respect to a manual classification of the experiment data done by a board of 5 experts in Spatial Data Infrastructures. The result quality measures used are the V-Measure score [63] and the Adjusted-Mutual-Information (AMI) [64]. V-Measure calculates the harmonic mean between homogeneity and completeness of the clusters. A cluster is homogeneous if it only contains members of a single class and complete if all the members of the class are in the cluster. Exact partitions are both homogeneous and complete, having a score of 1. With respect to AMI, it quantifies the mutual dependence between two sets of clusters according to the information they share. That is, it measures how one of the sets of clusters allows knowing about the other. The mutual information between two partitions is calculated as the sum of the probabilities that each collection resource has of belonging to any pair of clusters multiplied by the

logarithm of the observed/expected ratio of belonging to the clusters. This metric is then adjusted to take values between 1 for complete similarity and 0 for complete dissimilarity. In both cases, we calculate the measures comparing the clusters generated in each experiment with respect to the clusters created in the manual classification.

5. Dataset Description

The National Spatial Data Infrastructure of Spain (IDEE) is the official body that coordinates the cooperation of Spatial Data Infrastructures launched by public administrations at national, regional, and local level. In 2021, it integrates the collaboration of the governments of 19 autonomous communities, 14 national agencies, and 39 city councils (https://www.idee.es/resources/documentos/Responsables_nodos_IDE.pdf, accessed on 26 November 2021). Through the IDEE geoportal (name given to the portal of this type of infrastructure), it is possible to access thousands of resources (datasets and services) about a myriad of themes with coverage that range from the whole country to a municipality.

We have selected this collection because it contains a complete collection of the geospatial resources published in Spain. Specifically, this collection has been curated through a harvesting process that retrieves the contents of the catalogues running at the different SDI initiatives that belong to either national governmental offices or regional governments.

We downloaded 4824 metadata records describing these resources, but not all of them were suitable for the analysis in this paper. For comparison between the previously described processes, we have selected a subset of 630 metadata records describing datasets in this infrastructure. These records, compliant with the ISO 19115 geographic metadata standard [2], contain descriptive textual information on a range of themes such as cadastre, environment, and infrastructures. They have been selected because they are all in Spanish (there are many other records using different Spanish official languages), and none of them is tagged as part of an explicit series. In this subset, many resources cover small areas about equivalent themes, but, since they have been created by different providers, they do not have any explicit relation in their metadata. That is, it contains many resource sets that can be organised as quasi-spatial dataset series, making it very appropriate for the comparison of algorithms trying to identify such series. Additionally, the size of the collection allows its manual analysis to provide a baseline for the results comparison.

Table 2 shows some relevant features of the title and abstract properties of the 630 metadata records used in the experiments. The mean words per field and the standard deviation show that most of the analysed text values are short. Although the longest abstract contains 712 words, the majority have less than 250 words, and a relevant set less than 10 words.

Table 2. Relevant features of the datasets used for the experiments.

Property	Total Words	Mean Words	Min Words	Max Words	Std Dev
Title	8010	12.71	1	33	8.52
Abstract	87,482	138.86	6	712	119.24

Figure 4 shows an example of an original metadata record in XML format and translated into English. As it is shown, it is usual that such descriptions contain thematic information about the nature or purpose of the data combined with other factors such as formats, spatial references, and other technical characteristics of the data. Many of the terms frequently occurring in these descriptions are common but have no relevance for the desired thematic aggregation, so they can cause the generation of undesired clusterings.

The performance of each experiment has been evaluated with respect to a manual classification performed by a board of experts that has identified 80 quasi-spatial dataset series. The biggest manual cluster contains 119 elements, and there are 111 metadata records with no relations. This classification was done in a two-step process. First, the metadata of the resources was manually reviewed and grouped by similarity in their description (their content is equivalent). Then, the data on each identified cluster has been

visualised to determine how the contained resources are spatially distributed. The decision about the correctness of the identified sets has been done by consensus of the board of experts. We have identified that the main themes of the resources are related to nature protection (vegetation, soil erosion, floods, climate), agricultural activities (crops, cattle, forestry, hydrography, dams, irrigation), industry (distribution, pollution), and political organisation (administrative divisions, land uses). Some of the identified clusters are a collection of meteorological images from the LINDE satellite covering different Spanish municipalities, a compilation of pressures that produce the arrival of marine trash in each different Spanish sea demarcation, or a cluster with the areas of flood risk in rivers and coasts.

```
<gmd:MD_Metadata xmlns:gmd="http://www.isotc211.org/2005/gmd">
 <gmd:identificationInfo>
  <gmd:MD_DataIdentification>
   <gmd:citation>
    <gmd:CI_Citation>
     <gmd:title>
      <gco:CharacterString>Orthophoto of 2005 of the province of Castellón in RGBI and
      Alicante province RGB of 50cm. resolution</gco:CharacterString>
     </gmd:title>
    </gmd:CI_Citation>
   </gmd:citation>
   <gmd:abstract>
    <gco:CharacterString> Orthophoto of the Valencian Community, province of Castellón
    corresponding to {the} year 2005. Flight carried out between September
    - October 2005. With a pixel size of 50 centimeters and distributed by MTN50 sheets.
    (The orthophotography does not cover the entirety of 1: 50,000 sheet) Integral product
    of the NationalObservation Plan of the Territory of Spain: - Joint financing between
    the General State Administration (66%) and the Autonomous Communities (34%) - Hiring
    by the Autonomous Communities - Coordinated by the National Geographic Institute
    Digital Aerial Camera: DMC Panchromatic Focal = 120 mm.</gco:CharacterString>
   </gmd:abstract>
  <gmd:MD_DataIdentification>
 <gmd:identificationInfo>
</gmd:MD_Metadata>
```

Figure 4. An excerpt of an ISO 19115 metadata record extracted from the IDEE geospatial catalogue (translated to English from Spanish).

6. Experimental Results

This section compares the results of the different clustering techniques. Because of the number of experiments, we show the process configurations with the best performance for each different feature representation and clustering algorithm and the list of the ten best ones. Additionally, since one of our objectives was to identify if a general solution without data cleaning is viable, we also show the best results of such configurations.

Table 3 summarises the experiment configurations used and the acronyms shown in the result tables to indicate a specific configuration. In total, 5760 process configurations have been tested. They are all the possible combinations of the following elements: The "Data source" used in the experiments has been the title, abstract, or the title and abstract together. The "Cleaning" processes used are: deleting the text within parentheses (PT), converting all the text values to lowercase (CS), removal of stop words (SW), removal of places (P), and application of stemming (ST). These five cleaning processes generate 32 different cleaning combinations. As "Feature model", we have used: Word2Vec, GloVe, and ELMo word embeddings aggregated using the Mean (M) and Weighted Mean (WM) and TF-IDF, Sentence BERT, and the Universal Sentence Encoder that directly provide sentence representations. Finally, the "Clustering" algorithms include: DBSCAN and OPTICS clustering computed with both the Cosine (Cos) and Euclidean (Eucl) distance, KMEANS, and agglomerative clustering (AG). In the case of K-Means, the number of clusters to create has been manually set to the number identified in the manual classification.

Table 3. Alternatives for each process step in the clustering pipeline.

Process Step	Features	Num Configs
Source	All combinations of Title (T) and Abstract (A)	3
Cleaning	All combinations of Parenthesis texts (PT), Case (CS), Stopwords (SW), Place Names (P), and Stemming (ST)	32
Feature Model	TF-IDF, Word2Vec (M & WM), GloVe (M & WM), ELMo (M & WM), FastText. Sentence BERT, Universal Sentence Encoder	10
Clustering	K-means, DBSCAN (Cos & Eucl), OPTICS (Cos & Eucl), Agglomerative (AG)	6

Table 4 shows the best configuration for each feature representation and clustering technique. The order column shows the experiment rank in terms of V-Measure result. The results have almost the same ordering with both measures and high similarity values. They show that it is possible to identify automatically quasi-spatial dataset series in the collection data with a high precision although, depending on the feature representation and clustering technique used, the quality of the results vary.

Table 4. Process configurations with the best V-Measure for the different feature representations and clustering algorithms.

Order	Source		Cleaning					Feature Model	Clustering	V-Measure	AMI
	T	A	CS	PT	SW	P	ST				
1		X			X	X		**Word2Vec (M)**	**AG**	0.9409	0.8308
2		X			X	X		**GloVe (M)**	AG	0.9405	0.8302
13		X		X		X		**Word2Vec (WM)**	AG	0.9384	0.8225
23		X	X	X	X	X		**FastText (M)**	AG	0.9369	0.8148
39		X	X		X	X		**FastText (WM)**	AG	0.9360	0.8122
50		X			X			**GloVe (WM)**	AG	0.9353	0.8166
251		X	X		X	X		FastText (WM)	**DBSCAN** (Eucl)	0.9272	0.7799
279		X	X	X		X		**UNIVENC**	DBSCAN (Eucl)	0.9254	0.7798
373	X	X	X		X	X		**SBERT**	DBSCAN (Eucl)	0.9178	0.7671
433	X				X	X	X	**TFIDF**	DBSCAN (Cos)	0.9153	0.7791
442	X	X		X			X	TFIDF	**OPTICS** (Eucl)	0.9149	0.7526
662	X	X			X	X		Word2Vec (M)	**KMEANS**	0.9065	0.7609

Whereas Word2Vec and GloVe with agglomerative clustering are the best solutions, classical TF-IDF representation and clustering solutions such as DBSCAN, OPTICS, or KMEANS perform worse in all the cases. It is not surprising that word-embeddings representation works better than TF-IDF because they are richer representations, but it is important to note how the pure sentence-embeddings solutions perform worse than word-embeddings summarisation. We think that the cause of this is the difference between the collections used for training the embeddings and the terminology used in the test collection. Since some geospatial terms are specialised and technical, they may not appear in the training data of the sentence-embeddings models used in the experiment. The agglomerative clustering results were as expected because it was explicitly designed to avoid forcing individual elements inside a cluster as the other techniques do. However, the execution time is much bigger than the other techniques. Using an i5-4590 processor, DBSCAN had a mean cost of 0.03 s, K-means of 0.16 s, OPTICS of 0.83 s, and the agglomerative clustering of 30.77 s. The agglomerative clustering has a cost three orders of magnitude higher than the fastest solution, which becomes a relevant issue with large metadata collections.

Additionally, it is important to note that, in most cases, the use of the titles of the metadata records is not relevant in terms of results. Their content tends to have a negative impact on the results. We think this is caused by the terminology they contain. Even if the descriptions are very different, titles tend to be similar with the same terms repeating in

many resources. This causes small distortions in the generated clusters that worsen the obtained results. Finally, the cleaning steps show that stop words and place names are the elements in the metadata that affect, in a greater way, the clusters generated. This is also natural as they are common words in all the metadata records, so they affect the type of aggregation generated (e.g., clustering by place instead of by theme).

Table 5 shows the 10 best configurations without data cleaning. It can be observed that cleaning improves results, but the difference is small. The first eight configurations use a word-embeddings representation and an agglomerative clustering. The last two use DBSCAN with Euclidean distance. In this case, ELMO, FastText, and Word2Vec have similar results in both measures being small differences in the order. The fact that the two best ELMO results use mean embeddings (M) instead of weighted mean (WM) seems to indicate that word context in ELMO helps to correct the over-representation of common terms, which, in the rest of the techniques, is adjusted using the embeddings weighted mean. The last two results are a bit far from the rest in terms of performance, but they have the advantage of using a general clustering algorithm with a much lower execution time than the agglomerative one.

Table 5. Process configurations without data cleaning step with the best V-Measure.

Order	Source		Cleaning					Feature Model	Clustering	V-Measure	AMI
	T	A	CS	PT	SW	P	ST				
35	X	X						ELMO (M)	AG	0.9361	0.8146
44		X						FastText (WM)	AG	0.9357	0.8130
48		X						ELMO (M)	AG	0.9354	0.8169
61	X	X						FastText (WM)	AG	0.9349	0.8094
99	X	X						Word2Vec (WM)	AG	0.9331	0.8154
105		X						ELMO (WM)	AG	0.9329	0.8035
117	X	X						ELMO (WM)	AG	0.9323	0.7994
305		X						Word2Vec (WM)	AG	0.9239	0.7942
391		X						GloVe (WM)	DBSCAN (Eucl)	0.9169	0.7397
435	X	X						GloVe (M)	DBSCAN (Eucl)	0.9153	0.7313

Finally, Table 6 shows the 10 process configurations with the highest V-measure score among all the experiments performed. It can be observed how the use of Word2Vec of GloVe with agglomerative clustering always produces the best results independently of the other steps. Although the AMI score ordering is a bit different, the change in order is minimal.

Table 6. Process configurations with the best V-Measure.

Order	Source		Cleaning					Feature Model	Clustering	V-Measure	AMI
	T	A	CS	PT	SW	P	ST				
1		X			X	X		Word2Vec (M)	AG	0.9409	0.8308
2		X			X	X		GloVe (M)	AG	0.9405	0.8302
3	X	X			X	X		Word2Vec (M)	AG	0.9405	0.8281
4		X			X			Word2Vec (M)	AG	0.9404	0.8298
5		X			X			GloVe (M)	AG	0.9402	0.8293
6	X	X	X		X			Word2Vec (M)	AG	0.9391	0.8241
7		X	X		X	X		Word2Vec (M)	AG	0.9390	0.8262
8	X	X	X		X	X		Word2Vec (M)	AG	0.9390	0.8229
9		X	X		X	X		GloVe (M)	AG	0.9388	0.8232
10	X	X			X			Word2Vec (M)	AG	0.9388	0.8222

7. Discussion

We have tested multiple clustering solutions on a manually tagged collection to determine if the desired quasi-spatial dataset series can be generated automatically. The results have shown multiple suitable configurations with similar performance.

We can state that the use of word-embeddings representation improves the generation of the desired quasi-spatial dataset series, with respect to classical TF-IDF, even when no data cleaning is performed. Similarly, sentence embeddings could be used for feature representation at a small loss of performance. The results obtained with word-embeddings and sentence-embeddings representations always outperform the equivalent TF-IDF representations. This shows that they express better the information contained in the metadata records. However, it is also necessary to be careful with these solutions as sentence embeddings have proven to be dependent on the training data, making difficult to determine how they would behave with other collections with different terminology.

With respect to the clustering, even though agglomerative clustering outperforms the rest of the analysed techniques and deals well with the problem of one-element clusters, its execution time may discourage its use for big collections. In those cases, DBSCAN is faster and has proven to have a close performance.

The proposed quasi-spatial dataset series generation presents some limitations that have to be taken into account in the obtained results. Firstly, the process is completely dependent on the metadata quality. It requires a complete description of the resources, so the similarity in the definitions can be calculated. This may seem obvious, but currently there are many geospatial metadata collections with short descriptions where the proposed process could not be applied. Secondly, due to the nature of the algorithms, even though the quality of the generated aggregations is good, it is not perfect. Therefore, the results have to be interpreted by the users to determine if they make sense or not. Finally, the current proposal does not provide any intra-clustering ordering of the results, since it is not able to identify the nature of the identified clusters. A cluster may contain resources distributed along the space containing similar content, focus on the same area but with different creation times, or both of them simultaneously. A solution for this problem needs to be studied in future work.

8. Conclusions

This paper has shown how spatial fragmentation in geospatial catalogues can cause ineffectiveness in "concept at location" searches. We have summarised existing IR problems and described the existing dissonance between the continuous nature of geospatial information and the digital library based structure of these metadata catalogues. To solve this problem, we have proposed the automatic identification of quasi-spatial dataset series to provide aggregated results that can be used to improve query result lists.

We have shown how current clustering techniques can be used to generate good quality quasi-spatial dataset series using, as baseline, a Spanish metadata collection manually tagged. The results show clearly that the use of word embeddings with agglomerative clustering is the best solution, but it can be replaced with DBSCAN if execution time is a relevant factor.

As future work, we want to extend the proposed approach towards the direction of providing a continuous layer solution. The identified quasi-spatial dataset series can be added to the corresponding catalogue IR system using a spatial metadata automation tool. In this way, they could be provided as query results that make it easier for users to find data that meet their needs. For this purpose, we plan to develop an enrichment pipeline that allows integrating the heterogeneous resources of a quasi-spatial dataset series into a single resource. This will be useful not only for improving search capabilities of geospatial catalogues but also for data analysis. For example, these integrated layers would make it possible to identify areas with no data about a theme or to find areas with better or worse data quality. An equivalent problem to spatial fragmentation is the temporal fragmentation of data. Given the bigger relevance of spatial aspects, we have only focused on spatial

fragmentation, but we want to analyse the temporal management problem to determine if the same solutions proposed for spatial aspects can be applied. As part of this process, it would be necessary to analyse if additional metadata elements could be used and how to deal with multilingual catalogues. Another area of improvement is the identification of the nature of the clusters and presentation of the results. If the spatial or temporal relation present in each cluster can be identified, the content of each cluster can be presented in an ordered way that would simplify the task of analysing the content of the results.

Author Contributions: Conceptualization, Javier Lacasta and Francisco Javier Lopez-Pellicer; methodology, Javier Lacasta and Francisco Javier Lopez-Pellicer; software, Javier Lacasta and Francisco Javier Lopez-Pellicer; validation, Javier Lacasta, Francisco Javier Lopez-Pellicer, Javier Nogueras-Iso, Rubén Béjar, and Javier Zarazaga-Soria; formal analysis, Javier Lacasta and Francisco Javier Lopez-Pellicer; investigation, Javier Lacasta and Francisco Javier Lopez-Pellicer; resources, Javier Lacasta and Francisco Javier Lopez-Pellicer; data curation, Javier Lacasta and Francisco Javier Lopez-Pellicer; writing—original draft preparation, Javier Lacasta and Francisco Javier Lopez-Pellicer; writing—review and editing, Javier Nogueras-Iso and Rubén Béjar; visualisation, Javier Nogueras-Iso and Rubén Béjar; supervision, Javier Zarazaga-Soria; project administration, Javier Zarazaga-Soria; funding acquisition, Javier Zarazaga-Soria. All authors have read and agreed to the published version of the manuscript.

Funding: This work is part of the project T59_20R supported by the Regional Government of Aragon (Spain) and the project PID2020-113353RB-I00 supported by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033/).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and codes that support the findings of this study are available at "figshare.com" with the DOI: 10.6084/m9.figshare.13705945 (accessed on 26 November 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nebert, D. (Ed.) *Developing Spatial Data Infrastructures: The SDI Cookbook*; Global Spatial Data Infrastructure (GSDI). 2004. Available online: http://gsdiassociation.org/images/publications/cookbooks/SDI_Cookbook_GSDI_2004_ver2.pdf (accessed on 29 November 2021).
2. ISO 19115-1:2014-Geographic Information—Metadata—Part 1: Fundamentals. International Organization for Standardization (ISO): Geneva, Switzerland, 2014. Available online: https://iso.statuspage.io/#!iso:std:53798:en (accessed on 26 November 2021).
3. Da Silva Santos, L.B.; Wilkinson, M.D.; Kuzniar, A.; Kaliyaperumal, R.; Thompson, M.; Dumontier, M.; Burger, K. FAIR data points supporting big data interoperability. In *Enterprise Interoperability in the Digitized and Networked Factory of the Future*; ISTE: London, UK, 2016; pp. 270–279.
4. Hubner, S.; Spittel, R.; Visser, U.; Vogele, T.J. Ontology-based search for interactive digital maps. *IEEE Intell. Syst.* **2004**, *19*, 80–86. [CrossRef]
5. Larson, J.; Olmos, M.A.; Pereira, M. Are geospatial catalogues reaching their goals? In Proceedings of the 9th AGILE Conference on Geographic Information Science: Shaping the Future of Geographic Information Science in Europe, Visegrád, Hungary, 20–22 April 2006; pp. 1–8.
6. Fugazza, C.; Tagliolato, P.; Frigerio, L.; Carrara, P. Web-scale normalization of geospatial metadata based on semantics-aware data sources. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 354. [CrossRef]
7. Dareshiri, S.; Farnaghi, M.; Sahelgozin, M. A recommender geoportal for geospatial resource discovery and recommendation. *J. Spat. Sci.* **2019**, *64*, 49–71. [CrossRef]
8. Ivanova, I.; Brown, N.; Fraser, R.; Tengku, N.; Rubinov, E. Fair and standard access to spatial data as the means for achieving sustainable development goals. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.-ISPRS Arch.* **2019**, *42*, 33–39. [CrossRef]
9. Giuliani, G.; Cazeaux, H.; Burgi, P.Y.; Poussin, C.; Richard, J.P.; Chatenoux, B. SwissEnvEO: A FAIR National Environmental Data Repository for Earth Observation Open Science. *Data Sci. J.* **2021**, *20*. [CrossRef]
10. ISO 19131:2007. Geographic Information—Data Product Specifications. International Organization for Standardization (ISO). Available online: https://iso.statuspage.io/#iso:std:iso:19131:ed-1:en (accessed on 26 November 2021).
11. Larson, R.; Frontiera, P. Ranking and representation for geographic information retrieval. In Proceedings of the Extended Abstract in SIGIR 2004 Workshop on Geographic Information Retrieval, Sheffield, UK, 29 July 2004; pp. 1–3.
12. Zhan, Q.; Zhang, X.; Li, D. Ontology-based semantic description model for discovery and retrieval of geospatial information. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2008**, *32*, 141–146.

13. Zhang, Y.; Chiang, Y.Y.; Szekely, P.; Knoblock, C.A. A semantic approach to retrieving, linking, and integrating heterogeneous geospatial data. In Proceedings of the Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities, Beijing, China, 4–5 August 2013; pp. 31–37.
14. De Andrade, F.G.; de Souza Baptista, C.; Davis, C.A. Improving geographic information retrieval in spatial data infrastructures. *GeoInformatica* **2014**, *18*, 793–818. [CrossRef]
15. Li, W.; Goodchild, M.F.; Raskin, R. Towards geospatial semantic search: Exploiting latent semantic relations in geospatial data. *Int. J. Digit. Earth* **2014**, *7*, 17–37. [CrossRef]
16. Fugazza, C.; Pepe, M.; Oggioni, A.; Tagliolato, P.; Carrara, P. Raising semantics-awareness in geospatial metadata management. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 370. [CrossRef]
17. Fugazza, C.; d'Aragona, P.T.A.; Oggioni, A.; Carrara, P. Decentralized geospatial metadata management. *Earth Sci. Inform.* **2021**, *14*, 1579–1596. [CrossRef]
18. Miao, L.; Liu, C.; Fan, L.; Kwan, M.P. An OGC web service geospatial data semantic similarity model for improving geospatial service discovery. *Open Geosci.* **2021**, *13*, 245–261. [CrossRef]
19. Li, Y.; Jiang, Y.; Yang, C.; Yu, M.; Kamal, L.; Armstrong, E.; Huang, T.; Moroni, D.; McGibbney, L. Improving search ranking of geospatial data based on deep learning using user behavior data. *Comput. Geosci.* **2020**, *142*, 104520. [CrossRef]
20. Aggarwal, C.C.; Zhai, C. A Survey of Text Clustering Algorithms. In *Mining Text Data*; Springer: Boston, MA, USA, 2012; Chapter A: Survey of Text Clustering Algorithms, pp. 77–128.
21. Ma, L.; Zhang, Y. Using Word2Vec to process big text data. In Proceedings of the 2015 IEEE International Conference on Big Data, Santa Clara, CA, USA, 29 October–1 November 2015; pp. 2895–2897.
22. Li, C.; Lu, Y.; Wu, J.; Zhang, Y.; Xia, Z.; Wang, T.; Yu, D.; Chen, X.; Liu, P.; Guo, J. LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering. In Proceedings of the Companion Proceedings of the Web Conference 2018, Lyon, France, 23–27 April 2018; pp. 1699–1706.
23. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *2*, 3111–3119.
24. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1532–1543. Available online: https://aclanthology.org/D14-1162/ (accessed on 26 November 2021).
25. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
26. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: New Orleans, LA, USA, 2018; Volume 1, pp. 2227–2237. Available online: https://aclanthology.org/N18-1202/ (accessed on 26 November 2021).
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
28. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
29. Arora, S.; Liang, Y.; Ma, T. A Simple But Tough-to-Beat Baseline for Sentence Embeddings. In Proceedings of the International Conference on Learning Representations. Available online: https://openreview.net/pdf?id=SyK00v5xx (accessed on 26 November 2021).
30. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. *arXiv* **2014**, arXiv:1405.4053.
31. Riemers, N.; Gurevych, I. Sentence Embeddings using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084.
32. Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; Bordes, A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv* **2017**, arXiv:1705.02364.
33. Cer, D.; Yang, Y.; Kong, S.; Hua, N.; Limtiaco, N.; John, R.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C. Universal Sentence Encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 169–174. Available online: https://aclanthology.org/D18-2029/ (accessed on 26 November 2021).
34. Kusner, M.; Sun, Y.; Kolkin, N.; Weinberger, K. From word embeddings to document distances. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 957–966.
35. Zhang, C.; Tao, F.; Chen, X.; Shen, J.; Jiang, M.; Sadler, B.; Han, J. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. *arXiv* **2018**, arXiv:1812.09551.
36. Hu, K.; Luo, Q.; Qi, K.; Yang, S.; Mao, J.; Fu, X.; Zheng, J.; Wu, H.; Guo, Y.; Zhu, Q. Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. *Inf. Process. Manag.* **2019**, *56*, 1185–1203. [CrossRef]
37. Diaz, J.; Poblete, B.; Bravo-Marquez, F. An integrated model for textual social media data with spatio-temporal dimensions. *Inf. Process. Manag.* **2020**, *57*, 102219. [CrossRef]
38. Li, Y.; Cai, J.; Wang, J. A Text Document Clustering Method Based on Weighted BERT Model. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; Volume 1, pp. 1426–1430.

39. Arenas-Márquez, F.J.; Martinez-Torres, R.; Toral, S. Convolutional neural encoding of online reviews for the identification of travel group type topics on TripAdvisor. *Inf. Process. Manag.* **2021**, *58*, 102645. [CrossRef]
40. Zola, P.; Ragno, C.; Cortez, P. A Google Trends spatial clustering approach for a worldwide Twitter user geolocation. *Inf. Process. Manag.* **2020**, *57*, 102312. [CrossRef]
41. Newman, D.; Hagedorn, K.; Chemudugunta, C.; Smyth, P. Subject metadata enrichment using statistical topic models. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, Vancouver, BC, Canada, 18–23 June 2007; pp. 366–375.
42. Lacasta, J.; Nogueras-Iso, J.; Muro-Medrano, P.R.; Zarazaga-Soria, F.J. Thematic clustering of geographic resource metadata collections. In *International Symposium on Web and Wireless Geographical Information Systems*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 30–43.
43. Thomas, R.E.; Khan, S.S. Improved clustering technique using metadata for text mining. In Proceedings of the 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 21–22 October 2016; pp. 1–5.
44. Rajan, A.; Mittas, N.; Mehrotra, D. Clustering the Patent Data Using K-Means Approach. In *Software Engineering. Advances in Intelligent Systems and Computing*; Hoda, M., Chauhan, N., Quadri, S., Srivastava, P., Eds.; Springer: Singapore, 2019; Volume 731, pp. 639–645.
45. Rakib, M.R.H.; Zeh, N.; Jankowska, M.; Milios, E. Enhancement of short text clustering by iterative classification. In *International Conference on Applications of Natural Language to Information Systems*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 105–117.
46. Cai, Z.; Wang, J.; He, K. Adaptive density-based spatial clustering for massive data analysis. *IEEE Access* **2020**, *8*, 23346–23358. [CrossRef]
47. Lou, W.; Su, Z.; He, J.; Li, K. A temporally dynamic examination of research method usage in the Chinese library and information science community. *Inf. Process. Manag.* **2021**, *58*, 102686. [CrossRef]
48. Misztal-Radecka, J.; Indurkhya, B. Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems. *Inf. Process. Manag.* **2021**, *58*, 102519. [CrossRef]
49. Ahmad, M.; Ali, A. Mapping National Spatial Data Infrastructure Initiatives. 2019. Available online: https://www.google.com/maps/d/viewer?mid=1596RIb8g_n0LPyi55-N1E2PuDw4&ll=24.147211357953225%2C-86.74911452879445&z=2 (accessed on 26 November 2021).
50. Kalantari, M.; Syahrudin, S.; Rajabifard, A.; Subagyo, H.; Hubbard, H. Spatial Metadata Usability Evaluation. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 463. [CrossRef]
51. Hennig, S.; Belgui, M. User-centric SDI: Addressing users requirements in third-generation SDI. The Example of Nature-SDIplus. *Geoforum Perspekt.* **2011**, *10*, 30–42.
52. Masó, J.; Pons, X.; Zabala, A. Tuning the second-generation SDI: Theoretical aspects and real use cases. *Int. J. Geogr. Inf. Sci.* **2012**, *26*, 983–1014. [CrossRef]
53. Lacasta, J.; Lopez-Pellicer, F.J.; Espejo-García, B.; Nogueras-Iso, J.; Zarazaga-Soria, F.J. Aggregation-based information retrieval system for geospatial data catalogs. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1583–1605. [CrossRef]
54. Latre, M.A.; Lacasta, J.; Mojica-Abrego, E.; Nogueras-Iso, J.; Zarazaga-Soria, F.J. An Approach to Facilitate the Integration of Hydrological Data by means of Ontologies and Multilingual Thesauri. In *Advances in GIScience. Lecture Notes in Geoinformation and Cartography (LNG&C)*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 155–171.
55. Ingersoll, G.S.; Morton, T.S.; Farris, A.L. *Taming Text: How to Find, Organize, and Manipulate It*; Manning: Shelter Island, NY, USA, 2012.
56. Porter, M.F. Snowball: A Language for Stemming Algorithms. 2001. Available online: http://snowball.tartarus.org/texts/introduction.html (accessed on 26 November 2021).
57. Cardellino, C. Spanish Billion Words Corpus and Embeddings. 2016. Available online: https://crscardellino.ar/SBWCE/ (accessed on 26 November 2021).
58. Che, W.; Liu, Y.; Wang, Y.; Zheng, B.; Liu, T. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. *arXiv* **2018**, arXiv:1807.03121.
59. Hartigan, J.A. *Clustering Algorithms*; John Willey & Sons: New York, NY, USA, 1975.
60. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; Simoudis, E., Han, J., Fayyad, U., Eds.; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 1996; pp. 226–231.
61. Verma, M.; Srivastava, M.; Chack, N.; Diswar, A.K.; Gupta, N. A comparative study of various clustering algorithms in data mining. *Int. J. Eng. Res. Appl.* **2012**, *2*, 1379–1384.
62. Voorhees, E.M. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Inf. Process. Manag.* **1986**, *22*, 465–476. [CrossRef]
63. Rosenberg, A.; Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*; Eisner, J., Ed.; Association for Computational Linguistics: Prague, Czech Republic, 2007; pp. 410–420. Available online: https://aclanthology.org/D07-1043/ (accessed on 26 November 2021).
64. Vinh, N.X.; Epps, J.; Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **2010**, *11*, 2837–2854.