TESIS DE LA UNIVERSIDAD DE ZARAGOZA
2025
222

Sergio Martín Segura

Improving Access and Discovery of Spatial Resources in Catalogues

Director/es
López Pellicer, Francisco Javier
Zarazaga Soria, Francisco Javier

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad de Zaragoza
Servicio de Publicaciones

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

IMPROVING ACCESS AND DISCOVERY OF SPATIAL RESOURCES IN CATALOGUES

Autor

Sergio Martín Segura

Director/es

López Pellicer, Francisco Javier
Zarazaga Soria, Francisco Javier

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Ingeniería de Sistemas e Informática

2025

Universidad Zaragoza
1474

Tesis Doctoral

Improving Access and Discovery of Spatial Resources in Catalogues

Autor

Sergio Martín-Segura

Director/es

Francisco Javier López Pellicer

Francisco Javier Zarazaga Soria

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2025

— To Mom and Dad,
who chiseled the foundations of who I am today.

Acknowledgements

I also want to thank the rest of my environment: my brother Diego; my partner Marina; and my friends Daniel, Claudia, Adriana, and many others, for being and invaluable support and a place to rest.

To my academic environment: my supervisors Javier and Javier, and my group mates Javier, Javier, Rubén, and Miguel Ángel, for teaching me valuable lessons and guiding me through this stormy journey.

I also want to thank my former colleagues: Borja, for warning me about everything; Víctor, for always arguing with me; and Alejandro, Jorge, and Yineth, for giving me the best years I have had in the group (so far)... and the ones that are still here: Héctor, for being an inspiration; Mario, for being a reference of work ethics; and Dagoberto, Abul, Hala, and Umair, for teaching me the true meaning of hospitality.

I also want to thank my hosts and colleagues during my stay at the Universidade da Coruña: Miguel, Alejandro, Victor, and David, as well as the rest of the LBD group, and my "kids" from the residence: Iván, Joel, and Daniel, for being the fire extinguisher that put out the imminent burnout.

Finally, I want to thank the people who have reviewed this thesis, providing me with their patience and wisdom to help me polish it as much as I could.

Resumen

Las Infraestructuras de Datos Espaciales (IDE) continúan siendo, en muchos contextos, las principales fuentes de datos geoespaciales. Con el constante crecimiento del volumen de estos datos, la capacidad de encontrar el recurso espacial adecuado se ha vuelto cada vez más crítica. Sin embargo, este proceso suele ser complicado.

En esta tesis, abordamos la problemática del descubrimiento de recursos espaciales en catálogos de datos geoespaciales. Para ello, realizamos una serie de estudios empíricos que identifican los problemas que limitan la efectividad de los catálogos y analizan su impacto real. Demostramos que, en muchos casos, los registros de metadatos de los catálogos no son suficientes para localizar el recurso deseado. Proporcionamos una definición formal del fenómeno conocido como Metadata Reference Rot, analizando sus componentes, y evidenciamos que afecta significativamente a los metadatos espaciales: menos del 75% de los registros contienen al menos una URL de distribución accesible.

Además, demostramos que los sistemas de recuperación espacial actuales, basados en la representación de la *Minimum Bounding Box* (MBB), suelen generar resultados falsos positivos. Los resultados empíricos evidencian una precisión muy variable entre catálogos, y ninguno logra métricas suficientemente altas.

Como alternativa, proponemos un nuevo enfoque para mejorar la precisión de las búsquedas espaciales. Presentamos el DGGS Footprint, un método que utiliza Discrete Global Grid Systems (DGGS) para representar la extensión espacial de un conjunto de datos mediante una lista de celdas que intersectan con su huella real. Esta representación permite realizar búsquedas espaciales más precisas que los métodos tradicionales, como MBB o Convex Hull. Nuestro estudio empírico confirma que el DGGS Footprint mejora significativamente la precisión de las búsquedas, alcanzando un promedio superior al 96% en todos los catálogos analizados.

Abstract

Spatial Data Infrastructures (SDIs) remain, in many contexts, the primary sources of geospatial data. With the constant growth of geospatial data volumes, the ability to identify the correct spatial resource has become increasingly critical, yet often challenging.

This thesis addresses the problem of discovering spatial resources in geospatial data catalogues. Through a series of empirical studies, we identify key issues limiting the effectiveness of these catalogues and evaluate their real-world impact. Our findings reveal that, in many cases, the metadata records in spatial catalogues are insufficient for locating the desired resource. We introduce a formal definition of Metadata Reference Rot and analyze its components, demonstrating its significant prevalence in spatial metadata: fewer than 75% of metadata records contain at least one accessible distribution URL.

Furthermore, we show that current spatial retrieval systems, relying on *Minimum Bounding Box* (MBB) representations, are prone to generating false positive results. Empirical results highlight considerable variability in catalogue precision, with none achieving sufficiently high metrics.

To address these limitations, we propose an alternative approach for improving search precision. We introduce DGGS Footprint, a novel method leveraging Discrete Global Grid Systems (DGGS) to represent the spatial extent of datasets. By encoding the dataset's footprint as a list of intersecting cells, this approach enables more accurate spatial searches compared to traditional MBB or Convex Hull methods. Our empirical evaluation demonstrates that DGGS Footprint significantly enhances search precision, achieving an average accuracy of over 96% across all studied catalogues.

Index

Chapter 1

Introduction

The history of geography and cartography spans the history of humanity. From the first maps drawn in the caves of Altamira, to the digital maps that accompany us on our mobile devices, geography and cartography have been fundamental tools in various fields such as navigation, urban planning, natural resource management, policing, economics, transportation, health, or scientific research. The last 20 years have witnessed a significant and growing proliferation of geospatial information, both in terms of its availability and its possibilities of use. This proliferation is due to aspects such as the rise of always connected mobile devices or sensorized vehicles, the availability of more complete, accurate and cheaper cartography, the rise of earth observation satellites, the possibilities of administrations and companies for the creation of thematic geospatial information, the popularization of free applications and services (e.g. Google maps, Google Earth, Bing Maps, Open Street Map, etc), agreements for standardization (Open Geospatial Consortium (OGC), International Organization for Standardization (ISO)) and their implementation, support for interoperability of commercial and open source components, as well as the involvement of national and European authorities (European Directive 2007/2/EC INSPIRE) to facilitate the discovery, access and use of geospatial information and services. Currently, spatial data is a fundamental pillar of the information society, being a key element in decision-making. Disciplines such as epidemiology, energy engineering, architecture, history, tourism, etc. generate and consume data with spatial information.

Geographic Information, also known as spatial data or geospatial data, is information that describes phenomena associated directly or indirectly with a location relative to the Earth's surface. Currently, there are large amounts of geographic data that have been collected (for decades) for different purposes and by different institutions and companies. For example, geographic information is vital for decision-making systems and resource management systems (natural resources, basic supply networks, cadastre, economy, agriculture, etc.) at different levels (local, provincial, regional,

national or even global) (Rao et al., 2015; Jaloliddinov et al., 2024). In addition, the volume of information is growing day by day due to various satellite deployment initiatives (Global Navigation Satellite System and the Copernicus program), databases and geoprocessing software, not to mention the growing interest of individuals and institutions. With the correct methods, it is even possible to georeference complex collections of a wide range of types of resources, including textual and graphic documents, digital maps, images, real-time observations or legacy databases of historical tabular records.

In the last 25 years, nations have made unprecedented investments in both information and means to collect, store, process, analyze and disseminate geospatial information. Thousands of organizations and agencies (different government levels, private sectors, non-profit organizations, research centers) around the world spend billions of Euros per year to produce and use geographic information (United Nations GGIM, 2021). This has been facilitated by the rapid advancement of capture technologies, which has made the capture of geospatial information relatively quick and easy. Additionally, the impact of the Internet on the distribution of geospatial information resources must be mentioned. As with other types of resources, quantities of geospatial information resources are available on the Internet. And in some cases it is even assumed that the Internet itself is the information store.

However, it is common that each new project or study involving the use of geospatial information requires the creation of new resources from scratch. This apparent lack of reusable resources is often motivated by the following circumstances: the lack of financial resources causes one-off efforts to be made without long-term maintenance; some organizations, despite being public, are reluctant to distribute high-quality geographic information; data collected by different institutions is often incompatible; the lack of knowledge, in most cases, of what is currently available; the poor quality and documentation of what is available through the Web; and the increasing complexity and disparity of search and retrieval systems over the Internet.

Despite the potential uses of geospatial information and the investments in its creation, it is not sufficiently exploited (European Commission, 2003; Kügeler and Jirka, 2021; Quarati et al., 2021). It is said that 'information is power' but with the increasing amounts of data that are created and stored (often, not in an organized way) there is a real need to document the data for future use (to make them as accessible as possible and to a wider audience). Information is defined as the set of data plus the context for its use. Data without context is not as valuable as documented data. This need is of extreme importance in the case of geographic information. Once created, geographic data can be used by multiple systems and for different purposes.

It is clear that there is a need to create distributed solutions that facilitate the search, evaluation and access to data. In this reality of the late 90s, the concept of Spatial Data Infrastructures (*Spatial Data Infrastructure*) was born. *Spatial Data Infrastructure*s are infrastructures oriented to the optimization of the creation, maintenance and distribution of geographic information at different organizational levels (e.g. local, regional, national, global) and involving both public and private institutions (Nebert, 2004; Masser, 2019). The first formal definition of the term *Spatial Data Infrastructure* was formulated in the Federal Register of the United States in 1994 (U.S. Federal Register, 1994): 'An *Spatial Data Infrastructure* means the technology, policies, standards, and human resources to acquire, process, store, distribute, and improve the use of geospatial data'. The definition given by the GSDI (Global Spatial Data Infrastructure Association) [1] is also very similar: 'A coordinated approach of technology, policies, standards, and human resources necessary for efficient acquisition, management, storage, distribution and use of geospatial data in the development of a global community'.

Spatial Data Infrastructures have become as important as basic supply infrastructures (electricity, water, gas), transportation or telecommunications. This consideration, already with some history in countries like the United States (U.S. Federal Register, 1994) or Canada, received a very important boost in the European context since the launch of the INSPIRE Directive (INfrastructure for SPatial InfoRmation in Europe) by the European Commission (Parliament and Council, 2007). INSPIRE aims to establish the technical and policy foundations to create a true European common space of geographic information, in which everyone knows how to access Web services related to geographic information regardless of the country or European region in which they are located. By delegation, the member countries and, in Spain, the Autonomous Communities have the mandate to facilitate the development of the infrastructure in their area of influence. INSPIRE was born with a focus on environmental issues, but with the ambition to be developed in other areas (agriculture, transport, etc.).

Due to the fact that the concept of *Spatial Data Infrastructures* comes from the domain of geographic information, the first *Spatial Data Infrastructures* were built, from a technical point of view, based on the concepts and experiences provided by traditional Geographic Information Systems (GIS). The term GIS is commonly used to refer to software packages that are able to integrate spatial and non-spatial data to obtain the information necessary for decision-making. However, these GISs used as tools to perform particular analyses in isolated projects expanded to distributed and

[1] `http://gsdiassociation.org/`

cooperative environments, not only from a technical perspective but also taking into account cooperation policies between different public or private organizations and at different levels (local, regional, national or global).

Nowadays, the development of spatial data infrastructures constitutes a multidisciplinary application context that combines the experience and knowledge of different disciplines. In particular, Digital Libraries provide a very important knowledge base (Cantán et al., 2009; Béjar et al., 2009). There is a great deal of experience in technology for the distribution of digital resources that can be used as a basis for the concepts of *Spatial Data Infrastructures*, as well as processes and methods.

According to the classic *Spatial Data Infrastructure* model (Coleman and Nebert, 1998), among the components of one of them the following should be included:

- Technology. *Spatial Data Infrastructures* must be developed on the technological components created from the experience gained working with generic information technology. One of the most important challenges is the integration of all this experience, especially that provided by GIS.

- Standards and guidelines. Standards constitute the link between the different components of an *Spatial Data Infrastructure* providing common languages and concepts that make communication and coordination possible. Additionally, it is necessary to establish guidelines that are followed by all the actors involved in the *Spatial Data Infrastructures*. These guidelines should include different aspects such as architectures, processes, methods and standards.

- Human resources. The development of an *Spatial Data Infrastructure* must be based on the needs of its users, both end users and data producers. And on the other hand, the work of implementing and maintaining an *Spatial Data Infrastructure* must be carried out by qualified teams of researchers and developers. All these people make up the human resources that are necessary for the development of *Spatial Data Infrastructures*.

- Institutional agreements. It is necessary to establish political decisions that allow the creation of an institutional framework. The agreements must serve to establish *Spatial Data Infrastructures* at the local level, and to coordinate the creation of regional, national or even global *Spatial Data Infrastructures*. Although political aspects are not the core of *Spatial Data Infrastructures*, they do exert a great influence on their development (Nogueras-Iso et al., 2004; Zarazaga-Soria et al., 2004).

– Spatial databases and metadata. *Spatial Data Infrastructures* must provide access to geographic data, stored in spatial databases, and properly documented through a series of metadata.

– Data networks. *Spatial Data Infrastructures* must be open systems deployed over data networks that provide the access channel to services accessed from remote systems.

We will now proceed to mention the elements of an *Spatial Data Infrastructure* that are most relevant to this work: spatial metadata, geospatial data catalogues, and data access services. The selected elements make up only a part of what an *Spatial Data Infrastructure* is in its entirety, but they are the ones most related to the development of the thesis.

Metadata is commonly defined as 'structured data about data' or 'data that describes the attributes of a resource' or more simply 'information about data'. They describe the content, quality, condition, and other characteristics of a resource, constituting the mechanism to characterize data and services so that users (and applications) can locate and make use of that data and services.

As mentioned in (Nebert, 2004), geographic metadata helps people involved in the use of geographic information to find the data they need and to determine the best way to use it.

Concerning existing geographic metadata standards, the most important one is the international standard ISO 19115 (ISO, 2003; Maganto et al., 2008) for geographic information metadata. The ISO created the 211 committee (ISO/TC 211) in 1992 with responsibilities in Geographic Information and Geomatics. This committee has been responsible for preparing a family of standards in this context. The ISO19115 standard was approved in May 2003 and defines elements that allow the description, among others, of identification, extent, quality, spatial representation scheme, reference systems used, and data distribution form. Although this standard is mainly oriented towards the cataloguing of geographic datasets (including series or individual features/entities) in digital format, it can also be extended to other forms of geographic data, such as maps, textual documents, or non-geographic data. It is also worth mentioning the less widespread, but equally important, use of generic metadata standards such as Dublin Core (DCMI, 2004). Dublin Core is a standard for the description of information resources in cross-domains, that is, the description of all kinds of resources regardless of their format, area of specialization or cultural origin. This standard consists of fifteen basic descriptors that are the result of international and interdisciplinary consensus. Dublin Core has become an important part of the Internet

infrastructure and that includes its use in the description of geographic resources.

An important component of an *Spatial Data Infrastructure* is the *Geospatial Data Catalogue* (GDC). *Geospatial Data Catalogues* are the solution for publishing descriptions of geographic information resources (metadata) in a standardized way, making it easier for users to locate the data of interest (Kottman, 1999). The catalogues present a Geographic Information Retrieval (GIR) challenge: provide access to relevant information resources in response to queries with geographic constraints (Larson, 2009). The goal of the GIR is to retrieve documents thematically and geographically related to a given query (defined as 'concept at location' by Hubner et al. (2004)) whether the spatial context is defined by place names and references or by coordinate-encoded locations. In the context of GDCs, the search engines implement GIR algorithms that combine textual and spatial search over explictly georeferenced search items. This way, a user can input a textual query while also selecting an area of interest and the system will filter out the datasets unrelated to that area.

They serve two subtly different use cases: the discovery and access to the indexed resources. In the context of Information Retrieval, access covers use cases in which a user seeks to locate a known resource while discovery implies the desire to obtain resources that satisfy a need, but are not known at that time. An example of an access search would be 'PNOA 2024' (Plan Nacional de Ortofotografía Aérea) while an example of a discovery search would be 'orthophotos 2024'. In this work, we will address both use cases, but with a special focus on discovery, as it is the one that presents the most problems today.

Another important area in the development of *Spatial Data Infrastructures* is the one corresponding to the geospatial data access components. The Distribution Information is made up of all the information necessary to locate and access the geographic data. It often includes one or more URLs that point to the data, as well as information about the data format, projection, resolution, quality, timeliness, coverage, license, etc.

Once the data of interest has been located, it is necessary to visualize and evaluate the data. Then, if this data is desired, advanced users will require access to the data in its original format. The Web Map Server (WMS) component is a map server that offers graphical views (or maps) of the geographic data information through service interfaces accessible online (Beaujardière, 2002; Fernández et al., 2000). With this type of component, it is possible to evaluate and meet many of the users' needs without requiring the complete download of the data. But if this final access were necessary, the components called Web Feature Server (WFS) (Vretanos, 2002) and Web Coverage Server (WCS) (Evans, 2003) could be used. The WFS is a server that allows access

to discrete phenomena (features) in GML (Geography Markup Language) format (Cox et al., 2003). On the contrary, the WCS is a server that allows access to continuous phenomena or coverages in raster formats. In the implementation of these components, it is vital to use technology capable of handling large volumes of data efficiently and effectively. Aspects related to the transparent partitioning of information, optimization of compression algorithms, and retrieval and concatenation of data come into play in order to be able to visualize them from any application, both locally and via the Web.

Parallel to the spatial data exchange standards, generalist distribution methods such as HTTP (HyperText Transfer Protocol) or even FTP (File Transfer Protocol) downloads coexist. These are also used by industry and agencies as they facilitate access through any web browser. These services have interfaces and schemes that are much less rigid than OGC services, which can be an advantage or a disadvantage, depending on how mature the quality practices of the manager are. In the case of HTTP, the protocol establishes a series of attributes that allow:

- Content negotiation: the client can request a resource in one or more specific formats, and the server can return the resource in the format that best suits the client's needs.
- Error control: each server response includes a status code that indicates whether the request has been processed correctly or if there has been an error. The values of this field are profusely documented and standardized.
- Authentication: the HTTP protocol allows user authentication to control access to restricted resources.

On summary, the use of both systems, Catalogue and Access, combined would result in the following workflow (see Figure 1.1).

1. The user accesses the catalog and enters the search values: text, area of interest or other extra attributes.
2. The catalog returns a list of results for that specific search.
3. The user selects the resource they find most interesting and accesses the metadata page of that resource.
4. On the metadata page, the user can find additional information about the resource and, if interested, can access the resource distribution. The resource distribution can be a link to a WMS, WFS, WCS service, or a link to a downloadable file.

5. The user chooses the distribution that interests them the most and accesses the original resource.

6. It is then when the user can visualize the actual data and check if it is the resource they were looking for.

7. If not, they will have to return to point 3 and select another resource.

1.1 Challenges

Despite the potential value that the community agrees SDIs can provide (Dangermond and Goodchild, 2020), their development has proceeded unevenly. On one hand, academia and certain interested agencies drive their evolution, while on the other hand, the rest of the stakeholders just comply—often skeptically and with low resources—with the mandates and agreements. Academia and Industry are drifting apart: the academia has stopped studying the catalogues while the industry has stopped following the academia. This disconnection has led to a distancing from the real problems of the catalogues, which has led to a cooling of the topics and a decrease in the number of publications that address these topics. An evidence of this is the many years it took for INSPIRE to be fully implemented (Kotsev et al., 2021). The lack of resources and skepticism have hindered the development and maturity of SDIs and their catalogues, preventing them from achieving the utility and ubiquity that search engines and similar information systems have attained.

There is no 'Google for spatial data', and users know it. The experience of a user looking for spatial data is usually frustrating. First, popular implementations like GeoNetwork[2]—used by the Spanish Official catalogue[3], the French[4] one and many others[5]—or CKAN[6]—used by the official Open Data Portal of the USA[7], the Australian

[2] https://www.geonetwork-opensource.org/
[3] https://www.idee.es/csw-inspire-idee/srv/eng/catalog.search
[4] http://catalogue.geo-ide.developpement-durable.gouv.fr/catalogue
[5] https://docs.geonetwork-opensource.org/4.2/fr/annexes/gallery/
[6] https://ckan.org/
[7] https://catalog.data.gov

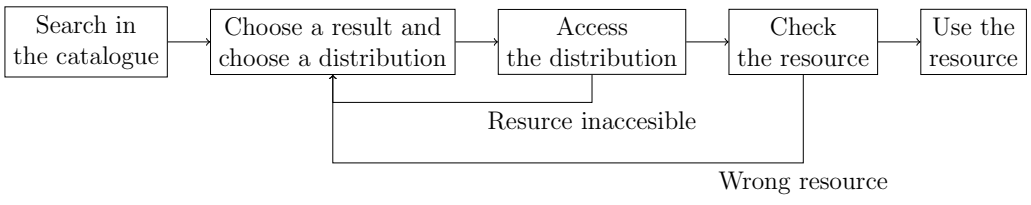

Figure 1.1: Step-by-step activity flow of a user searching for a resource in a catalogue

one[8] and many others[9]—do not even implement basic search engine features such as ranked search or fuzzy search. On one hand, ranked search shows the results ordered by their relevance to the query. On the other hand, fuzzy search matches with similar words to the ones searched, which improves the results and the tolerance to typos. This means that if a user searches for "temperatures" and the metadata record contains "temperature," it will not be present in the search results.

On top of the problems of textual search are the problems of spatial search. When a user selects the geographic area where they want to search, the catalog rarely offers a list of results that satisfies the question. Usually, there are a considerable number of results that are not spatially related to the search area. However, once the user has managed to find the metadata of a resource that may interest them, sometimes that metadata is not even useful to access the resource. There are scenarios where the user needs a specific format that the metadata says it offers, but when they review the different distributions of the metadata, none of them offers that format. Another common problem arises when the download URL of the resource is broken, or has changed; or is active but the site it points to is not the resource but a web page that sometimes leads to the resource and sometimes does not. In those cases, that distribution is useless because it does not give us access to the resource. In the worst case, if all distributions are useless, the metadata itself becomes useless as a means of accessing the resource.

We can find tangible examples of these problems in real open data catalogs. For instance, let us imagine a user interested in ornithology that is looking for bird data in Kansas City. To do this, they access the catalogue of the aforementioned official Open Data Portal of the USA, enters "birds" in the search bar, and selects the area of Kansas City to spatially filter the results. In the fourth position of the list of results (see Figure 1.2), they find a resource called "Atlantic Offshore Seabird Dataset Catalog, Atlantic Coast and Outer Continental Shelf, from 1938-01-01 to 2013-12-31 (NCEI Accession 0115356)"[10]. The title itself suggests that they found a case of an incorrect result of the spatial search that we mentioned before. To verify this, they accesses the resource and searches among the distributions of the resource to find a way to access it online or download it (see Figure 1.2). They follow the distribution hyperlink named "Dataset Landing Page" only to find that the link is broken (see Figure 1.3). Finally, using a generic search engine such as DuckDuckGo, they manage to find an active link that points to the resource, download the dataset, and open it in

[8] https://data.gov.au/home

[9] https://ckan.org/showcase

[10] https://catalog.data.gov/dataset/atlantic-offshore-seabird-dataset-catalog-atlantic-coast-and-

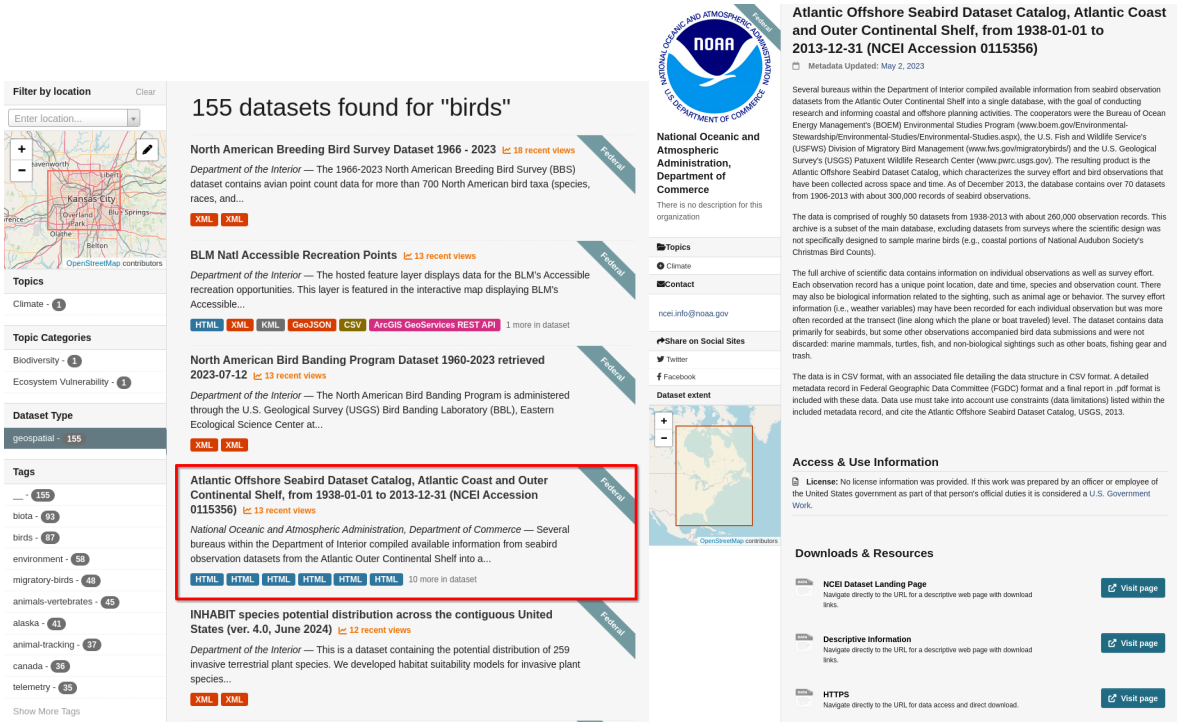

Figure 1.2: USA Open Data Portal: Results list of the search "birds" and resource page of "Atlantic Offshore Seabird Dataset Catalog"

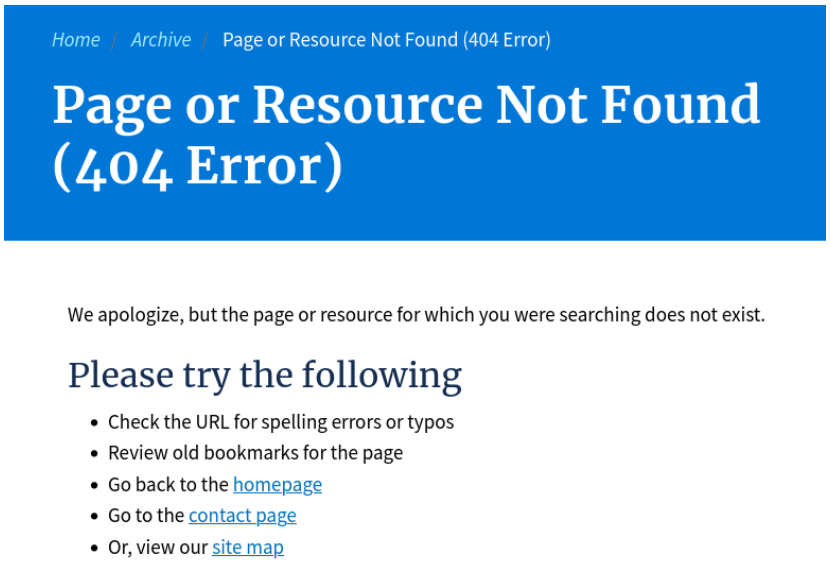

Figure 1.3: USA Open Data Portal: Result of accessing the download link of "Atlantic Offshore Seabird Dataset Catalog"

a spatial data viewer. To their surprise, the dataset does not contain a single piece of data remotely close to the area they had defined, as the name already suggested (see Figure 1.4).

1.2 Research Question and Methodology

Throughout the course of this research we sought to identify, study, and address the causes of the problems exposed in the previous section. This led us to the following research questions:

- **RQ1**: Do spatial metadata records provide an effective way of accessing spatial resources?
- **RQ2**: Are spatial search results relevant or suitable?
- **RQ3**: How can we improve the quality of spatial search results?

To answer these questions, we have designed a methodology backed by the realization of empirical studies. For two of the experiments, we analyzed the current state of the technology and the infrastuctures, and for the third one, we proposed and evaluated a novel methodology. Each experiment has followed the same structure:

- First, we identified the topic we wanted to study and measure and stablished the research questions.
- We carried a study on the state of the art of the topic and the current state of the research techniques.
- We designed the experiment and the metrics we wanted to measure following state-of-the-art techniques in the field of Information Retrieval and Geographic Information Retrieval when they were applicable and designed and proposed novel methodologies for problems not yet addressed by the literature.
- We collected real datasets and metadata records from real geospatial data catalogues and performed the data transformations required to conduct the experiments, in a controlled environment, seeking to maximize the reproducibility.
- We applied the metrics we designed and analyzed the results.
- Finally, we drew conclusions from the results and discussed the present and future implications of the findings.

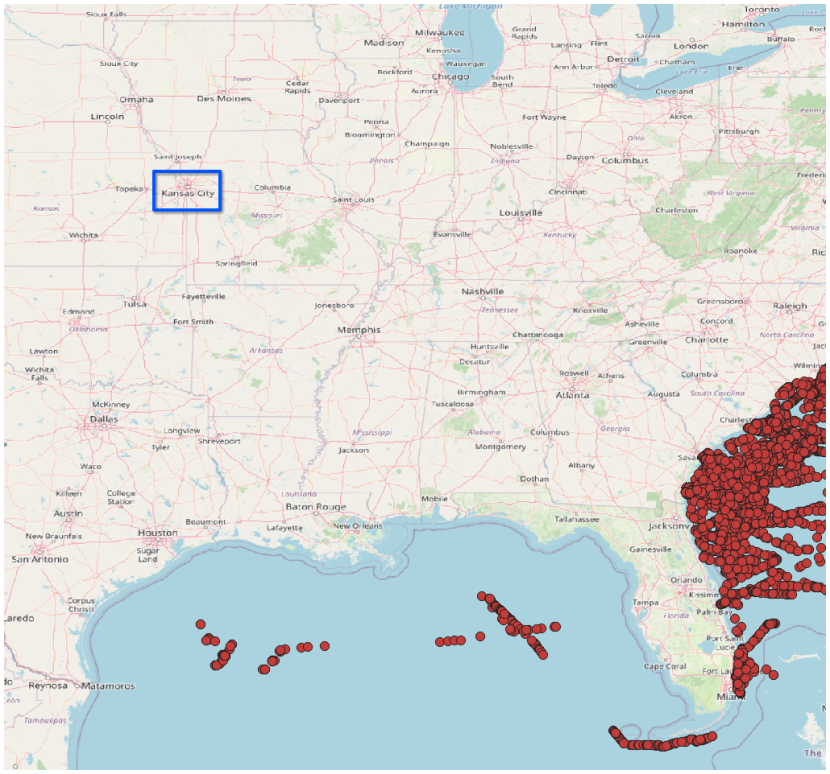

Figure 1.4: Contents of the downloaded dataset of "Atlantic Offshore Seabird Dataset Catalog". Data points are represented in red. The search area has been manually marked in blue for reference.

From a broader point of view, the two main high-level problems addressed, spatial search and access to resources, are strongly interrelated. Not only because both affect the same task: finding a spatial resource, but also because the repercussions of one aggravate those of the other. In a context where metadata does not offer effective access to the resources they describe, the task of automatic data collection becomes a challenge—as will be seen throughout this thesis—thus complicating the analysis of these phenomena. It also affects the available solutions: it is difficult to establish mechanisms that compensate for search problems if access to all resources cannot be obtained automatically.

Despite this interrelation, the first two research questions (RQ1, RQ2) could then be answered in parallel, studying on the one hand the phenomena of *Reference Rot* and on the other the precision of the search results. However, the study of search precision (RQ3) requires a collection of data that, in the case that our hypotheses are true, will be difficult to obtain. This is why we have decided to study the phenomena of *Reference Rot* first and reuse the results of this study for the study of the search results. With the results of both studies, we will be able to validate our hypotheses and propose an answer to the last research question (RQ3) and the problems found. Figure 1.5 depicts the situation we described in a visual and structured way, that also

serves as a map to navigate the structure of this thesis.

1.3 Thesis Structure

The structure of this thesis follows the chronological order of the studies conducted. Chapter 2 addresses the first research question (RQ1) related to the issue of access to resources, studying the phenomena of link rot and content drift on spatial metadata. We conduct an empirical study over 26 main european spatial catalogues, part of the INSPIRE Priority Datasets, to measure the presence of those phenomena. Chapter 3 addresses the second research question (RQ2) related to the issue of spatial search based on *Minimum Bounding Boxs*, studying the precision of these search result. We conduct an empirical study over real spatial resources extracted from the previous catalogues, comparing the results yielded by the MBB method vs their actual spatial extent. In Chapter 4, we address the last research question (RQ3) by studying the feasibility and advantages of a novel method to describe the extent of a spatial resource using DGGS tiles. In this section, we not only propose a new technique but also a roadmap for the adoption of this system. Finally, this thesis ends with a reflection on the work done in the form of conclusions in Chapter 5, as well as a proposal for future lines of continuation.

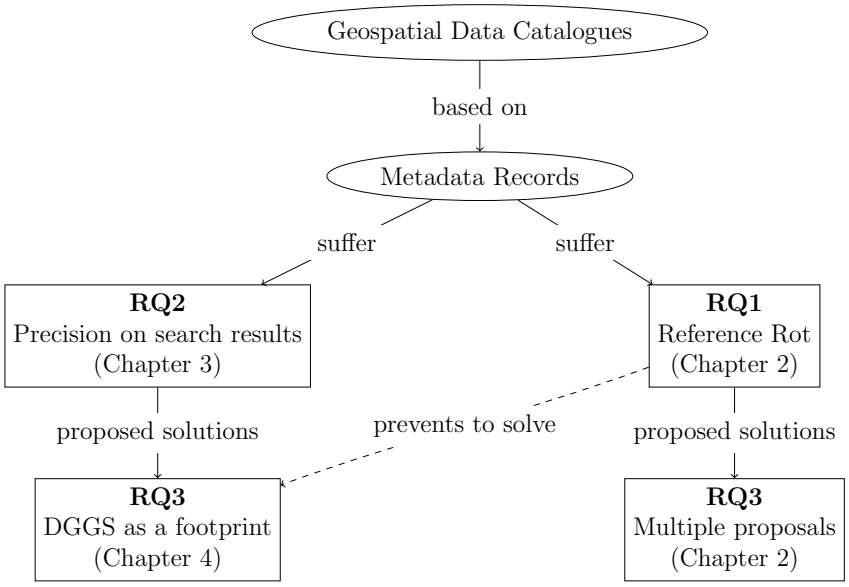

Figure 1.5: Research problems and their interrelations

Chapter 2

MeasURLng data access problems

2.1 Introduction

As stated in Chapter 1, an *Spatial Data Infrastructure* relies entirely on the spatial metadata records (Nebert, 2001) and the accessibility of a spatial resource depends on a chain of requirements. If data producers do not publish their metadata in these catalogues, users will not be able to locate any resource. However, if published metadata contain broken links in their distribution information or point to other undesired data, users will not be able to locate any resource. As there are no automatic checking mechanisms, the records rely on being properly curated and maintained (Tyler and McNeil, 2003). Therefore, spatial metadata quality has been studied from different perspectives and frameworks by maintainers, stakeholders, and researchers (Ureña-Cámara et al., 2019; Quarati et al., 2021).

The way the metadata records refer to their distribution locations, via distribution URLs (Uniform Resource Locator), meets the conditions outlined for being susceptible to suffering *Reference Rot*. *Reference Rot* (Klein et al., 2014) denotes the combination of two problems: *Link Rot* and *Content Drift*. *Link Rot* (or *Broken Link*) occurs when an URL no longer gives access to resource representations since the resource has been moved or deleted. For example, a metadata record that has an URL pointing to a web map service that was shut down years ago will suffer from *Link Rot*. *Content Drift* occurs when an URL returns resource representations that do not represent the resource that was intended to be referenced by that URL. For example, a metadata record that describes a Web Map Service whose URL now points or redirects to a different site or resource. *Content Drift* ranges from simple text corrections that change the meaning of a sentence to all kinds of updates to the resource.

Those two phenomena have concerned different academic communities including digital libraries managers and web experts. Previous *Link Rot* estimates vary dramatically across studies (i.e., 20% of Science, Technology, and Medicine

articles (Klein et al., 2014), 58% of web citations in Agricultural Library (Sife and Bernard, 2013) or 27% of *American Political Science Review* links (Gertler and Bullock, 2017)).

Studies about *Reference Rot* have focused on domains such as academic journal citations, legal texts and digital libraries. However, the extent of *Reference Rot* in geospatial metadata, to the best of our knowledge, has not been analyzed in the detail shown in this chapter.

The geospatial domain has its own peculiarities, which requires the development of a specific methodology for its analysis. All *Link Rot* studies mentioned before limit their scope to *basic broken hyperlink checking* without analyzing the content of the responses. This *naive approach* has two problems when applied to spatial metadata: (1) it does not check if the returned content matches the expectations declared in the metadata (*Content Drift*) and (2) when dealing with spatial web services, which require deeper understanding of the geospatial protocols, it faces *false positives* (an accessible resource link returns an `HTTP error` status) and *false negatives* (an inaccessible resource link does not return an `HTTP error` status but an `HTTP OK` status).

In this chapter, we will cover the following topics:

1. We propose a method to study the presence of *Reference Rot* in *Spatial Metadata Records* that considers the content of the linked resources to improve the *naive Link Rot* checking approach, and uses its type as an indicator of *Content Drift*. This method can be applied to other catalogues as well;

2. We have detected and measured the presence of *Reference Rot* in 18,054 metadata records and its 22,738 distribution URLs from 26 officially registered INSPIRE Discovery Services of European Union and European Free Trade Association countries;

3. We have identified a lack of good practices among the publishers implementing the ISO 19115 standard and the INSPIRE *Implementation Guidance* as one of the potential causes for *Content Drift*.

The rest of this chapter is organized as follows. First, we collect some related works in Section 2.2. In Section 2.3, we analyze the multiple dimensions of the problem and the challenges we will face. In Section 2.4, we describe the methodology we designed to detect *Reference Rot* in spatial catalogues along with the details of the experiment. In Section 2.5, we present the results of the execution over real catalogues with a brief analysis. In Section 2.6 we discuss these results. Finally, we expose the conclusions and future works in Section 2.7.

2.2 Related Work

Since the early days of the Web, researchers realized that *Link Rot* was one of its notorious problems (Ingham et al., 1996; Nielsen, 1998). Various studies have been conducted over time on different corpora: web, digital libraries metadata, academic electronic journals, legal documents, and so forth. (Harter, 1997; Koehler, 1999; Davis and Cohen, 2001; Casserly and Bird, 2003; Tyler and McNeil, 2003; Wren et al., 2006; Dimitrova and Bugeja, 2007; Sife and Bernard, 2013). First experiments reported different degrees of incidence of *Link Rot*, varying from 18.3% for URLs in dermatology journals from 1999 to 2004 (Wren et al., 2006), to 81% for three-year-old references in undergraduate term papers in 2000 (Davis and Cohen, 2001). The consensus at that time was that the half-life of a hyperlink is directly correlated to its age.

Other studies focused on the causes behind the *Link Rot* (Rhodes, 2002; Ingham et al., 1996) and concluded that the causes are: (1) the authors and metadata curators are not aware of the risks of link rot in their resources; (2) a lack of URL maintenance policies; and (3) a lack of synchronization between authors and metadata curators.

Rajabifard et al. (2009) pointed out in 2009 that creating and storing spatial datasets and their metadata separately creates two independent collections that had to be carefully managed and updated to keep them synchronized. At the same time, Olfat et al. (2010) highlighted the need for automatic methods for managing metadata due to the effort involved for administrations. However, these systems do not solve the lack of synchronization when the referenced resource is not managed by the catalogue owner.

The literature on *Content Drift* began long before the term was coined when studying the evolution and dynamics of the web content (Brewington and Cybenko, 2000; Cho and Garcia-Molina, 2000; Koehler, 2002; Fetterly et al., 2004; Ntoulas et al., 2004; Adar et al., 2009). One of the main findings was that some types of data are more likely to change than others. For example, HTML (HyperText Markup Language) pages change more frequently than PDF files.

The Hiberlink Project (Sanderson et al., 2013) introduced the term *Reference Rot* to aggregate these two phenomena that affect the availability of linked resources: *Link Rot* and *Content Drift*. Researchers associated with the project continued the studies with the new terminology (Klein et al., 2014; Burnhill et al., 2015).

The closest approach for measURLng *Reference Rot* in open metadata is found in *non-spatial metadata quality assessment* studies and systems. Methodologies and frameworks, such as *Open Data Portal Watch (ODPW)* by Neumaier et al. (2016), the *Metadata Quality Assessment (MQA)* by the European Data Portal (European Data Portal, 2021), and the Dataset-Service Linkage Service by INSPIRE (INSPIRE, 2020).

Both ODPW and MQA work with general purpose metadata that follow the Data Catalogue Vocabulary (DCAT) schema (W3C, 2020), an RDF vocabulary designed to facilitate interoperability between Open Data catalogues published on the web. Which means that none of them work natively with any spatial metadata standard such as ISO 19115. ODPW limits its analysis to metadata correctness and conformance to the standard, so it does not perform any *Reference Rot* analysis. MQA does use a simple *naive* HTTP request to check *Link Rot* from the status of the distribution hyperlinks, so it suffers from the limitations described in Section 2.3. The aforementioned INSPIRE service aims to establish a relationship between the metadata of datasets and the services that serve the same content. Unlike the previous ones, this system works with ISO 19115 metadata. The process needs to access the resource locations. Therefore, it finds which links are broken.

Nogueras-Iso et al. (2021) conducted a study similar to this one in which they analyzed the general purpose (non-spatial) Open Data Portal of the Spanish Government, with 22,406 records and 112,874 distributions. In this study, they performed a *naive* detection of broken links and a basic comparison of declared and obtained data types, based only on the file extensions of the resources. They found that 8.21% of analyzed URLs were broken and only 52.61% of the resources matched their declared type.

2.3 Reference Rot in Geospatial Metadata

The ISO 19115 data model for describing metadata distributions allows a resource to have zero, one or many distribution URLs. Distributions contain information about its distributors, its online locations, and its formats. The online locations are the place where the access URLs are declared. The formats define the expected data types or protocols in which the resource will be served.

Distribution Link Rot happens when the URL included in a specific distribution cannot retrieve any content. This manifests as a *connection error* (i.e., `invalid URL` or `connection timeout`) or an *HTTP error status code* (`4XX` or `5XX`) and implies that the consumer will not be able to retrieve the resource from that specific distribution URL. Overlooking *Link Rot* leads not only to a detriment in the usefulness of the distribution URLs but also of the metadata itself. *Metadata Link Rot* happens when the metadata only contains broken links, and the consumer has lost all chances of obtaining the resource in any way. This *broken metadata record* may have some historical or archival value, but it is useless for data sharing.

A *naive* approach of using a simple HTTP request for detecting *Link Rot* may

be enough for checking direct download links, but it may report *false positives* when the linked resource is a web service endpoint that requires some specific protocol. For example, OGC web services always require a set of mandatory parameters that are sometimes not included in the distribution URLs. The INSPIRE *Implementation Guidance* suggests including full URLs with all the needed parameters, such as `GetCapabilities` as a method. Despite that, we do not consider the cases that do not follow this as broken links if they point to a working and accessible service endpoint. In these cases, a *naive* HTTP request approach will wrongly report *Link Rot*. Knowing the protocol allows us to create a valid URL to test the availability of the service again.

Besides, as and OGC specification does not enforce the implementation of appropriate HTTP response status codes, some implementations use the `HTTP OK` status (`200`) for error responses too (i.e., service error, not available, required arguments, etc.). There are also hyperlinks that return an empty page with an `HTTP OK` status. We consider this scenario invalid as they are not serving any content. In both cases, a *naive* HTTP request only based on HTTP status code will not detect the *Link Rot*. In this study, we will consider these scenarios as a category of interest called *Wrong OK status* so we can analyze its presence as a special type of *Link Rot*.

Content Drift happens when the resource retrieved by the URL does not match the expected/declared one. This mismatch may be semantic (the content is not the expected, changing its meaning) or syntactic (the content is not presented in the expected manner, changing how we consume it). In this study, we will focus on the syntactic mismatches, using the data format as an indicator for measURLng this fenomenon. This way, we detect *Distribution Content Drift* as a mismatch between the expected resource format and the real one found in the URL. *Metadata Content Drift* happens when none of the declared formats is found on any of the distributions.

ISO 19115 is a flexible standard which allows a high degree of freedom by design but has some difficulties for establishing the expectations about the distributions formats. Even though the *Spatial Data Infrastructures* like INSPIRE suggest implementation restrictions in its *Implementation Guidance*, metadata publishers do not always follow the best practices which makes the automatic understanding of the metadata records more difficult.

First, the standard by itself does not enforce any controlled vocabulary such as MIME (Multipurpose Internet Mail Extensions) Types (Freed et al., 1996) for declaring the formats. This means that the publishers are free to populate that information however they want, making it difficult to automatically identify. This makes the *Content Drift* analysis hard, but also prevents the metadata from being useful in scenarios where the user wants to search records in a specific format because it cannot

filter by any keyword. The INSPIRE *Implementation Guidance* recommends using an `gmx:Anchor` tag to declare the encoding format using a controlled vocabulary but most publishers prefer to use the free text field. Besides, the relationship between the distributions and their expected formats is not enforced by any means. The standard does not propose any kind of *"URL-to-Format"* relationship mechanism. The number of declared types does not even have to match the number of distributions. This prevents us from expecting any specific type from a specific distribution. Finally, the data retrieved from the URL may not match the declared format when the URL points to an intermediate medium, such as a web page or a feed. This is a common practice in many public catalogues, spatial or not.

2.4 Materials and Methods

To measure the presence of *Reference Rot* in Spatial Metadata Records and its distribution links, we examined the records obtained from different Spatial Data Catalogues. Specifically, we focused on the following questions:

1. What is the percentage of metadata records with curation issues related to *Reference Rot*?

2. What is the percentage of spatial resources inaccessible using their metadata records due to *Link Rot*?

3. What is the percentage of spatial resources accessible using metadata records with misleading format descriptions due to *Content Drift*?

4. What is the percentage of spatial resources with only *indirect access* (accessible through intermediate third-party web sites)?

The ISO 19115 metadata records were harvested from service catalogues implementing the OGC CSW (Catalog Services for the Web) standard (Nebert et al., 2007). The whole process involved the following steps:

1. *Link extraction.* Identify URLs that may give access to the reference resource in the metadata record.

2. *Format extraction.* Identify distribution formats for returned representations according to the metadata record.

3. *Request phase.* Perform HTTP requests using the extracted URLs and produce a preliminary estimate of *Link Rot*.

4. *Type Guessing phase.* Analyze the successful HTTP responses, guess the format of the returned representations, and produce a preliminary estimate of *Content Drift*.

5. *False positives and false negatives removal.* Identify potential *Reference Rot false positives* and *false negatives* and manage them properly.

6. *Metadata Reference Rot assessment.* Evaluate the *Reference Rot* at *Metadata level* considering the *Link Rot* and *Content Drift* of all distributions.

7. *Indirect Access Resource.* Evaluate how many resources can only be accessed indirectly.

2.4.1 Metadata Harvesting

The first phase of the process began by obtaining the metadata records that would be analyzed. These records were extracted from catalogues offering an OGC CSW endpoint by using the operation `GetRecords`. This method is mandatory in OGC CSW compliant catalogues. This operation allowed us to harvest all the metadata records in a given metadata schema. The requested output format was ISO 19115 XML (eXtensible Markup Language). The retrieved metadata records were stored for further processing.

2.4.2 Link and Format Extraction

Stored records were parsed to extract each potential distribution URL, all declared distribution formats, and the date the metadata record was created.

The declared formats were located in the `gmd:distributionFormat` and `gmx:Anchor` nodes. The `gmd:distributionFormat` contained a free text description while `gmx:Anchor` contained an URL describing a controlled data type or data model specification. The date, used to verify that the documents were recent and still relevant, was found in `gmd:dateStamp`.

In order to fix the lack of a standard vocabulary to describe the distribution formats, we developed a list of well-known synonyms and aliases for popular formats. In this manner, we normalized them to a common internal limited list of keywords. For example, `ogc wms`, `web map service`, and `ogc:wms` would all be mapped to `wms`. This list is based on the most common keywords found in the metadata records. The process of obtaining the list used in the experiment below is detailed in Section 2.4.5.

Some metadata records use keywords such as `n.d.` or `unknown` as a "declared" type. We considered this to be equivalent to not declaring anything, so we ignored

them, and in the cases where they are the only "declared" type, we considered this metadata as if it had not declared anything.

2.4.3 Request Phase

In this phase, we performed an HTTP request to each extracted distribution URL. Once the URL access was completed, the *request status* and the *response body* were stored for further analysis. In *request status* we stored a specific code for each identified URL syntactic problem, network failure or HTTP status code. The *response body* contains the HTTP raw response content. The results of this phase gave us a preliminary estimate of the number of distributions affected by *Link Rot* as connection errors and unsuccessful HTTP response status codes reveal potential *Link Rot*. Nevertheless, some errors may occur due to a temporary service failure. For this reason, URLs that failed due to network or server problems were given a two-day grace period to recover its service before a second attempt. This grace period was based on similar *Link Rot* studies mentioned in Section 2.2.

Spatial resource sizes may vary from a few kilobytes to hundreds of megabytes. In this phase, we decided to fetch only the first 5000 bytes of each response. This is because the type guessing tool that we used in the next phase is based on Magic Number recognition (Kessler, 2002). This technique only requires a fraction of the file to detect its file signature and we have found 5000 bytes enough for most cases. The details about how the type guessing tool works and how it is affected by this 5000 bytes limit is explained in Section 2.4.4 and Section 2.4.5.

2.4.4 Type Guessing Phase

In this phase, we analyzed the content of the HTTP responses for each distribution URL using the tool *Libmagic* [1] to guess its file format. Then we mapped the inferred file format to our controlled vocabulary (see Section 2.4.2) so we could compare it with the list of expected ones declared in the metadata.

Libmagic works with many supported data types (i.e., HTML, PDF, PNG). However, for detecting more specific spatial domain formats (i.e., GML, GeoJSON, OGC WMS Capabilities) we applied various strategies.

When the guessed format is XML, HTML, or plain text we first tried to parse them as XML and, if successful, we looked for specific *XML Nodes* and *XML Namespaces* that denote its type. We also tried to detect other known text patterns that denote other spatial formats such as GeoJSON. Finally, we tried to detect some common *OGC*

[1] `https://man7.org/linux/man-pages/man3/libmagic.3.html`

Error messages that most OGC Service implementations return. This is necessary and useful as explained in the next Section 2.4.5.

Compression algorithms like ZIP allowed us to decompress the first bytes of a file without having the whole content (stream decompression). This allowed *Libmagic* to detect the magic number of a compressed file even when partially downloaded. However, if the compressed archive contained more than one file, only the first ones could be detected. For example, we can detect within a ZIP file a compressed ESRI Shape File using only the first 5000 bytes if we decompress its content and find the header of a `.shp` file. However, if the ZIP file contains other attached contents, such as a PDF documentation, that were added before the Shape Files, the first bytes may only be enough to detect the attachments but not the spatial files. Besides that, we took advantage of the fact that ZIP includes the name of the files as plain text to look for specific file extensions in order to reinforce the type detection. In the cases where these strategies were not enough to detect the type of compressed files, we marked them as "special cases" and addressed them in the next phase (see Section 2.4.5).

The results of this phase can give us a preliminary estimation of how many distributions suffer from *Content Drift* when we compare them with the declared types extracted in Section 2.4.2. We can directly compare these keywords because we used the same controlled vocabulary.

2.4.5 Spatial Specific Cases

In Section 2.3 we explained how a *naive* approach to *Reference Rot* measURLng may report wrong results for Spatial Metadata Catalogues. In this phase, we explain the methods used to manage these situations.

Incomplete Service URLs

A common case of *Link Rot false positives* may happen when the distribution URL only contains an OGC Web Service endpoint, missing some of the mandatory parameters. As the OGC standard enforces implementation of at least a `GetCapabilities` function, we can build a `GetCapabities` request to check the URLs we previously detected as *OGC Errors* in the type guessing phase. Then, the new URLs were requested and guessed again.

Non-Matching Data Type Declarations

Content Drift false positives may happen when we cannot guarantee that there is an issue even though the types do not match (when have "undecidable content"). These cases must be individually identified so we can handle them correctly:

- Intermediate web HTML portals or Atom feeds. That is, Metadata declared `gml` and distribution URL returned an `html` web page that may (or may not) contain a direct link to the resource.

- Combinations of Web Services and compatible spatial data formats. That is, Metadata declared `wms` and distribution URL returned a `png`.

- The distribution URL returned a compressed file whose content type we could not guess. That is, distribution URL returned a compressed `gml` which was identified as `zip` instead of `gml`.

- Any distribution whose type was guessed as `xml`, but we could not specify the schema. This covers some marginal results like undetected OGC errors or other unsupported formats where we cannot assure that they were not the expected result.

We have designed a strategy to detect pairs of declared and guessed data types that may be correct. It is based on a list with a *target type* and their allowed potential *matching types*. Table 2.1 shows the available cases and their respective decisions. The method makes the decision of the first matching case, evaluating from top to bottom. It does not matter if the *target type* is the declared or the guessed one. That is, the same rule matches a declared `gml` with a guessed `wfs service` and a declared `wfs service` with a guessed `gml`, so the same decision will be made for both.

It is worth highlighting the inclusion of the pair `WMS-GML` as an undecidable case. Even though WMS is primarily an image service, it can serve GML via `GetFeatureInfo` method. We did not consider declaring a WMS Service with a GML format good practice, but we cannot say it is incorrect.

Wrong OK Status

We already mentioned that some implementations of OGC Services wrongly returned an `HTTP OK` status code even when the content suggests an error. We considered as *wrong OK status* the situations where an *OGC Error* had been detected in the type guessing phase; the HTTP status did not indicate the error; and the distribution had failed the *retry* with the new `GetCapabilities` URL too. We also considered as *wrong OK status* the `empty` responses that returned an `HTTP OK` status code.

2.4.6 Metadata Reference Rot Analysis

The previous analysis provides *Reference Rot* metrics per distribution URL. Nevertheless, each distribution represents a different way of obtaining the same resource

Target Type	Matching Type	Decision
Any format	Same format	ok, No content drift
wfs	gml	ok, no content drift
feed or html	Any format	No direct access
wms	jpeg, png, gif, tiff, svg, bmp, img, pdf, rss, kml or kmz	Undecidable (service detected)
wms	gml	Undecidable (service detected)
wfs	shp, geojson, kml or csv	Undecidable (service detected)
wcs	jpeg, png, gif, tiff, bmp, arcgrid	Undecidable (service detected)
compressed	May contain any format	Undecidable (compr. detected)
xml	wms, wfs, kml, gml or kmz	Undecidable (xml detected)
None declared	Any format	No expectations
"unknown"	Any format	No expectations
Any format	Different format	Content drift

Table 2.1: Decision Table.

described in the metadata. This implies that the analysis for the whole metadata record must consider not only the presence of individual issues, but also their joint effect.

- Regarding the *Link Rot*, a degraded metadata record with at least one valid distribution URL should still be able to somehow provide access to the described resource. The worst case scenario happens when the resource is completely lost because all its distribution URLs are broken.

- Regarding the *Content Drift*, the scenarios are much more diverse. From the usability and interoperability standpoints, a declared type that is not served in any distribution is far worse than a distribution whose type was not correctly declared. That is because an "extra" distribution whose type is not declared does not benefit from interoperability, but does not mess up any expectations either. On the other hand, when a type was declared, we expected to find at least one distribution serving that type. That is why we wanted to target the declared types that were not served by any of the accessible distribution URLs.

We classified each metadata record based on the combination of metadata-wide *Link Rot* and *Content Drift* analysis (see Figure 2.1). This allowed us to analyze the status of the records and obtain a single overview of any metadata collection. The main categories are:

- Resource Found: The record has at least one directly accessible URL and its type is correctly declared;

- No direct access: The record has at least one accessible URL but none of them provides direct access to the resource;

– More data needed: Covers any of the scenarios explained in Section 2.4.5 where the available information is not sufficient to give a robust answer about the status of the metadata;

– *Content Drift*: None of the declared types match the types found in the accessed URLs;

– No expectations: The record does not declare any data type;

– *Link Rot*: All URLs are broken;

– Without Links: The resource has no URLs.

Some categories are divided in subcategories to cover specific scenarios. The *Resource Found* category is divided based on how many declared formats are available:

– No metadata decay: All declared formats are available in accessible distributions;

– Some metadata decay: Some, but not all, declared formats are available in accessible distributions. It should be noted that this category cannot be applied to metadata records with only one distribution.

The *No direct access* category specifies if the intermediate medium is:

– Web page: An HTML page that may contain an URL to the resource;

– Web feed: An Atom or RSS that may contain an URL to the resource.

The *More data needed* category contains:

– Compressed file detected: Some distribution served a compressed file whose content could not be identified. This means that it may contain a resource that matches a declared type;

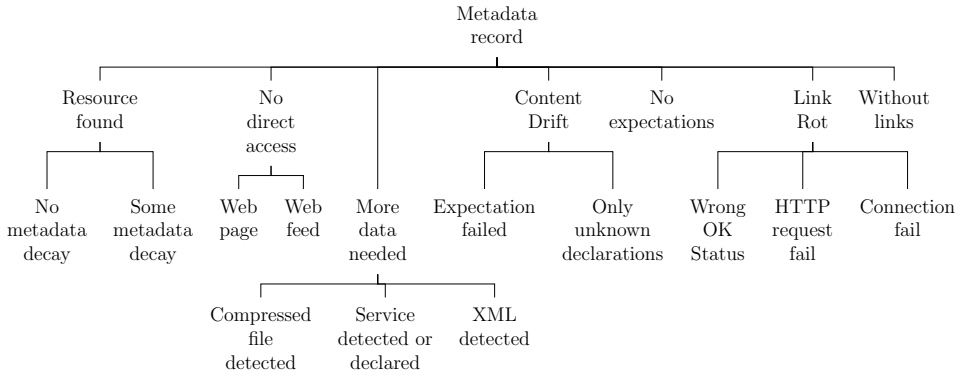

Figure 2.1: A tree displaying all *Metadata Reference Rot* Categories.

– Service detected or declared: Some *OGC Service* is declared, and a compatible data type is available in some distribution. It also includes the cases where a compatible data type is declared, and the *Service Capabilities* is available in a distribution;

– XML detected: Some distribution served an XML file whose schema could not be identified. This scenario covers the data types and schema that were not considered when designing the experiment.

The *Content Drift* category is divided based on the cause of the mismatch:

– Only Unknown Declarations: None of the declared formats could be identified. This happens when the text description is unrecognized or unclear;

– Expectation Failed: None of the accessible resources matches any of the identified declared formats.

The *Link Rot* category is divided based on the error types:

– HTTP request fail: All URLs are broken but some obtained an HTTP error response from the server;

– Connection fail: All URLs are broken and none of the succeed to connect to any server.

2.4.7 Experiment

For this experiment, we used 26 out of 35 officially registered INSPIRE Discovery Services of European Union and European Free Trade Association countries at INSPIRE Geoportal[2]. We chose these catalogues because they are curated and carefully maintained to comply with the INSPIRE Directive, so they are expected to be of high quality. Nevertheless, some of the listed catalogues were not included due to access problems or huge differences in metadata policies denoted by the size of the catalogue or some other practices. One example of these problematic cases is the Italian catalogue, which published as many metadata records as all the other catalogues together. It also applied some practices that dramatically distorted the results such as listing 7717 different metadata records that pointed to the same OGC WMS service URL and not declaring it correctly. From the 26 catalogues harvested, the analysis found 18,054 metadata records from different producers with a total of 22,738 different hyperlinks. The extraction process was executed between 1 September and 3 September 2021.

[2] `https://inspire-geoportal.ec.europa.eu/harvesting_status.html`

As each metadata record can have zero, one or many distribution URLs, we analyzed the number of distributions provided by each metadata record. The results can be seen in Table 2.2. The most common case is a metadata record (28.64%) with one distribution URL. Next, the second most common case is a metadata record (20.79%) with 7 distribution URLs. This is because the *Belgian Catalogue (Flanders)* has over 3500 metadata records with this characteristic. Then, we have 35.85% metadata records that have between 2 and 6; and 10.16% that have 8 or more. There are also outliers. For instance, a metadata record in the catalogue of *Luxembourg Catalogue* contains 422 distribution URLs. Finally, 4.56% of metadata records have zero distributions.

By analyzing the date of the records, we see that most of them are recent; 91.99% of records are less than 4 years old (2018 (3.58%), 2019 (7.55%), 2020 (21.68%), 2021 (59.19%)). Less than 1.70% of metadata records were created more than 10 years ago.

As explained in Section 2.4.2, to establish a comparison between types we have associated a set of *uncontrolled natural language type definitions* with a controlled keyword. By taking the most common keywords found, we achieved a great coverage of all cases: less than 1.5% of uncontrolled cases in type inference and 11.13% of uncontrolled cases (`OTHER`) in type declarations. Many of the uncontrolled types were not identified, not only because of the diversity of the keywords to express the same format, but also because of the generic or vague terms used. This issue could be solved if metadata publishers used the mechanism proposed by the *Implementation Guidance*. Some examples of confusing declarations:

– "*vettoriale*" used 649 times in the *Italian Catalogue*;

– *"aaa"* used 178 times in the *Danish Catalogue*;

– *"volgens afspraak"* (according to appointment) used 101 times in the *Danish*

Distribution Count	Count	Rate
0	823	4.56%
1	5171	28.64%
2	3243	17.96%
3	1506	8.34%
4	832	4.61%
5	556	3.08%
6	335	1.86%
7	3754	20.79%
+7	1834	10.16%

Table 2.2: Distribution count on metadata records.

Catalogue;

- *"online"* used 79 times in the *Austrian Catalogue*;
- In the *British Catalogue*: *"geographic information system"* (44 times), *"paper"* (32 times) and *"digital"* (32 times). In total, the catalogue has more than 500 different text declarations.

2.5 Results

This sections is divided into subsections for each different metric studied, covering bot link rot and content drift related measurements. This includes statistical overviews over the collected data and its nature.

2.5.1 Link Rot

Table 2.3 shows the detailed response status code count for each unique distribution URL. By unique, we mean that a URL that appears in two different metadata records is not counted twice. This reveals that only 89.59% of the distributions are accessible, while the remaining 10.41% suffers from *Link Rot*. The most common `HTTP errors` are `404 Not Found` and `500 Server Error`. The most common `non-HTTP errors` are `Connection Error` and `Read Timeout`. 1.39% of the URLs returned an `HTTP OK` status code while the content of the resource suggests an OGC Error (see Section 2.4.5).

As metadata records may have more than one distribution, we need to study how the records are affected by *Link Rot*. The results show that only 74.84% of records have all its links accessible, while the other 14.3% have some of them broken. The remaining 10.86% do not provide access to any resource because: (1) 6.30% have all its URLs were broken (5.37%) or had *wrong OK status* (0.93%), (2) 4.56% have no distributions. Those percentages would be even higher if we decide to exclude the 5% records without distribution links from the calculation.

2.5.2 Resource Types

Table 2.4 shows the distribution types obtained in the type guessing process. It only counts the resources that received an `HTTP OK` status code, even though the process analyzed all responses for detecting *false positives* (a total of 20,687 resources). The second most common family is "Intermediate page" (20%), which represents HTML pages as explained in Section 2.3. The "Undecidable" category consists of *unguessable* compressed resources (1.72%), XML files with uncontrolled schema (0.21%) and other unrecognized files (0.01%). The *wrong OK status* category adds up to 1.52% of the

Code Type	Status Family	Status	Count	Ratio
Non-HTTP Errors (4.53%)	URL Error (0.13%)	Invalid URL	13	0.06%
		Invalid Schema	15	0.07%
	Connection Errors (4.38%)	Connect Timeout	29	0.13%
		Read Timeout	717	3.15%
		Connection Error	250	1.10%
	Other	Other connection exceptions	6	0.03%
HTTP Errors (4.49%)	5XX — Server Error (0.97%)	504 — Gateway Timeout	1	0.00%
		503 — Service Unavailable	18	0.08%
		502 — Bad Gateway	13	0.06%
		500 — Internal Server Error	188	0.83%
	4XX - Client Error (3.52%)	499 — *NGINX non-official error*	10	0.04%
		410 — Gone	74	0.33%
		406 — Not acceptable	2	0.01%
		405 — Method not allowed	2	0.01%
		404 — Not Found	553	2.43%
		403 — Forbidden	33	0.15%
		401 — Unauthorized	49	0.22%
		400 — Bad Request	78	0.34%
HTTP OK (90.98%)	2XX — OK	200 - Success	20372	89.59%
	Wrong OK	*Wrong OK status*	315	1.39%

Table 2.3: Distribution URL status.

guessed records combining the `OGC Errors` and the few `empty responses`. The rest of the data types are, as expected, spatial datasets and services.

2.5.3 Reference Rot Presence Over Time

Figure 2.2 shows the presence of *Link Rot* and *Content Drift* over time, based on the dates extracted from the metadata. The identified categories are the following:

– All *Link Rot*: The records have only broken distributions (including *wrong OK status*);

Type Family	Ratio
Vector data	35.02%
Intermediate page	20.52%
Portrayal service	18.14%
Download service	13.96%
Raster data and Coverage	4.59%
Document	3.19%
Undecidable	1.73%
Wrong OK status	1.52%
Process service	0.91%
Geodatabase	0.42%

Table 2.4: Resource types (overview).

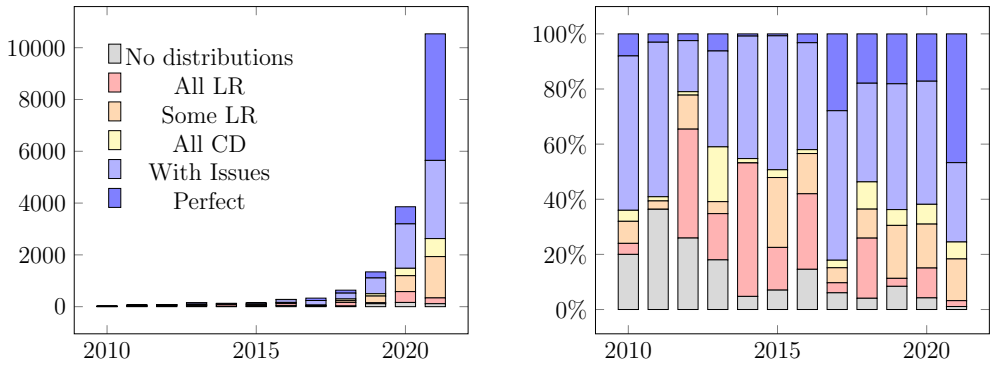

Figure 2.2: *Reference Rot* presence by year

– Some *Link Rot*: The records have some broken distributions (including *wrong OK status*);

– All *Content Drift*: All distributions are accessible, but none of its declared types are served;

– With Issues: Some of the declared types are not served, or the records have some undecidable distributions;

– Perfect: All distribution URLs are accessible, and all their declared types are served.

We can see that even the most recent ones have a considerable percentage of *Reference Rot* issues. If we see the evolution of the *Link Rot* in the past 4 years, the records with only broken links have decreased from 21.82% in 2018 to 2.13% in 2021. The overall *Link Rot* presence also decreased form from 32.3% in 2018 to 17.27% in 2021. This agrees with the related work that indicated that *Link Rot* risk is related to age of the document. Another interesting conclusion is that the records from 2021 have dramatically improved its accessibility, as 46.72% of them have neither broken links nor wrongly declared types (*perfect* records), compared with the rest of the years where they have never surpassed 27.88%. We can also see a growing trend in the number of records offering distribution URLs. The number of records with no distributions drops from 20% in 2010 to 1.08% in 2021. Finally, it is worth noticing that, in 2017, there are exceptionally good results. This may be explained since there are only 330 records from 9 catalogues from this year.

2.5.4 Metadata Wide Reference Rot

Figure 2.3 shows the metadata-wide *Reference Rot* categories we defined previously in Section 2.4.6.

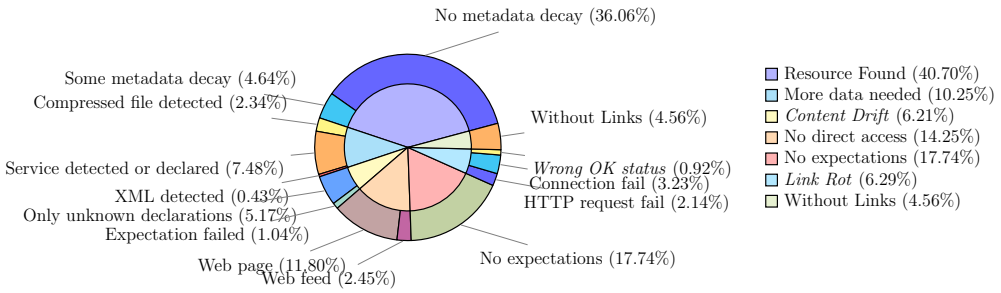

Figure 2.3: Metadata-wide *Reference Rot*

We can state that 40.70% of the metadata records offer some of their declared resource types, with 36.06% offering all of them. This means that they have the highest degree of interoperability and even an autonomous user agent can discover and access these resources in the expected format just from the metadata record.

If we consider the metadata that may provide indirect access (14.25%) and the ones that need more data to determine if they provide their resources (10.25%), which both together add up to 24.5%, the best case scenario rises to a potential of 65.2% found records. This means that only around 65% of the metadata records may provide at least one of their declared types. Regarding the other 35%, 4.56% of the records did not include any distribution hyperlink at all. This suggests that the purpose of the metadata are not the public distribution of the resource. Other 17.74% of the records did not declare any data types but offered at least one accessible hyperlink. This makes it impossible to expect anything about the resources type. The remaining 12.5% records belong to the proper *Link Rot* (6.29%) and *Content Drift* (6.21%) phenomena extensively described in this chapter. Of the 6.29% of metadata records with all their URLs broken: (1) 2.14% have at least one link that succeed to connect to a server; (2) 3.23% have no links that succeed to connect to anything; and (3) 0.92% have at least one link that received an `HTTP OK` status but whose content reveals a *wrong OK status*. Of that 6.21% of metadata records affected by *Content Drift*: (1) 1.04% declared different types than the ones served; and (2) 5.17% declared them in such a way that we could not identify them:

– The *Lettish Catalogue* and the *Belgian Catalogue (Wallonia)* barely declared any data types (81.25% and 99.05% of records, respectively). This prevents us from having any expectations and suggests a lower interest on *open data sharing* than others.

– The *Bulgarian Catalogue* has 81.25% of records with no declared type because most of the metadata records declare `unknown` as the only distribution data type.

- Some catalogues such as the *Romanian Catalogue*, the *Belgian Catalogue (Federal)* and the *Swiss Catalogue* contain between 23% and 36% of records with zero distribution URLs, which suggests a lower interest on *data sharing* too.

- The catalogues with more *broken records* are the *Belgian Catalogue (Brussels)* and the *Polish Catalogue* with 28.30% and 36.41%, respectively.

- The *Liechtensteiner Catalogue* has the highest percentage of indirect access records (80%). This is because it only contains 20 records, and 14 out of those 20 reference available HTML pages. However, other catalogues like the *Irish Catalogue*, the *British Catalogue* or the *Swiss Catalogue* also have a high percentage of non-directly accessible resources (between 43% and 45%).

The catalogues that have the highest percentages of well declared and accessible resources usually follow the same patterns in most of their records.

- The *Belgian Catalogue (Flanders)*, being the biggest collection, has around 4300 (of its 6648) records leading to `wfs` services declared as `gml`;

- The *Lithuanian Catalogue* mostly uses three resource types: `wms`, `shp` and `gml`;

- The *Greek Catalogue* only contains 80 records: All the accessible ones were `wfs` services serving `gml`.

2.5.5 Indirect Access

Section 2.5.4 showed that at least 12% of the metadata records may only provide indirect access to their resources thought an HTML Web Page. As each metadata record can only receive one category, the percentage may be even higher.

We have studied specifically the presence of HTML web pages as distributions in metadata. The results show that (1) 14.94% only point to HTML resources; (2) 16.77% point to one or more HTML resources but also point to resources of a different type; (3) 63.73% do not point to any HTML resource (this includes the 6.30% with *metadata Link Rot* from Section 2.5.1), and (4) 4.56% have no distributions. We can see that 31.37% of the metadata records have at least one HTML resource linked. This situation may be sufficient for a human, but an autonomous user agent needs a more sophisticated logic to browse those HTML pages and find the desired spatial resource (if available).

2.6 Discussion

The *Spatial Data Infrastructure* literature always highlights that a *Spatial Catalogue* is an essential component for discovering and sharing datasets and services. In this chapter we have spotted some issues that question the usefulness of the current status of catalogues for discovering and accessing the spatial resources they describe.

We have found that metadata affected by *Link Rot* cannot give access to its resource. This implies that the catalogues are advertising resources that they cannot provide. Even in well curated catalogues with recent records, such as the ones analyzed in this chapter, more than 10% of the distribution URLs were broken, resulting in more than 6% of the metadata records having completely lost access to its resource. The amount of *Link Rot* presence in the records of last years (2020–2021) suggests that the average life of some resources is shorter than what we expected. We also appreciate a growing trend in *Link Rot* as the metadata gets older. However, we cannot confirm that with a single time analysis (see Section 2.7 for further details).

The results show that *naive Link Rot checking* is not enough due to the nature of the spatial services reporting *false positives* and *false negatives* when the content is not considered. The *false positives* could be fixed by using the full `GetCapabilities` URL instead of just the endpoint as suggested in the *Implementation Guidance*. To fix the root cause of *false negatives*, the affected OGC Service implementations should make use of the appropriate HTTP response status codes. Even when they are not violating the OGC specification, they are technically incorrect by *standard composition* as they work over HTTP protocol too.

It is also interesting to note the fact that 17% of the records did not declare any resource type. This suggests that the publishers are not aware of its usefulness for discovery purposes. Even if we assume that the 24.5% of `No direct access` and `More data needed` records were declared correctly, it leaves us with 6.21% of records with no match or wrongly declared types. Issues like these do not prevent access to resources, but they may affect how they are consumed. This effect is more notorious when the user agent trying to access the resource is not a human but an autonomous system such as a crawler.

About one third of the metadata records contained at least one HTML page as indirect distribution medium while 17% have them exclusively. This extra layer of indirection implies that the consumer must browse and discover the effective distribution URL (sometimes this is difficult when there is a lack of context). It also hides the final distribution URL status, so it may report *Link Rot false negatives* when the resource is down but the page is up. Whether the link in the metadata link or

the link in the intermediate medium fails, the link will be broken. It also supposes an accessibility barrier for non-expert humans who access the catalogue and automatic user agents.

In Section 2.3 we explained the way ISO 19115 declares its distributions. We pointed that the freedom in its data model combined with the lack of good practices among metadata publishers lead to an unpleasant experience when trying to discover, find and access spatial resources. In Section 2.4.5, we saw the lack of consistent type declarations among some published data. Declaring resource data types without a standard vocabulary makes it difficult to search or filter resources in a specific format. In addition, not declaring the format of each individual distribution makes it impossible to determine which is the one we want or to assert that the content meets the expectations.

The Go FAIR principle A1 (GO FAIR, 2021) describes "components involving manual human intervention" as one of the accessibility barriers that an open service should avoid unless strictly necessary (in cases regarding confidential information). Intermediate mediums fit this description. They also highlight that, even when a resource is not freely accessible, it is desirable that "a machine can automatically understand the requirements, and then either automatically execute the requirements or alert the user to the requirements".

The break down results showed that each organization interprets their own rules and applies their own policies. A minimum degree of diversity is positive because it allows each institution to adapt its own workflows, but an excess dramatically impacts the data interoperability. When we compare the ISO 19115 standard with other metadata standards, such as DCAT vocabulary, we can see how they made this information more explicit. DCAT uses different distribution fields to identify whether the URL points to a service (`dcat:accessService`), a direct link to a dataset (`dcat:downloadURL`), or a link to an intermediate portal or web form that gives access to the resource (`dcat:accessURL`). This helps to establish a solid expectation about the outcome of the distribution URL and the way they are intended to be accessed. However, even using the model of DCAT, publishers still apply their own practices ignoring the guidelines (Nogueras-Iso et al., 2021). We consider the effort worthwhile as it would dramatically increase the resources' accessibility while facilitating *Reference Rot* verification.

The reality of the web demonstrates that hyperlinks are never persistent. The spatial catalogues, as *document-centric* systems, suffer from the same issues. Once a resource is published or updated, there is no mechanism that enforces anyone to register or notify the update to the catalogue. It is utopian to assume that metadata

authors will always be willing (or will be able) to maintain their metadata over time. To guarantee future availability, we need to be aware of that risk and adopt some measurements.

One of the simplest but most effective solutions proposed to prevent *Link Rot* is to do periodic link checking (Tyler and McNeil, 2003). This approach is interesting when the checking process is performed by the metadata owner so they can fix any issues the moment they are detected. It also benefits from metadata records that facilitates automatic checking.

Historically, other authors proposed architectures to prevent *Link Rot* on the Web, such as W3Objects (Ingham et al., 1996) or Hyper-G (Andrews et al., 1996), which tried to maintain referential integrity in ultra-large-scale web-based systems.

Systems such as Handle (Sun et al., 2003) and its subsystem DOI (ISO Central Secretary, 2012) have taken the approach of giving persistent identifiers (PID) to resources and providing resolving systems to locate them. Other authors such as Klump et al. (2016) discussed the relevance of giving DOI (Digital Object identifier) to geoscience data. An advantage of PIDs is that they are compatible with the architecture and the structure of the World Wide Web and can help to resolve the *Link Rot* problem. The only requirement is the availability of a resolution system.

All the methods mentioned above aim to solve *Reference Rot* for immutable resources. Several web archival systems have emerged to avoid *Link Rot* when web resources are deleted and *Content Drift* when web resources evolve. We find good examples in projects like the Wayback Machine of The Internet Archive (The Internet Archive, 2001) (nowadays, the largest web pages' snapshot archive), The Memento Protocol (Van de Sompel et al., 2013) (a protocol for accessing web page snapshots compatible with the Wayback Machine, among others) and WebCite (WebCite Consortium, 1998) (focused on archiving academic related material). Some of these systems are based on web crawlers while others rely entirely on the user requests. However, many of these systems are not widely used, so relying on them, as third-party systems, may not be the best solution.

2.7 Conclusions

In this chapter, we have developed a methodology for detecting *Reference Rot* in *Spatial Metadata Catalogues* that considers the content of the linked resources to improve the *naive Link Rot* checking approach, and uses its type as an indicator of *Content Drift*. We have applied this method over 26 officially registered INSPIRE Discovery Services. We have shown that the distribution URLs of spatial metadata records, even in well

curated metadata collections, are affected by *Reference Rot*.

The presence of *Reference Rot* in the analyzed corpus suggests that it is necessary to implement quality systems to prevent link decay. Automatic systems like the one implemented in the European Data Portal, which uses the MQA methodology, may be a good reference. However, we need to extend them to the spatial metadata domain and its peculiarities. We could also use search tools to try to locate lost resources that have been moved. Nevertheless, this work focuses more on detecting and notifying any issue to metadata owners than automatically recovering from existing problems.

This leads to the second conclusion. Publishers need to make a greater effort to follow the best practices and guidelines. The experiment has faced multiple challenges such as identifying and interpreting the declared types or detecting incomplete OWS (OGC Web Services) service URLs. This reveals gaps in the usefulness of current metadata for tasks beyond description and management, such as discovery and access to resources.

Since the used catalogues are the INSPIRE official Discovery Services, they are expected to have the best quality among all the available ones. Therefore, the identified problems can be seen as general issues affecting the spatial data access. Further studies may perform this analysis over larger and less curated catalogues to compare the results, expecting a lower quality. We have limited the analysis to a static snapshot of the metadata and its resources, taken on 1 September and 3 September 2021. Hence, we do not aim to study the evolution of *Reference Rot* in a set lapse of time but its presence in a specific moment. Future works may also consider the temporal perspective of the *Content Drift* and the evolution of *Link Rot* over time. The *Content Drift* has been evaluated by using the data type as the only indicator. Therefore, any other mismatch between the metadata and the resource, such as dates, spatial data extent and so forth, cannot be not detected. Future works may check more specific features such as: (1) if data are inside the declared bound box; or (2) if all distributions represent the same spatial dataset. We only fetched the partial content of the HTTP responses (the reasons were detailed in Section 2.4.3). This implies that the tool we developed to guess the data types may fail with some specific compressed files where the whole response body is needed to identify its content (more details in Section 2.4.4). Repeating the experiment fetching the whole response contents would increase the storage and time requirements but also the quality of the type guessing. This work could result in a more mature and robust *spatial data type guessing* tool that could be published as a standalone project.

Chapter 3

Measuring the precision of spatial search results

3.1 Introduction

Geographic Information Retrieval covers the process of providing access to relevant information resources in response to queries with geographic constraints (Larson, 2009). Its goal is to retrieve documents thematically and geographically related to a given query (defined as 'concept at location' by Hubner et al. (2004)) where the spatial context is defined by place names and references or by coordinate-encoded locations. GIR methods are used within both specialized systems such as Geospatial Data Catalogues (GDC), and general systems that encompass any form of spatial information, including Google, Facebook, Uber, among others.

In a catalogue, the query 'concept at location' implies two components, a thematic query component that compares the free-text query against the textual contents of the documents and a location query component that evaluates a user-defined query area against the spatial extent of the resource. The thematic component is covered by traditional *Information Retrieval* techniques, while the location is typically computed relying on the Minimum Bounding Box (MBB) (Nebert, 2001).

The MBB is defined as 'a rectangle, oriented to the x and y axes, which bounds a geographic feature or a geographic dataset. It is specified by two coordinates: xmin, ymin and xmax, ymax' (Caldwell, 2005). All commonly used standards in metadata catalogues such as ISO 19115 (ISO Central Secretary, 2014), GeoDCAT (European Union, 2016) or Dublin Core (DCMI Usage Board, 2020) propose the use of MBB as one of the most simple mechanisms to describe the spatial extent. The metadata models of these standards allow the use of more complex geometries or alternative geographic identifiers but the majority of catalogue tools provide spatial searches based on MBBs. One of these reasons is its ease of use when performing computations. Hence, catalogues

compute the spatial component of a query as the existence of an intersection between the metadata MBB and the user query area (QA).

However, simplifying the extent representation of a dataset can increase the risk of considering a dataset relevant when it is not, due to the overestimation of its area of interest. As the relevance is being computed only over the MBB, results may be considered relevant even when their actual spatial extent is not related to the query area. This causes the user to waste time trying to use (visit/download) a resource that is ultimately not useful to them, having to repeat the process again with an alternative resource. This ultimately leads to a loss of trust in the catalogue. Figure 3.1 shows an example of how a real dataset from a Spanish catalogue yields a false positive for a MBB-based search. Additionally, catalogues frequently contain metadata records that either do not define their resource locator or declare it incorrectly. This prevents access to the original resources, making it very difficult to verify the accuracy in the MBB describing the spatial extent of the described resources.

In this chapter, we address this issue and demonstrate that MBB representations present precision issues on searches across all catalogues. For doing that, we generated a test data collection with real geospatial metadata records and its datasets extracted from multiple curated European catalogues. Then, we applied a set of spatial queries over the metadata records and checked the results under the relevance criteria we defined, evaluating the results against their real datasets to find false positives. Questioning the decision of most GDC for still using dichotomic spatial retrieval criteria is outside the scope of this chapter. We just focused on the impact of the MBB simplification.

The remains of this chapter are structured as follows. Section 3.2 reviews the most relevant studies in the field of GIR and GDC. Section 3.4 describes the method we used to study the precision of the MBB-based retrieval systems. Section 3.5 presents the results of the experimens. Section 3.6 discusses the results and the implications of the MBB precision issue.

3.2 Related Work

The problem of precise spatial information retrieval has been addressed in multiple fashions. A document may have a spatial context encoded as textual references to places (books, articles, web pages) on digital libraries or from coordinate-based encoded locations (datasets, web services) in spatial data catalogues. These two formats require different approaches and technologies, but both share enough challenges to be covered by the geographical information retrieval (GIR) literature.

The study of GIR covers the whole process from extracting the spatial context to indexing and retrieving the relevant documents. During the last years the works have manly focused on the first step, proposing modern techniques for extracting and computing the spatial context from location names in textual and natural language corpora such as libraries and metadata collections using ontologies, keywords and natural language processing techniques. (Chen et al., 2020, 2022; Sharma et al., 2022). Hence, the study about the retrieval solutions and their performance has lately paced down. This work retakes the most relevant studies in this field to conduct an updated approach.

The SPIRIT (Purves et al., 2007) and the GeoCLEF (Gey et al., 2006) projects have been the reference in terms of GIR test document collections. However, both were text-based collections, not spatial data collections. Works like Frontiera et al. (2008) pointed to the lack of coordinate-based datasets as a severe problem for the study of coordinate-based GIR techniques.

The relevance judgments in most GIR systems, and particularly in catalogues, based on the MBB representation already mentioned before. However, those have been yet criticized and alternatives have been studied. Caldwell (2005) presents the most common issues that bounding boxes may suffer. Those include dramatic size growth due to reprojection to another datum or big holes and gaps due to the geometry of the shape. They propose a metric called *Bounding Box Factor* to evaluate 'the ratio of the area of the bounding box to the area of the feature' (p. 8). They conclude that the two most relevant factors affecting the *Bounding Box Factor* are the presence of multiple component parts and the shape/orientation of the feature. I.e., 'Conditions where component parts are small and widely separated lead to the most extreme cases' (p. 14). 'In these cases [overlapping bounding boxes], searches based on bounding boxes will result in excessive irrelevant information' (p. 14). It is worth mentioning that their *Bounding Box Factor* only applies to geometries with an area so no points or lines can be studied.

The literature has proposed multiple spatial similarity measurements and algorithms. Some of the most popular ones are Hill (1990), Walker et al. (1992), and Beard and Sharma (1997). Larson and Frontiera (2004) performed a systematic study of the aforementioned techniques. They used a custom-developed test collection of 2527 selected metadata documents from the California Environmental Information Catalog. The extent representations of their test data collection were extracted from the metadata documents from the MBB defined by the metadata producer or from the geometries of the administrative units declared in the metadata record. It is worth mentioning how, they point out that the MBB should be used as a first step to filter

some results and later apply a refinement step but 'in a geographic digital library environment, the end-user is the refinement step'. In their later work Frontiera et al. (2008), they introduced a probabilistic method for spatial retrieval based on logistic regression, computing the similarity score as the probability of a document being relevant to a particular query. The query areas they used were the same geometries of the extent of the resources. They generated the relevance judgments by manually reviewing the 3732 pairs of 'footprint and query area' with the help of overlapping logic in a database. However, they did not consider the content of the datasets for verifying their relevance. As they pointed out, their study was limited by the lack of a better public test collection of coordinate-based documents at that time. Even though their results are not directly comparable with ours because we do not compute a spatial rank, their methodology is still valuable as they face the same lack of test data collections as we do.

Inspired by Renteria-Agualimpia et al. (2015), Lacasta et al. (2017) proposed the use of the Hausdorff Distance, a geometry similarity metric, as spatial relevance indicator. This way, they obtained a spatial rank comparing the given query area with the MBB.

Degbelo and Teka (2019) performed another study comparing different thematic-spatial search strategies focused on Open Government Data. For the spatial component, they compared two approaches: the MBB overlap and the Hausdorff Distance proposed by Lacasta et al. (2017). For the thematic component, they used query expansion using two linguistic knowledge bases: WordNet and Concept Net. They used a custom collection of documents harvested from the United Kingdom Open Data Portal.

Xu et al. (2021) proposed a different similarity measurement based on position graphs. However, they used a custom collection consisting only of complex geometries from buildings and lakes rather than real and diverse spatial datasets.

Cai (2011) noted that very few techniques have demonstrated their feasibility through prototypes. He also pointed out that there is a lack of evidence of the benefits of creating high-cost infrastructures to support more complex geometric and thematic models. Our work aims to address this later issue by evaluating the cost in the precision of current real GDC.

3.3 The problem of the Minimum Bounding Box

A user that searches for datasets in a GDC expects that the results would be related to the provided query thematically and spatially. As mentioned before, the common approach to determine if a spatial resource is relevant or not in GDC is the geometric

intersection between the provided QA and the MBB declared in the metadata record of this resource.

From a user's perspective, a dataset is related to its query if the spatial extent of the dataset is, at least partially, contained inside its QA. A user should not be concerned about the mechanisms behind the GIR system or the existence of MBBs. This may lead to false positive results where the MBB intersects with the query but their content is empty on that particular area as the example shown in Figure 3.1.

As Caldwell (2005) pointed out, the 'bounding box factor' varies depending on the distribution of the features inside the dataset. The more spread the features inside a dataset are, the more empty areas the MBBs will cover, hence the more false positives may cause. On the contrary, the more condensed and uniform the footprint is, the less likely to have gaps and holes it will have. This property can be extrapolated to data collections of geometries with no area such as points and lines. Although the 'bounding box factor' is not computable for all resources, we can guess that the odds of being a false positive result will be affected by the size of the QA and the size of the gaps between the features. Even assuming a perfectly continuous footprint, such as a raster image dataset with orthophotos, Roth (2011) studied how the shape and orientation of the geometry affect the empty space inside the MBBs.

Each false positive result impacts the overall precision of the response list and ultimately the performance of the catalogue. As most GDC only have access to the metadata document and not the resource itself, they cannot provide a preview of the contents of the resource, so the users need to manually access or download it and check its content by themselves. Hence, every false positive result comes with a potential loss of time, or ultimately, a loss of confidence in the dataset owner or the spatial catalogue itself.

3.4 Method

We have defined a method for studying the precision of the bounding box based spatial retrieval systems used in most GDCs inspired by the study of Frontiera et al. (2008). They identified three key elements that compose an *Information Retrieval* test collection:

- A test data collection of spatial resources.
- A test query collection of representative QAs.
- A criteria to determine if a resource is relevant to a query or not.

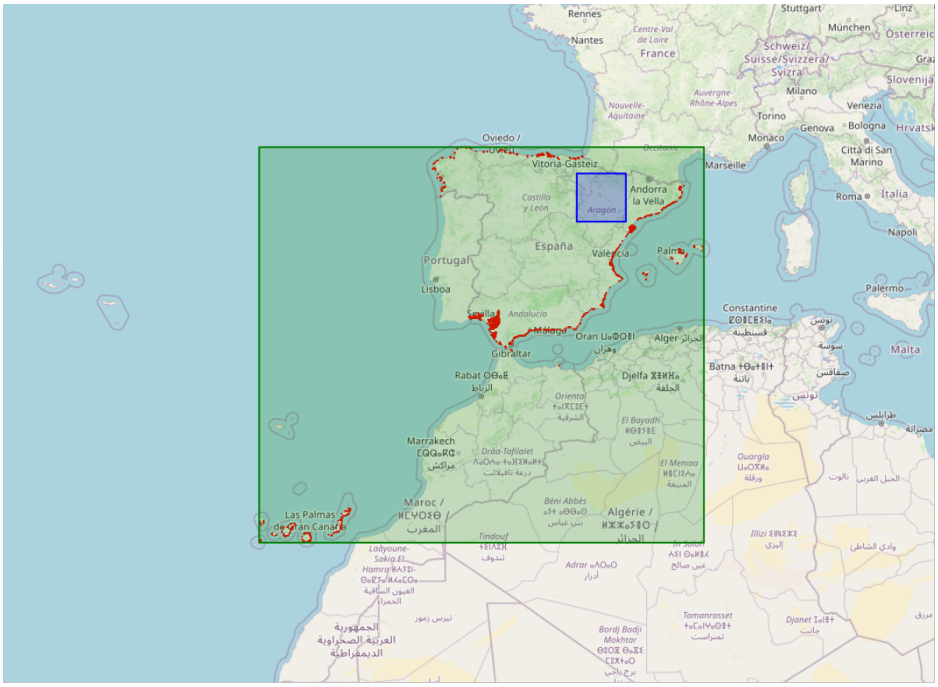

Figure 3.1: Example of a false positive result, where the bounding box of the resource (green big rectangle, associated with a dataset describing the Spanish coast lines) intersects with the query area (blue smaller rectangle) but there are no features (red geometries around the coast) inside that query area.

The method performs a classical MBB based retrieval process and then evaluates which results are true positives and which ones are false positives using our own criteria. By applying this process over different test collections we get a deeper and more diverse understanding of phenomenon. We use the classic *Information Retrieval* performance precision metric from Perry et al. (1955), defined as: the fraction of retrieved documents (true positives + false positives) that are relevant (true positives).

In our work, we are not comparing the studied method against others, but evaluating its performance in different scenarios. This makes it difficult to establish a baseline or a minimum acceptance threshold. The report (JGSI, 2002) used by Ureña-Cámara et al. (2019) in their metadata quality study suggested an Acceptance Quality Limit (AQL) of 5%. The concept of AQL represents the maximum amount of defects that would be considered acceptable. However, to the best of our knowledge, there is no explicit reference AQL for search results for either the general domain or the spatial domain.

Most studies use average precision to compare their method with the existing ones. These studies apply different techniques of retrieval over the same data, query and judgments collections and measure and compare the Mean Average Precision to determine which performs better (Purves et al., 2018). We applied the same method over different test collections to answer the following questions:

– Is this problem present in all catalogues?

– How big the problem is?

– Does it affect all of them equally?

The results and metrics to answer those questions include not only the average precision of each catalogue, but the distribution of that precision along the QAs too. We carefully studied the means, variances and percentiles of the precisions to have a better caparison between the different catalogues.

3.4.1 Test Data collection

One of the main challenges for studying the alternatives for representing the spatial extent is to count on a metadata collection including proper references to the associated spatial resources. In Chapter 2 we already pointed out how difficult it is to access some spatial resources described in a catalogue, due to problems of availability, wrong format descriptions, etc. In their work, they harvested the metadata records from multiple priority catalogues of European countries. They extracted the distribution URLs contained in metadata records from which the spatial resources should be downloaded and analyzed their availability and real data format once accessed.

Their study resulted in a valuable collection of spatial metadata records that proved to be accessible in specific data formats. Reusing this collection allowed us to speed up the task of gathering the right resources for the experiment, while avoiding the many challenges they faced in their work, such as incorrect format declarations or incorrect URLs.

Studying Multiple Catalogues

The literature pointed out that the shape and distribution of the geometries are the most determining factors on how much unused space an MBB has. As this empty space is the sole source of the false positive results, we have the hypothesis that different countries with different geographies and publishing policies will report different results when compared.

We chose to evaluate the catalogues separately for two reasons. The first one is that combining all the resources in one single index dramatically reduces the overall precision. That is because there are many resources whose MBB covers other countries due to overseas territories. Those resources wrongly appear always as results in foreign country queries. E.g., France-only datasets whose MBB extends until French islands will always intersect with Portugal resulting always in a false positive result. The

second one is that, as stated before, evaluating them separately replicates their real behaviour where each country publishes their own catalogue which they are responsible for. This approach also allows comparing how the precision behaves in different kinds of countries. One positive side effect this division has is that the computational costs of the experiment are reduced. That is because the evaluation of the intersections between the queries and the resources is, in the worst-case scenario, a Cartesian product. It is easier to evaluate multiple smaller collections than a big one.

Filtering the catalogues

From the catalogues we analyzed in Chapter 2, we discarded the ones that do not satisfy these minimum conditions:

- They cover the whole country: We discarded regional-only catalogues because they have a different nature than the national ones.
- They have at least two 'Administrative Units at Level 3': We required at least two administrative units because otherwise there will only be one query area covering the whole country.
- They contain more than 10 supported datasets: We have considered 10 supported datasets as the minimum amount for being statistically representative.

Six catalogues satisfy all the requirements proposed. The resulting list along with its respective resource count appears in Table 3.1.

Filtering the spatial resources

Valid datasets are the ones that are automatically accessible and well-formed without errors, i.e., no manual user intervention was needed for downloading or fixing them. We have implemented our system to support the following dataset formats: Shape File, GML, WKT and GeoJSON. Those were the most common vectorial downloadable formats in the original data collection. The spatial web services were discarded for two reasons. First, because of the higher cost in time for consistently harvesting all of its contents. Second, because a user can detect a false positive in a web service faster than in a downloadable dataset as they do not have to wait until the whole resource is downloaded. We also discarded the raster formats because, as we previously stated: 'the more condensed and uniform the footprint is, the less likely to have gaps and holes it will have'. Introducing raster datasets would add another degree of variability which could make the results more difficult to interpret. E.g., comparing catalogues with different proportions between raster and vector resources.

Table 3.1 shows the count of how many resources were successfully processed and analyzed in the study. Those resources were selected because they were automatically accessible and well-formed without errors. I.e., no manual user intervention was needed for downloading or fixing them.

3.4.2 Test Query Collection

The query collection is a set of spatial query areas (rectangles) with different sizes and locations. We considered four techniques to compare, some of them inspired by the literature and others designed by us. The first one is subdividing the space in a continuous grid to guarantee that every tile of the grid has the same area (Sahr et al., 2003). The second one is reusing the MBBs of the resources as the query areas like Frontiera et al. (2008) did. The third one is enclosing in a rectangle the administrative units of the studied regions using a system like the Nomenclature of Territorial Units for Statistics (NUTS) [1]. The fourth one is using a real GDC search log with the real QAs that the users requested in the proportion they were used.

Each method has its own trade-offs. We have drawn up a comparison table to evaluate some of the key features that we found interesting (see Table 3.2). The characteristics we have chosen to evaluate allow us to understand how valuable or representative each method is.

- Spatial coverage: The queries cover the whole study area leaving no gaps. This avoids missing any located issues.

- Uniform Distribution: The quereis cover the area uniformly avoiding overlapping. It has statistical benefits.

- Consistent Area: All the areas have the same area (magnitude). It has statistical benefits.

[1] `https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-20-092text`

Country	Code	Res Count	Query Count
Austria	at	118	35
Portugal	pt	91	25
Spain	es	82	59
France	fr	65	101
Finland	fi	39	19
Slovenia	si	17	12

Table 3.1: Finally selected catalogues

- Real case scenario: The areas reproduce plausible query scenarios a user may request. This gives results closer to what a real-world system will have.
- Relevance weight: The areas represent the interest of some regions over others. If one area is more interesting than the other, its statistical weight will be higher.

From the four options, we found that the Real Log was the most interesting, but it required access to multiple search logs in multiple catalogues, which was not feasible. The MBB option has been used in previous studies, but it does not offer any of the characteristics we are looking for. The DGGS option offered a good level of coverage and distribution but lacked any relationship to the real-world query areas. Hence, we chose the third one: enclosing every NUTS administrative unit at level 3 with its minimum bounding rectangles. This representation method simulates quite closely a real case scenario while only losing the ability to ponder the most common queries.

3.4.3 Relevance Criteria

The relevance criteria is the way we determine if a given resource (from the test data collection) is relevant or not with respect to a given query (from the test query collection). It evaluates one result for one particular query and not the whole result list.

Different studies have used different criteria in the literature. Some of them are automatic such as distance measurement, MBB overlapping logic, etc. while others involve a human manual process. In our study, we defined: *A resource is relevant to a query only if the real footprint of the resource overlaps with the query area.* We consider 'overlapping' any kind of spatial relationship in which the intersection of the two areas is not zero (intersect, within, crosses, etc.). This criterion can be automatically verified for metadata records including a reference to the online resource, but in many cases these references are not available.

3.4.4 Loading data and computing results

Essentially, the operations for calculating the precisions are, for each QA:

	Grid	MBB	NUTS	Real Log
Spatial coverage	x		x	
Uniform Distribution	x			
Consistent Area	x			
Real case scenario			x	x
Relevance weight				x

Table 3.2: Query area approaches comparison

– Get the retrieval results: Find all the resources whose MBB intersects the QA.

– Get the relevance judgments (find false positives): Check for each result if none of their features intersect the query.

Ureña-Cámara et al. (2019) performed a metadata quality study where they found that the MBB of around 7% of the metadata records in their study collection was wrongly defined by the metadata producer. To prevent any wrong results from defective MBBs we computed the MBB ourselves instead of reusing the ones declared in the metadata.

For computing all those intersections we converted them to a common compatible system. We have implemented a data processing pipeline in Python that downloads, stores, transforms, and loads the resources into a spatial database where we can apply the intersection operations between all the geometries. The conversions were made with the OGR2OGR command line tool [2]. It is a spatial dataset conversion tool that allows us to read from multiple formats and write into the same spatial database. We chose SpatiaLite as the database because it is portable, easy to back up, and natively supported by OGR2OGR.

Apart from the spatial features and the computed MBBs, we also load the NUTS geometries with the same tool. We specifically used the geometries facilitated by the Geographic Information System of the Commission (GISCO) in their site [3]. It contains multiple levels of NUTS with a resolution of 60 meters. Each geometry has its unique identifier and hence its MBB. Finally, we compute which dataset's MBB intersect with which NUTS QA and whether those datasets contain spatial features within them or not, obtaining tuples with the following structure.

```
(catalogue_id, query_area_id, dataset_id, mbb_insersects, feature_intersects)
```

Catalogue ID contains the code of the catalogue to which the metadata record belongs. *Query Area ID* contains the ID of the NUTS used for that query. *Dataset ID* contains the ID of the dataset retrieved for that query. *MBB Intersects* expresses if the MBB intersects with the QA. *Feature Intersects* expresses if the dataset contains spatial resources inside the QA.

Every result whose `mbb_intersects` is `true` but its `feature_intersects` is `false` should be considered a false positive result. We aggregate this information by Catalogue and by Query Area to study it in the following section.

[2] `https://gdal.org/programs/ogr2ogr.html`

[3] `https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts`

3.5 Results

This section covers not only the precision metrics obtained from the experiments but also other relevant metrics about the data collection built during this study. This aims to provide a better understanding of the results and the context in which they were obtained.

3.5.1 Preeliminary Retrieval Analysis

Figure 3.2 and Figure 3.3 show the intersection relationship between the QAs and the MBBs before evaluating the relevance judgments. These figures help to understand the different distributions along the space that the resources have in relationship with their administrative units. First, Figure 3.2 shows the total amount of resources intersecting each NUTS QA. The difference in the size of the test data collections is notable. This is why we also included the Figure 3.3 where we normalized the results by the number of resources each catalogue has and by the maximum of intersections in a single QA where found. They show how scattered and heterogeneous the distribution of the datasets is. A QA with a value of 1 means that all the MBBs intersect that area; a value of 0 means that this region is not covered by any MBB. The more compact the box plot is, the more homogeneous the distribution is. The higher the values of the box plot are, the less scattered the MBBs will be.

This way we can see how differently some catalogues behave. In the catalogue of Slovenia (SI), most QAs intersect with all 17 datasets which means that this catalogue does not publish regional-only resources, only country-wide ones. The least covered QAs are the outer regions of Pomurska and Koroška. The catalogue of Finland (FI) is very homogeneous too, with only a few regions being less covered by their resources. The least covered one is the island of Åland. The catalogue of France (FR) has an unusually low ratio when the values are normalized by the total resources in the collection. That is because it is the more scattered one. France has lots of overseas territories where they publish regional-only datasets. That means that there is no region in France where more than 40% of datasets MBB intersect. However, if we observe the values normalized by the maximum number of intersections found in the catalogue, we see that more than 95% of the QAs have the same number of intersections so it is very homogeneous in those regions. The least covered regions are the overseas island of Mayotte, Guadeloupe and both regions of Corse (Corsica). The catalogues of Portugal (PT) and Spain (ES) share similar distributions. Both have overseas territories, but less than France, but have a more diverse intersection count among their QAs. Their respective less covered areas are, for Spain: La Palma,

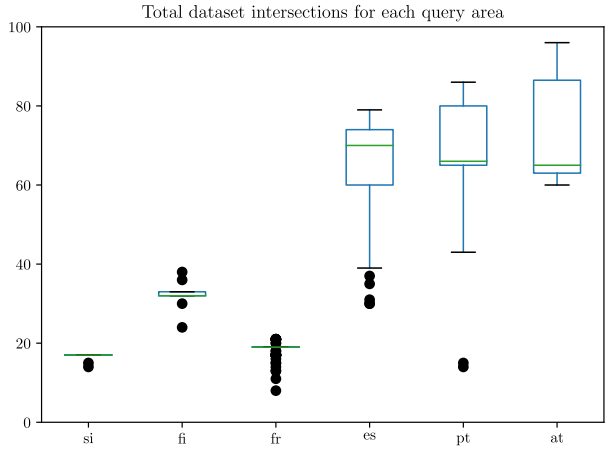

Figure 3.2: Total dataset intersections (%) for each QA

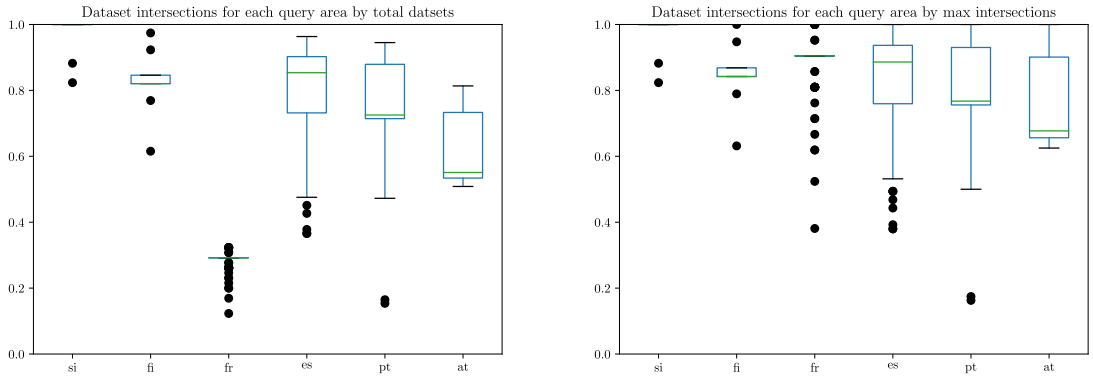

Figure 3.3: Dataset intersections by total datasets and by maximum intersections

La Gomera and El Hierro, the farthest overseas Canary Islands; and for Portugal: Madeira and Azores, both overseas islands too. Finally, the catalogue from Austria (AT) is more scattered but more homogeneous than ES and PT because it has mostly regional datasets, hence its intersection count by total do never exceed 90%, but also is never below 50%.

This analysis concludes that the regions the least covered by the MBBs are the outer ones for all catalogues. This is not a universal rule: a catalogue could potentially publish regional-only datasets focusing more on the outer territories and leaving the interior regions without datasets. But the reality is that the combination of regional-only and country-wide datasets results in the interior regions being more covered.

3.5.2 Precision analysis

Table 3.3 and Figure 3.4 show the precision of the QAs after calculating the false positives using the relevance judgement. Table 3.4 represent the percentile at 90% or more precision score. I.e., how many QAs have more than 90% precision? A percentile of 80% means that 80% of the QAs had more than 90% precision (the higher, the better).

At a glance, there are some first conclusions we can make. All catalogues have regions with 100% precision. This means that the users that search on those QAs will not perceive any problems when accessing the resources. On the other hand, all catalogues have regions with less than 90% precision. This means that on any catalogue, a user may search for a QA and receive 1 non-relevant result out of 10. The mean values have a range between 88% and 97%. However, the minimum values are below 75% on three catalogues. Given the fact that GDCs do not implement spatial ranking and hence considering that the order or the results are random: a user that searches on those conflicting query areas may have up to a 77% chance of finding a non-relevant resource as the first result.

The best-performing catalogues are from Slovenia, Finland and Austria. They all

catalogue	count	mean	std	min	25%	50%	75%	max
si	12	0.93	0.08	0.76	0.88	0.94	1.00	1.00
fi	19	0.97	0.04	0.88	0.97	0.97	1.00	1.00
fr	101	0.74	0.15	0.53	0.63	0.74	0.89	1.00
es	59	0.73	0.19	0.33	0.55	0.74	0.92	1.00
pt	25	0.88	0.13	0.62	0.79	0.95	0.98	1.00
at	35	0.94	0.05	0.83	0.93	0.95	0.97	1.00

Table 3.3: Precision statistical description

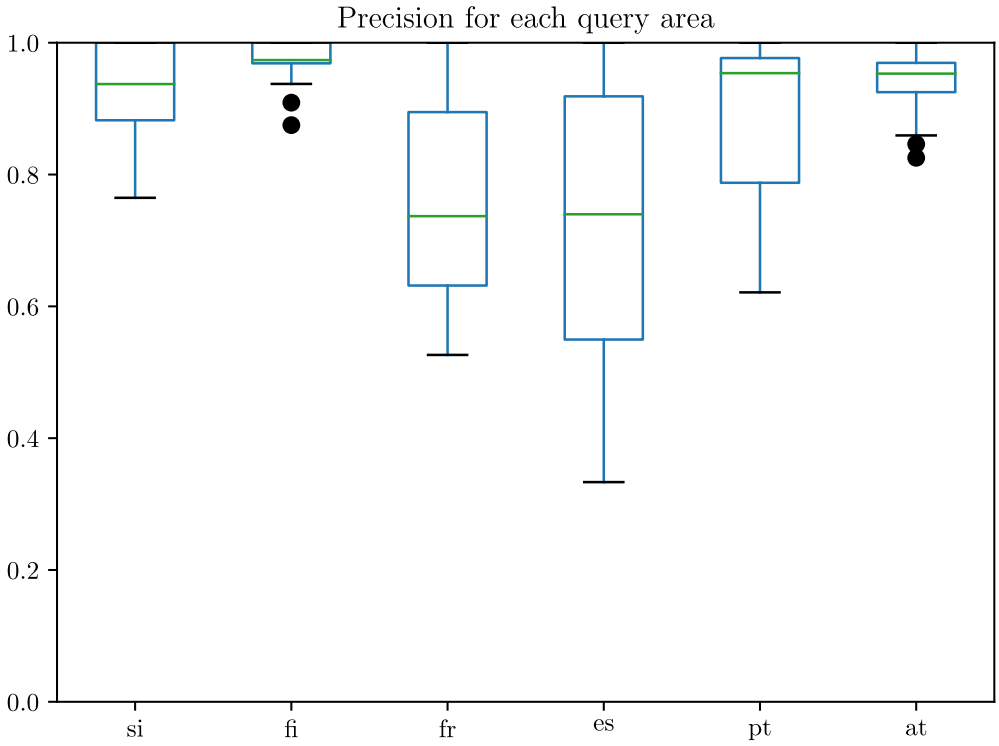

Figure 3.4: Precision for each QA

catalogue	total	count@+90%	percentile@+90%
si	12	7	0.58
fi	19	18	0.95
fr	101	21	0.21
es	59	18	0.31
pt	25	16	0.64
at	35	28	0.80

Table 3.4: Percentile at +90% precision (the higher, the better)

share the best average precisions and minimum values. However, Portugal has a better percentile at 90% precision than Austria. Portugal is the best performing among the 3 worst ones. Half of its QAs have a precision higher than 95%. However, Portugal has much a lower 25% percentile than Slovenia, Finland and Austria The worst performing ones are France and Spain, with both medians at 74% precision. Spain has the lower minimum and the lowest precisions but also a higher 75% percentile than Portugal. I.e., the results in Spain have more variance than in Portugal.

3.6 Discussion

We have considered the presence of spatial features inside the query area as a minimum requirement for a resource to be considered spatially relevant in that area. The empirical results have shown that all the studied catalogues have scenarios where they return results that are not truly spatially relevant.

We previously discussed how difficult it is to estimate how bad this issue is; if a 10% of false positives is a reasonable error rate; what is the cost for the users in terms of time wasted; what is the cost for the GDCs in terms of reputation lost. If we set the 5% AQL used by Ureña-Cámara et al. (2019) in their study, none of the catalogues would have passed the acceptance limit on average, being the Finnish catalogue the best-performing catalogue with exactly 95% of its QAs reporting a precision above 10%.

The results suggest that the precision metrics vary depending on the distribution and coverage of the resource collection, the geometry of the country, etc. It is also difficult to study how each factor affects the precision individually as the volume of information we could gather is not large enough. Answering that question requires an even deeper study of the distribution of the features inside each dataset and its interactions with their MBBs.

This problem has other considerations. As many GDCs do not implement any spatial ranking system, the false positive results may have the same chance to be in the first position as the true positive ones. That is because the ranking if applied, is only computed from the thematic component.

It may also limit the evolution of potential initiatives. Lacasta et al. (2017) mentioned the need of a Pan-European catalogue to satisfy the users that need continuous data across countries. The INSPIRE directive (INSPIRE MIG, 2017) is an example of an initiative in that direction. However, this idea can not be implemented yet until this issue is addressed because some regions will inevitably suffer this problem at a much bigger scale. E.g., The MBB of France's mainland covers approximately

one-third of Switzerland and it increases to more than half of the country if the MBB includes Corsica.

Another aspect that we have not been able to study under the scope of this work is how this phenomenon affects other kinds of spatial resources such as raster datasets or spatial web services. A bigger study that implements new transformation pipelines for the new data resources could complement the information this work has found.

This problem has not a straightforward solution. One possible approach could be, instead of using dichotomic criteria, introducing any ranking system to at least push the false positive results to the bottom of the result list. Multiple solutions have been studied yet in the literature Frontiera et al. (2008); Lacasta et al. (2017); Xu et al. (2021).

Another solution, that could be combined with the ranking system or implemented alone, could be finding better ways to represent the footprint of the resources. The literature already has proposed multiple of them: Convex Hulls, Two MBBs, etc. Evaluating the trade-offs of each one and measuring its performance under real workloads could lead.

Any of the previous solutions will, at some point, be limited by some lack of precision in the extent of representation. A different approach could be a system inspired by Gao et al. (2016) where the contents of the resource are directly indexed in the spatial index. It will prevent any false positive as it will retrieve resources based on its contents indexed in that area. However, it will not benefit from the anisotropic properties it had for text documents so another ranking system should be designed if needed. That is because the more you mention a place in a news article the more bound to that place it is, but this relationship can not be applied to spatial resources. The amount of features a dataset has concentrated in one area does not indicate more interest in that area.

Implementing a system like that will face a big challenge. Unlike their domain, where they had access to the whole resource (the news article), in the current GDC architecture the metadata and the resource are two independent artifacts with independent life cycles and the GDC does not receives or has direct access to the resource at any moment. The metadata records by themselves can not always provide the access to the spatial resources they describe, as we pointed out in Chapter 2.

Cai (2011) pointed out how there was a lack of evidence about the benefit versus the cost of adopting more sophisticated GIR solutions in real-world production systems. We believe that our work brings new evidence that reinforces the need for a better system. However, future proposals must consider the cost of implementation to facilitate its adoption and finally improve the discovery of real spatial resources.

3.7 Conclusions

In this work we have measured how using the Minimum Bounding Box (MBBs) as the only indicator for the spatial coverage of an information resource affects the precision of the spatial search results. The criteria we have used to assert the relevance of a resource to a search area is if the real footprint of the resource overlaps the search area.

As the MBB spatial search is widely used in most geospatial data catalogues (GDC), we have studied the average precision of six real relevant European GDCs. We observed how all of them suffered from precision issues on at least some of the test query areas (QA). The results indicate that the average precision of the search scenarios is between 73% and 97%. This means that in some catalogues the use of MBB instead of the extent to describe resources can lead to an average failure rate of around 25% for user queries. Only two catalogues reported a precision above 90% in 95% and 80% of their tests respectively, while, in the most affected catalogue, only 21% of the search scenarios achieved precision greater than 90%.

We have discussed how these false positive results affect the real user experience when trying to discover new spatial resources in a GDC and how it may limit future initiatives such as multi-country catalogues. We have also glimpsed some possible directions in which future works can move towards better retrieval techniques while also considering the cost of the solutions as a limiting factor for their adoption.

Chapter 4

Improving the precision of spatial search results

4.1 Introduction

In Chapter 3, we have empirically verified that current spatial data catalogues suffer from precision problems in their spatial search results. The results suggest that the representation of the spatial extent of a resource is a critical factor. By using a rough representation such as the *Minimum Bounding Box*, the area covered by the metadata is overestimated, leading to more spatial matches than there actually are. Hence, a reasonable approach would be to find a better representation technique for the spatial extent of a resource. By doing this, the likelihood of reporting false positives would be reduced. Figure 4.1 shows how different methods provide different accuracies.

The literature has studied multiple alternate methods for representing the spatial influence of any resource (see Section 4.2). Among all the studied ones, the best fitting candidate until now is the Convex Hull (CH), defined as the smallest convex polygon that encloses a collection of geometries (de Berg et al., 2008, p. 2). As the Convex Hull of any collection is, by definition, contained inside its MBB, the accuracy of its representation and its area will always be better or equal. However, we propose a different approach.

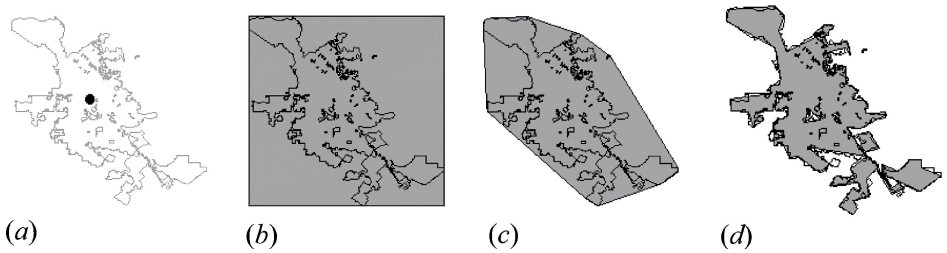

Figure 4.1: Set of possible geographic footprints for the city of San Jose, California: (a) single point; (b) min. bounding box (MBB); (c) Convex Hull (CH); (d) generalized polygon. (from Frontiera et al. (2008))

We propose a different method that involves discretizing the world using a grid system and representing the extent of a dataset as the list of grid tiles it covers. We chose to implement this approach using a Discrete Global Grid System (DGGS), a hierarchical system for partitioning the surface of the Earth into a finite number of discrete cells or tiles with unique identifiers. Hence, we call this representation DGGS Footprint and is defined as: the list of DGGS tiles that intersect with the actual footprint of a dataset. This approach allows for more precise spatial queries, as a big bounding box is replaced by smaller tiles, better fitted to the actual geometry of the spatial features.

To measure the improvement of our method, we conducted an empirical study comparing our DGGS-based spatial search against MBB and Convex Hull. For doing that, we generated a test data collection with real geospatial metadata records and their datasets extracted from multiple curated European catalogues. For each dataset, we computed the MBB, Convex Hull, and DGGS tiles, and we used these representations to search for datasets that overlapped a user-specified search area. We measured the precision of each method by examining whether the retrieved data actually contained data features within the user-specified search area.

In this chapter, we demonstrate how using DGGS Footprints provides more precise results than MBB and Convex Hull. We also study and propose a progressive transition from existing MBB-based representations to DGGS based ones.

The remains of this chapter are structured as follows. Section 4.2 presents previous works related to spatial information retrieval and spatial data catalogues. Section 4.3 describes our proposed spatial representation technique. Section 4.4 describes the method we designed for comparing the improvements of our proposed technique. Section 4.4.2 displays the results of our study. Section 4.5 proposes a roadmap for transitioning from current spatial representation techniques to our proposed one. Section 4.6 offers a comprehensive summary of the key findings, contributions, and implications of our research.

4.2 Related Work

The works related to the quality of the spatial search have already been covered in Section 3.2. In this section, we will complement that with the solutions and proposals that other authors have considered.

4.2.1 Alternative extent representations

In 2004, (Larson and Frontiera, 2004) stated that 'Minimum bounding ellipse, minimum bounding N-corner convex polygon, and Convex Hull, have been investigated in the context of spatial databases and GIS applications, but not for GIR, where the MBB still represents the state of the art'. Since then, very few studies have proposed specific methods for extent representations for spatial retrieval. Roth (2011) proposed a method for reducing the amount of uncovered space by using two bounding boxes instead of one. Their algorithm efficiently assigned the two best-fitting rectangles that minimize the uncovered space. He considers other shapes like Oriented Bounding Rectangles, Convex Hulls, or Multi Hulls, but none of them have the balance between simplicity and efficacy as they require more complex calculations. Gao et al. (2016) studies the limitations of the MBB retrieval method for text-based documents. They claim it to be space-redundant, as they cover a wider area tha one described in the text; and isotropic, as they do not describe the density or scale of the information. Hence, they propose a system based on point sets for representing footprints instead of bounding geometries, where each point represents each identified location in the text. This allows them to keep their density and scale. They compared their method with MBB and convex hull based techniques with a test collection of 700 news articles crawled from the Chinese platform Sina News. and 50 predefined spatial queries. Their work suggests the idea of indexing coordinate-based points directly instead of MBBs, but only in the text-based domain and only for point geometry types. This is aligned with the proposal of the Purves et al. (2007) and the Gey et al. (2006) information retrieval conferenes that propose the indexing of extracted georeferences in terms of coordinates.

Despite the extensive use of Convex Hull for spatial data analysis (Longley et al., 2005), and even being used as benchmark method (Frontiera et al., 2008), to the best of our knowledge, there is currently no research proposing the use of Convex Hull. Neither Convex Hull nor any other of the presented methods have been used as an alternative representation of extent in metadata records.

4.2.2 Geocoding based representations

While the proposed methods use vector geometries to define resource extent, literature and industry also rely on rasterization/tessellation/encoding as an indexing method for spatial retrieval. Geocoding solutions like CUID and AUID (Béjar et al., 2019), GeoHash (open) (Gustavo Niemeyer, 2008), What3Words (proprietary) (What3words Limited, 2023) and Open Location Code by Google (Open) (Philipp Bunge et al., 2018)

provide a simple way of encoding geographic coordinates into a compact string format. However, these methods have never been proposed by the literature as a footprint description for a metadata record in catalogues. Any of these options could fulfill the requirements of a hypothetical system similar to the one we propose as each of them provides a mechanism of assigning unique identifiers to predefined areas.

4.3 Proposed method for indexing dataset extents with DGGS Tiles

As mentioned in the introduction, DGGSs are hierarchical systems for partitioning the surface of the Earth into a finite number of discrete cells or tiles. DGGSs have grown in popularity to the point that they are currently being standardized by the OGC (Purss et al., 2016), and are being used as mechanisms to facilitate the efficient integration of heterogeneous spatial data (Béjar et al., 2023).

All DGGSs provide a mechanism for assigning a unique identifier (ID) to each cell on each hierarchy level (DGGS-ID) (Sahr et al., 2003) with each tile at a certain level corresponding to a smaller area than its parent tile. Our method proposes that, when publishers record the spatial extent of the created resource in the metadata, they should also compute the list of DGGS-IDs that intersect with the resource and include it in the metadata record.

To obtain this DGGS Footprint, publishers will compute the intersections of their resource geometries with the reference DGGS tiles and preserve the IDs of the matching ones. This process essentially involves 1) the transformation of a vector spatial resource into a raster representation and assign to each pixel a value of 'true' or 'false' based on whether it covers the original resource or not; 2) the selectction of the cells marked as true; 3) obtain the IDs of the covering pixels (tiles). As a result, the computed DGGS Footprint will look like Figure 4.2.

4.3.1 Indexing and Spatial Search

Upon receiving a new metadata record, the spatial data catalogue we propose will add the DGGS Footprint to its inverted index of DGGS-IDs pointing back to the metadata, enabling efficient retrieval during a search. This index could be implemented as a spatial index where the DGGS-IDs are converted back to tile geometries or could be a text index where the DGGS-IDs are stored as plain text. Our proposal is to index them as text, provided that an agreement is reached to use the same DGGS for all metadata records. The reasons behind this choice will be further discussed in Section 4.5.

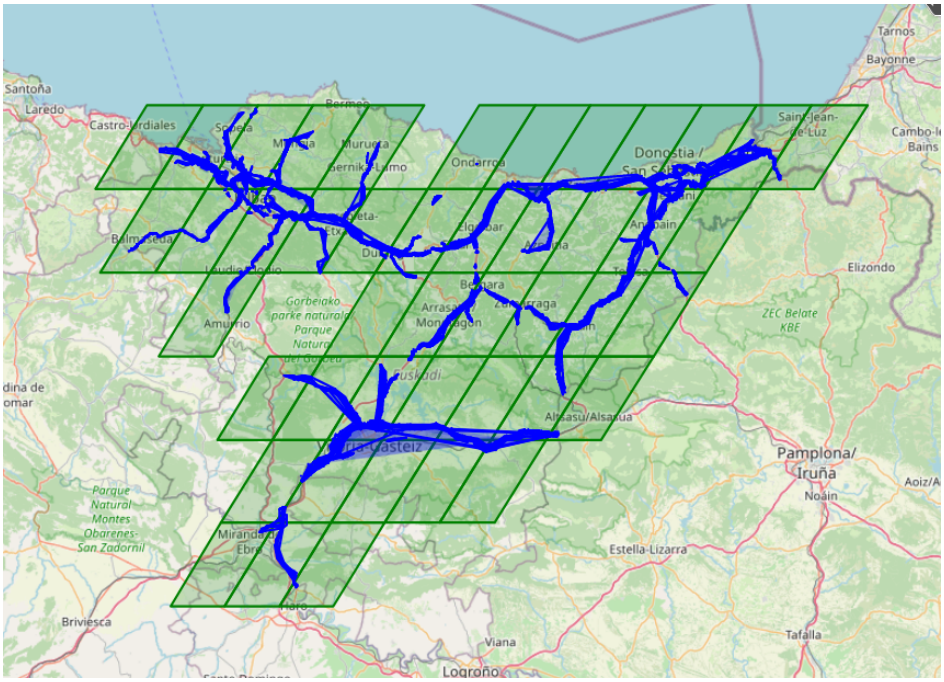

Figure 4.2: Example of rHEALPix DGGS tiles that cover a dataset

The catalogue will still be able to offer spatial searches over indexed textual IDs due to the following property: The DGGS Footprint is an approximation that always overestimates the area because the tiles will always envelop the geometries. Hence, if two geometries intersect (e.g. a feature and the query area), they will have at least one tile in common.

To perform a spatial search, the system will first compute the DGGS Footprint of the user-defined spatial query area and then use the DGGS-IDs index to return the spatial resources that share at least one DGGS-ID with it. As a result, all the spatial resources that will be retrieved will have at least one DGGS-ID in common with the query area, ensuring their coverage.

4.3.2 rHEALPix

Currently, there are several Discrete Global Grid Systems: Uber H3, DGGrid, Google S2, OpenEaggr or rHEALPix. Among all of them, the chosen system to conduct this experiment is rHEALPix. Its names stand for rearranged Hierarchical Equal Area isoLatitude Pixelization (rHEALPix) and it is a cubic geodesic DGGS based on cells that are squares once they are projected Gibb (2016). The use of squared cells simplifies the assignation of unique IDs to each cell, making them congruent with their upper hierarchies (see Figure 4.3). This means that, unlike other non-congruent shapes like hexagons, all the area covered by a square cell will also be covered by its parent cell. The system is also compliant with the OGC current DGGS standard.

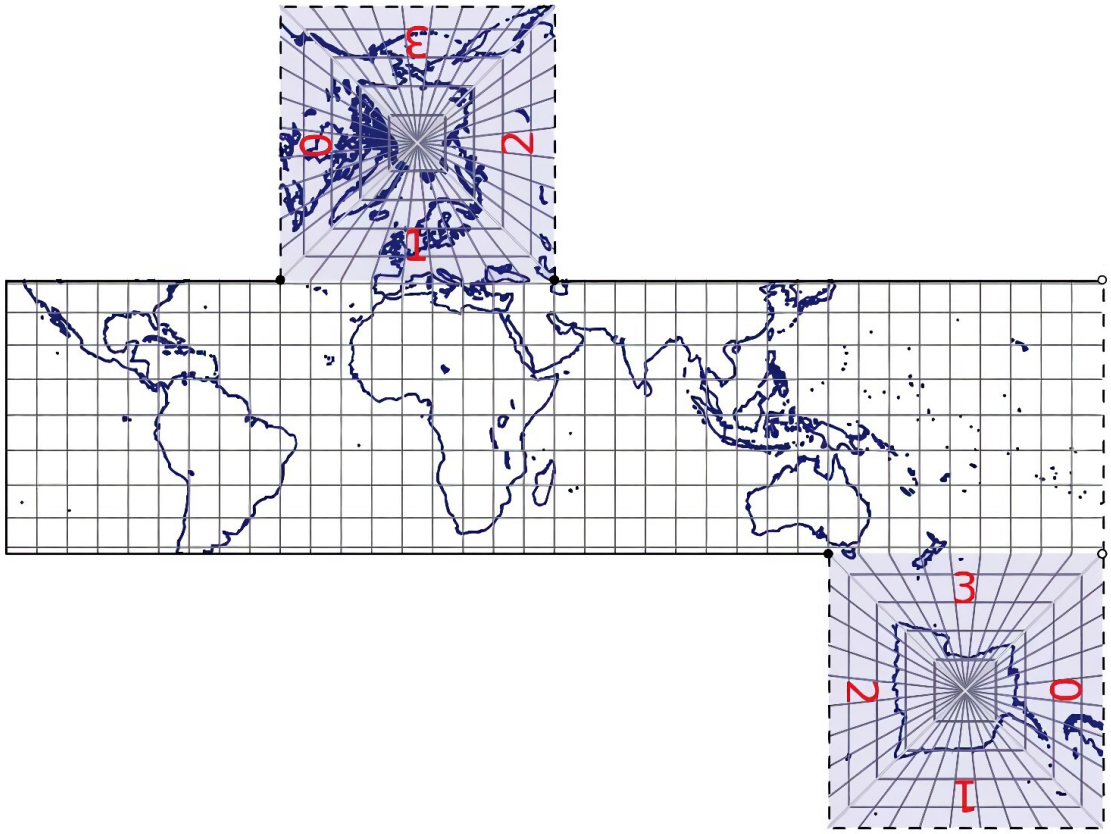

Figure 4.3: (1,3)-rHEALPix projection of the world. (From Gibb (2016))

The reasons why we choose this system over others is further discussed in the Section 4.5. The specific settings we used are the `(1,0)-rHEALPix` projection with `nside=3` and a level of hierarchy of 6, generating tiles of 13.735m^2.

4.4 Empirical study of precision improvements

To justify the necessity or suitability of the proposed method, it is essential to empirically evaluate the advantages it offers over existing solutions. For this reason, we apply the same method over the three spatial search approaches we chose to study (MBB, Convex Hull and DGGS Footprint) and compare the results among them. The chosen method is the same we designed in Section 3.4, but extended with the Convex Hull and DGGS methods.

4.4.1 Computing the footprints and the results

To collect the new metrics, we performed the following process:

1. Compute the three alternative footprint representations of the spatial extent described in each metadata record: MBB, Convex Hull and DGGS.

2. Perform the spatial queries against the set of catalogue resources with the three alternative representations of the spatial extent. This produces three different result lists for each search scenario.

3. Evaluate each result as true positive or false positive for that query area by checking if its real footprint intersects with the query area.

4. Compute the average of the results for each catalogue and compare the three representation methods.

The data processing was also made in Spatialite databases where the original resources were loaded in. Using the Spatialite operators we computed the Convex Hull of each data resource as well as their DGGS Footprints by computing the intersections with the pre-loaded Level 6 rHEALPix Tiles. Ureña-Cámara et al. (2019) performed a metadata quality study where they found that the MBB of around 7% of the metadata records in their study collection was wrongly defined by the metadata producer.

At a precision level of 6, the area of each tile is 13.735m^2, which means that thousands of tiles will be needed to cover the surface of an entire county. Figure 4.4 shows the number of tiles required to represent the footprint of each dataset. Without any hierarchical aggregation, the resources require up to 4974 tile IDs. Since a dataset

covering thousands of tiles spans a wider area, it is more likely for these tiles to be aggregated into their parent tile. When applying the aggregation, the number of tiles required decreases until no resource needs more than 2500 tiles. Given that each encoded tile ID at level 6 requires 8 characters (consisting of 1 cube side letter, 6 subdivision numbers, and 1 separator), the total size of 2500 tiles would be 20Kb. As this is not the main focus of our work, we will only discuss further compression techniques in the Discussion section.

At the end of the processing, we generate a table to show the results of the experiments with the following structure:

- Catalogue ID: the code of the catalogue of the resource.
- Query Area ID: the ID of the NUTS used for that query.
- Dataset ID: the ID of the resource for that query.
- MBB Intersects: if the MBB intersects with the query area.
- Convex Hull Intersects: if the Convex Hull intersects with the query area.
- DGGS Intersects: if the DGGS Footprint intersects with the query area.
- True Positive: if the resource contains spatial resources inside the query area (The scenario in Figure 3.1 would be an example of a false positive).

Every result whose `method_intersects` is `true` but its `true_positive` is `false` if it is a false positive result for that method. We aggregate this information by Catalogue and by Query Area to compute the Average Precision, the quantiles and a customized metric we defined as 'the number of search scenarios yielding a precision level below 90%'. With this later metric, it is possible to quantify the percentage of search results that do not reach the acceptable level of precision, based on the 90% threshold. However, achieving a perfect score of 0% in this metric only indicates that all search scenarios exceeded the precision threshold of 90%, not that the searches were error-free.

4.4.2 Results

Figure 4.5 represents three box plots with the precision results of each catalogue using the three spatial representation methods (MBB, Convex Hull and DGGS). Table 4.1 summarizes the average precision of each catalogue and each representation method along with their amount of results below 90% precision. Upon the analysis of the

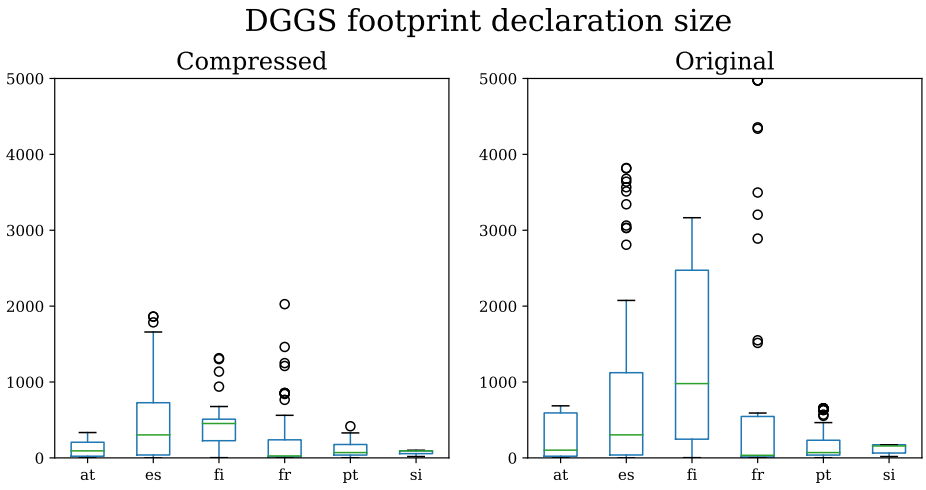

Figure 4.4: Number of tiles required to represent the footprint of each dataset (right: original, left: simplified)

results, it can be observed that our proposal of using DGGS is promising as there is a significant improvement in the search precision when the representation of the footprint of the resources is more detailed. First, Convex Hull, as a subspace of the MBB, always offers equal or better precision that the latter on all search scenarios. However, while there is a big improvement in specific search scenarios, the biggest increase in catalogue average precision is only a 4 percentage point rise. As we expected, the DGGS method outperforms the two alternatives giving an average precision score of over 96% was achieved for all catalogues with a precision increase of up to 24 percentage points.

If we analyze the results of each catalogue individually, we can see multiple behaviors. In the Austrian catalogue, the precision of MBB results is among the best ones, so there is less room for improvement. The Convex Hull and the DGGS barely improve the average precision. However, in the percentage of search results below 90%

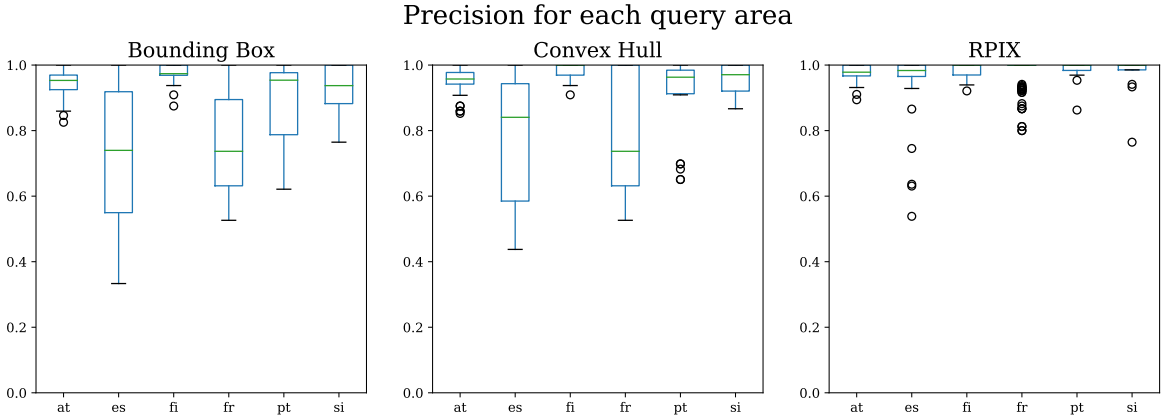

Figure 4.5: Precision dispersion comparison between the three sptial representation methods.

		Average Precision			Precision < 90%		
cat	count	MBB	CH	DGGS	MBB	CH	DGGS
at	35	0.94	0.95	0.98	0.20	0.14	0.03
es	59	0.73	0.77	0.96	0.69	0.59	0.08
fi	19	0.97	0.98	0.98	0.05	0.00	0.00
fr	101	0.74	0.77	0.98	0.79	0.70	0.09
pt	25	0.88	0.91	0.99	0.36	0.20	0.04
si	12	0.93	0.95	0.97	0.42	0.25	0.08

Table 4.1: Average precision and precision below 90% per catalogue

precision, the improvement is outstanding. In the Spanish catalogue, the precision of MBB results is significantly lower than others. The Convex Hull barely improves the results while the DGGS solves most scenarios. In the Finnish catalogue, similarly to the Austrian one, the results had less room for improvement. Hence, the methods behave only slightly better. In the French catalogue, we see a big improvement from MBB to Convex Hull in the upper 25% of the box plot, meaning that regions benefit from this approach while others did not improve that much. Similarly to the Spanish catalogue, the DGGS method still faced some problematic regions. However, the overall results present one of the biggest improvements, specifically with the MBB almost 80% of search scenarios yielded a precision bellow 90% decreasing to 9% of search scenarios when using DGGS. The Portuguese catalogue is the one that benefits the most going from MBB to Convex Hull, specially on the lower 50% of the box plot. However, the DGGS still manages to improve even more, specially in the problematic areas. We can consider that the Slovenian catalogue behaves similar to the Austrian and the Finnish one: slight but consistent improvement.

4.5 A transition to a DGGS-based system

The results show how every MBB based catalogue has precision issues and how DGGS Footprint improves them. Currently, there is a wide range of spatial information systems in operation, whose catalogues employ the MBB method for spatial search. However, these highly valuable systems need to remain available, despite their lower accuracy, while transitioning to a system that offers improved performance. Therefore, we are going to conduct an assessment of the needs and adaptations that would be necessary to achieve a proper transition, involving each step in the process: metadata creation, indexing, and searching.

4.5.1 Metadata creation

The success of this proposal relies on the adoption of existing technologies. However, as an alternative to MBB, it cannot be built on top of it, but besides it. Software companies implementing catalogues should facilitate the availability of tools to link the spatial extent of resources to DGGS tiles. This process would present the perfect opportunity to unify and reach a consensus on the use of a single DGGS system, ensuring compatibility among all DGGS-IDs. We propose rHEALPix because it is a well known system that offers:

- A simple unique ID assigning scheme: Other DGGSs mentioned in Section 4.2 offer more complicated ID systems that are not congruent among hierarchy levels, which difficult many cross-resolution operations.

- Congruent square-based tiles: rHEALPix uses squares that, unlike other shapes like hexagons, maintain a consistent coverage throughout their hierarchy levels. This facilitates many cross-resolution operations.

- Support for most modern tools and libraries: It has an implementation as a python library. In addition, it is supported in the common projection library PROJ contributors (2023). This facilitates the data transformation and integration with existing tooling systems.

The responsibility for incorporating and implementing this representation method should lie with the software developers, committees, catalogues and data producers. To ensure a smooth coexistence of both systems, during the creation of the metadata for their spatial resource, they should include information about the MBB as well as the DGGS Footprint. This approach involves all the producers who wish to adopt this system to facilitate the discovery of their resources and allows for distributing the workload of the transition. However, the resources of those producers who choose not to offer their DGGS Footprint would not be able to benefit from this. In these cases, the catalogues themselves could try to access the original resources and compute the DGGS Footprint on their own. However, we have already demonstrated in Chapter 2 the difficulties that exist in accessing some spatial resources solely through their metadata.

4.5.2 Indexing and retrieval

Technologically, there are various ways to coexist the classical system based on MBB with the DGGS-based system. In a context where a universal and shared DGGS configuration has been agreed upon, the DGGS-IDs are all mutually compatible,

allowing them to be directly indexed as textual keywords. These inverted indices of DGGS-IDs would coexist with the indices that index the MBBs. In the case where a metadata uses a different ID system, there would always be possible workarounds such as reprojecting to the standard DGGS (assuming a certain loss of precision).

In this way, the search process would be as follows:

- Calculate the DGGS-IDs of the query area provided by the user.
- Retrieve the relevant resources from the DGGS-IDs index that match the DGGS-IDs of the query area.
- Retrieve the relevant resources (that do not have a DGGS Footprint) from the MBB index using the classical method.

The result would be a more accurate list of results without disregarding spatial resources that have not yet provided their DGGS Footprint.

4.6 Conclusions

In this chapter, we have examined the advantages and disadvantages that would arise from using DGGS Footprint compared to MBB for spatial data retrieval. The main objective of this study was to quantify the improvement in the precision of spatial search results, as well as to examine the implications and requirements that such a transition would entail. Our findings demonstrate that the precision of search results improves significantly when a more precise representation of a resource footprint is used, being DGGS Footprint the best among the studied solutions. These results are important because, given the evidence of the existing problem, they present a promising alternative. In addition to measuring the results, this study also proposes a method of progressive transition to the new system, suggesting various solutions to the potential challenges that may arise. Furthermore, it also takes into consideration the drawbacks and costs associated with such a transition. In summary, this study has successfully demonstrated the advantages of using DGGS Footprints in spatial data catalogues, supporting the need to transition to more effective systems.

Future lines of work could, for example, apply this study to a different corpus: with a larger amount of data, with another type or format of data, or with another data domain. It would also be possible to explore other representation methods alternative that aim to achieve the same goal of improving the spatial representation of the resource. Finally, a more refined study of the method for obtaining DGGS Footprints would be desirable, as it is a technique that still has a lot of potential for optimization. This would also facilitate the adoption by metadata creators.

Chapter 5

Conclusions and Future Work

The current avaliability and needs for the use of geographic information in complex scenarios are demanding a better performance of Spatial Data Infrastructures and their Geospatial Data Catalogues. The main goal of this PhD thesis has been to address the problems that are affecting the ability of users to find and discover the resources they need. As a result, we demonstrated that current catalogues are not up to the task. This thesis has not only demonstrated the deficiencies of existing catalogues in facilitating effective resource discovery and access but has also proposed several lines of improvement to address these issues. These proposals are, in general terms:

- Improving the annotation of resources, adopting best practices or providing better tools to metadata producers to ensure that the information they reflect in the metadata corresponds to the reality of the spatial resource.

- Facilitating automatic access to resources, thus allowing the integration of various systems without the need for human intervention.

- Greater care of the health status of distribution links, ensuring that the links offered in the metadata are accessible and lead to the original resource in the format that the user expects. This point would benefit greatly from the previous point.

- Replacing spatial indexes based on MBB with more accurate DGGS based descriptions of the footprint of a spatial resource to improve the precision of search results.

Each chapter has presented its own local conclusions based on the research conducted. In this final chapter we will summarize the main ones and discuss the future work that could be carried out to continue the line of research presented here.

5.1 Research Contributions

The main hypothesis was that the current catalogues are suffering from a range of problems yet to be studied. To address that, we defined a sistematic approach to study them and obtain empirical evidence of their impact.

Chapter 2 addresses the problem of *Link Rot* and *Content Drift* in the context of *Spatial Data Infrastructure*, a novel approach that had not been studied before. In this study, we measured how broken or migrated links in spatial metadata negatively impact the ability to find the original resource. In this study, we analyzed the metadata of 26 European catalogues of the highest quality (the so-called, "priority datasets"). We answered the first research question (RQ1) by showing that the problem of *Link Rot* and *Content Drift* is widespread, with a significant percentage of broken or migrated links. In some cases, when all the links in a metadata record were inaccessible, it was virtually impossible to retrieve the described resource. In addition, it was discovered that a large number of distribution links did not lead directly to the original resources, but to intermediate websites or proprietary viewers that required human action to obtain the final resource. It was commented how this was a huge obstacle when automating processes of discovery, access, or even measuring the quality of the resources offered in the catalogue. We also partially answered the third research question (RQ3) by proposing a set of guidelines to mitigate the problem of *Link Rot* and *Content Drift*. The results have been published in the International Journal of Geo-Information (Martin-Segura et al., 2022) and presented at the 22nd AGILE Conference in Limassol (Martín Segura et al., 2019). Additionally, the findings were shared at the JIIDE Conference, with the proceedings published in the MAPPING journal (Martín Segura et al., 2018). The resulting code and dataset from the experiments are published in Figshare[1]

Chapter 3 addresses another of the problems that affect the discovery of spatial data in catalogues: the precision of search results. To measure this, an empirical study was conducted on the spatial resources (that were accessible) of the previous work. We answered the second research question (RQ2) by showing that traditional methods based on the description of the dataset footprint with a *Minimum Bounding Box* yield a large number of erroneous results. This is detrimental to the user experience, as that user not only does not find the resource they are looking for, but also lose time and confidence in the data provider. The results of this chapter has been presented as part of the article published in the International Journal of Geographical Information

[1] `https://figshare.com/articles/dataset/The_problem_of_Reference_Rot_in_Spatial_Metadata_Catalogues/16940365?file=31337734`

Science (Martin-Segura et al., 2024b).

Chapter 4 addresses the task of studying and proposing a new alternative indexing or search method to minimize the problems of precision. The proposed technique was the use of a DGGS to describe more accurately the area covered by a resource. With a more accurate description, errors due to overestimation of coverage should be reduced. To demonstrate this empirically, the study was repeated on the datasets of Chapter 3 comparing the reported precision when describing the footprint using *Minimum Bounding Box* with those reported using Convex Hulls and DGGS. The study showed that the results when using DGGS are much more precise than when using *Minimum Bounding Box* or Convex Hull. This partially answers the third research question (RQ3) by providing one way to improve the quality of the results. The chapter also includes a section that discusses extensively how the transition to the use of DGGS could be made, identifying the main challenges as well as the agents that should be responsible for such a transition. The results of this chapter has been presented as part of the article published in the International Journal of Geographical Information Science (Martin-Segura et al., 2024b). The resulting code and dataset from the experiments are published in Figshare[2] To futher support the transition to DGGS, a software library and command line tool was developed, published and registered as `DGGS Tools`[3] and presented in the 27th AGILE Conference in Glasgow(Martin-Segura et al., 2024a);

Finally, during the development of this thesis, two possible use cases have been addressed in parallel where these catalogues could be of great utility. The first of them, developed in collaboration with the Universidade de Coruña, consisted of the integration of a prototype of a modern catalogue into its already existing Software Product Line, with the intention of facilitating the integration of new data sources dynamically in the process of creating software products. This project allowed the implementation of the elements that we consider fundamental for a modern catalogue in a proof of concept.

Secondly, we had the opportunity to contribute to the development of a voice interface for a Web Map Viewer, which was released and registered as `VUI-Prototype`[4], and our research findings were published in Applied Sciences journal (Blanco et al., 2023). This was a first step for a second iteration, integrating Large Language Models

[2] `https://figshare.com/articles/dataset/An_empirical_study_of_the_limitations_of_Minimum_Bounding_Boxes_for_defining_the_extent_of_geospatial_resources_the_use_of_DGGS_and_other_alternatives_for_improving_the_performance_of_spatial_searches/23531610`

[3] `https://doi.org/10.5281/zenodo.10659260`

[4] `github.com/IAAA-Lab/VUI_Prototype`

to improve the comprehension capabilities and widen the range of possible actions. The results of this second iteration were presented in the CAEPIA 2024 Conference in Coruña (López-Franco et al., 2024). One of the new actions that are still being developed is the possibility of searching and loading data automatically in the viewer by voice. We consider this to be a *killer feature* that demonstrates the possibilities of a catalogue like the one we propose.

5.2 Future Work

The future work for this line of research has already been sparsely commented on in Section 2.7, Section 3.6 and Section 4.6. In summary, this thesis has presented several of the weak points that make current catalogues ineffective for discovery. The straight forward path would be to implement and deploy a catalogue that includes: 1) A better description of resources, 2) more responsible management policies for the health of hyperlinks 3) and an indexing and search system that minimizes erroneous results.

That could result in a very valuable tool for the entire community of spatial data users, leading to a new generation of spatial data catalogues with the potential to enable new use cases. Effective discovery and access to spatial resources would unlock the participation of autonomous agents that could integrate, analyze, or visualize this data. Tasks of quality measurement, analysis, and prospective of the datasets themselves could be automated and performed more efficiently. Access to large collections of datasets for research and education would be facilitated. These large collections could be used to train machine learning models, or to perform large-scale data analysis. Enrichment and reuse of open data would be facilitated, and the creation of new services and applications that take advantage of this data would be encouraged.

Another line of future work would be to continue developing the two use cases mentioned above—the Software Product Line of the Universidade de Coruña and the voice interface for a Web Map Viewer—as well as to explore new use cases.

As a conclusion, we find it necessary to transition to a new way of sharing and discovering spatial data, allowing users to find and access resources more efficiently and accurately. It is the work of the entire spatial data community, from data providers, software developers, researchers, and users, to work together to make this transition possible, but we believe that the benefits of this transition are worth the effort.

Chapter 6

Bibliography

Adar, E., Teevan, J., Dumais, S. T., and Elsas, J. L. (2009). The web changes everything: Understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 282–291, New York, NY, USA. Association for Computing Machinery.

Andrews, K., Kappe, F., and Maurer, H. (1996). The hyper-g network information system. In Maurer, H., Calude, C., and Salomaa, A., editors, *J.UCS the Journal of Universal Computer Science: Annual Print and CD-ROM Archive Edition Volume 1 1995*, pages 206–220. Springer Berlin Heidelberg, Berlin, Heidelberg.

Beard, K. and Sharma, V. (1997). Multidimensional ranking for data in digital spatial libraries. *International Journal on Digital Libraries*, 1(2):153–160.

Beaujardière, J., editor (2002). *Web Map Server Implementation Specification. Version 1.1.1.* Open Geospatial Consortium Inc (Open GIS Consortium Inc), OpenGIS project document 01-068r3.

Béjar, R., Lacasta, J., Lopez-Pellicer, F. J., and Nogueras-Iso, J. (2023). Discrete Global Grid Systems with quadrangular cells as reference frameworks for the current generation of Earth observation data cubes. *Environmental Modelling & Software*, 162:105656.

Béjar, R., Latre, M. Á., Lopez-Pellicer, F. J., Nogueras-Iso, J., and Zarazaga-Soria, F. J. (2019). On the problem of providing unique identifiers for areas with any shape on Discrete Global Grid Systems. In *Accepted Short Papers and Posters from the 22nd AGILE Conference on Geo-information Science*, Limassol, Cyprus.

Blanco, T., Martín-Segura, S., de Larrinzar, J. L., Béjar, R., and Zarazaga-Soria, F. J. (2023). First Steps toward Voice User Interfaces for Web-Based Navigation of Geographic Information: A Spanish Terms Study. *Applied Sciences*, 13(4):2083.

Brewington, B. E. and Cybenko, G. (2000). Keeping up with the changing Web. *Computer*, 33(5):52–58.

Burnhill, P., Mewissen, M., and Wincewicz, R. (2015). Reference rot in scholarly statement: Threat and remedy. *Insights the UKSG journal*, 28(2):55–61.

Béjar, R., Nogueras-Iso, J., Latre, M., Muro-Medrano, P., and Zarazaga-Soria, F. (2009). *Handbook of Research on Digital Libraries: Design, Development, and Impact*, chapter Digital Libraries as a Foundation of Spatial Data Infrastructures, pages 390–399. IGI Global, Singapore. ISBN 978-1-59904-879-6.

Cai, G. (2011). Relevance ranking in Geographical Information Retrieval. *SIGSPATIAL Special*, 3(2):33–36.

Caldwell, D. R. (2005). Unlocking the Mysteries of the Bounding Box. *Coordinates: Online Journal of the Map and Geography Round Table, American Library Association.*

Cantán, O., Nogueras-Iso, J., and Zarazaga-Soria, F. (2009). *Handbook of Research on Digital Libraries: Design, Development, and Impact*, chapter DL and GIS: Path to a new collaboration paradigm. Chapter XL, pages 390–399. IGI Global, Singapore. ISBN 978-1-59904-879-6.

Casserly, M. F. and Bird, J. E. (2003). Web Citation Availability: Analysis and Implications for Scholarship | Casserly | College & Research Libraries. *American Communication Journal*, 9(2).

Chen, L., Shang, S., Yang, C., and Li, J. (2020). Spatial keyword search: A survey. *GeoInformatica*, 24(1):85–106.

Chen, Z., Wang, X., and Liu, W. (2022). Reverse keyword-based location search on road networks. *GeoInformatica*, 26(1):201–231.

Cho, J. and Garcia-Molina, H. (2000). The Evolution of the Web and Implications for an Incremental Crawler. In *Conf. on Very Large Databases*, page 18.

Coleman, D. J. and Nebert, D. D. (1998). Building a North American Spatial Data Infrastructure. *Cartography and Geographic Information Systems*, 25(3):151–160.

Cox, S., Daisey, P., Lake, R., Portele, C., and Whiteside, A., editors (2003). *OpenGIS Geography Markup Language (GML) Implementation Specification, Version 3.0.* Open Geospatial Consortium Inc (Open GIS Consortium Inc), OpenGIS Project Document OGC 02-023r4.

Dangermond, J. and Goodchild, M. F. (2020). Building geospatial infrastructure. *Geo-spatial Information Science*, 23(1):1–9.

Davis, PM. and Cohen, SA. (2001). The effect of the Web on undergraduate citation behavior 1996-1999. *Journal of the American Society for Information Science and Technology*, 52:309–314.

DCMI (2004). Homepage of the Dublin Core Metadata Initiative. Dublin Core Metadata Initiative (DCMI), http://www.dublincore.org.

DCMI Usage Board (2020). DCMI Metadata Terms. Standard http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/, Dublin Core Metadata Initiative.

de Berg, M., Cheong, O., van Kreveld, M., and Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*. Springer, Berlin, Heidelberg.

Degbelo, A. and Teka, B. B. (2019). Spatial search strategies for open government data: A systematic comparison. In *Proceedings of the 13th Workshop on Geographic Information Retrieval*, GIR '19, pages 1–10, New York, NY, USA. Association for Computing Machinery.

Dimitrova, D. V. and Bugeja, M. (2007). Raising the Dead: Recovery of Decayed Online Citations. *American Communication Journal*, 9(2).

European Commission (2003). Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. Technical report.

European Data Portal (2021). Metadata Quality Dashboard - Methodology. https://www.europeandataportal.eu/mqa/methodology?locale=en#.

European Union (2016). GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe. Standard.

Evans, J., editor (2003). *Web Coverage Service (WCS), v1.0*. Open Geospatial Consortium Inc, OGC 03-065r6.

Fernández, P., Béjar, R., Latre, M. Á., Valiño, J., Bañares, J. A., and Muro-Medrano, P. R. (2000). Web mapping interoperability in practice, a java approach guided by the opengis web map server interface specification. In *Proc. of the 6th European Commission GI&GIS Workshop, The Spatial Information Society - Shaping the Future*, Lyon, France.

Fetterly, D., Manasse, M., Najork, M., and Wiener, J. L. (2004). A large-scale study of the evolution of Web pages. *Software: Practice and Experience*, 34(2):213–237.

Freed, N., Innosoft, Lannom, N., and First Virtual (1996). Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. RFC 2045, RFC Editor / RFC Editor.

Frontiera, P., Larson, R., and Radke, J. (2008). A comparison of geometric approaches to assessing spatial similarity for GIR. *International Journal of Geographical Information Science*, 22(3):337–360.

Gao, Y., Jiang, D., Zhong, X., and Yu, J. (2016). A Point-Set-Based Footprint Model and Spatial Ranking Method for Geographic Information Retrieval. *ISPRS International Journal of Geo-Information*, 5(7):122.

Gertler, A. L. and Bullock, J. G. (2017). Reference rot: An emerging threat to transparency in political science. *PS - Political Science and Politics*, 50(1):166–171.

Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., and Petras, V. (2006). GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In Peters, C., Gey, F. C., Gonzalo, J., Müller, H., Jones, G. J. F., Kluck, M., Magnini, B., and de Rijke, M., editors, *Accessing Multilingual Information Repositories*, Lecture Notes in Computer Science, pages 908–919, Berlin, Heidelberg. Springer.

Gibb, R. G. (2016). The rHEALPix Discrete Global Grid System. In *IOP Conference Series: Earth and Environmental Science*, volume 34 of *IOP Conference Series: Earth and Environmental Science*, page 012012.

GO FAIR (2021). FAIR Principles. https://www.go-fair.org/fair-principles/.

Gustavo Niemeyer (2008). Geohash.org is public! https://web.archive.org/web/20080305223755/http://blog.labix.org/#post-85.

Harter, S. P. . K. (1997). ARCHIVE: Electronic Journals and Scholarly Communication: A Citation and Reference Study. *Journal of Electronic Publishing*, 3(2).

Hill, L. L. (1990). *Access to Geographic Concepts in Online Bibliographic Files: Effectiveness of Current Practices and the Potential of a Graphic Interface*. PhD thesis, University of Pittsburgh.

Hubner, S., Spittel, R., Visser, U., and Vogele, T. J. (2004). Ontology-Based Search for Interactive Digital Maps. *IEEE Intelligent Systems*, 19(3):80–86.

Ingham, D., Caughey, S., and Little, M. (1996). Fixing the "Broken-Link" problem: The W3Objects approach. *Computer Networks and ISDN Systems*, 28(7-11):1255–1268.

INSPIRE (2020). Geoportal workflow for establishing links between data sets and network services. Technical report.

INSPIRE MIG (2017). Technical Guidelines for implementing dataset and service metadata based on ISO/TS 19139:2007. INSPIRE Maintenance and Implementation Group (MIG). Version 2.0.1. Standard.

ISO (2003). Geographic information - Metadata. ISO 19115:2003, International Organization for Standardization.

ISO Central Secretary (2012). Information and documentation – Digital object identifier system. Standard ISO 26324:2012, International Organization for Standardization, Geneva, CH.

ISO Central Secretary (2014). Geographic information – Metadata – Part 1: Fundamentals (2014). Standard ISO 19115-1:2014, International Organization for Standardization, Geneva, CH.

Jaloliddinov, J., Tian, X., Bai, Y., Guo, Y., Chen, Z., Li, Y., and Wang, S. (2024). Large-Scale Cotton Classification under Insufficient Sample Conditions Using an Adaptive Feature Network and Sentinel-2 Imagery in Uzbekistan. *Agronomy*, 14(1):75.

JGSI (2002). User's Manual for Spatial Data Product Specification Description. Technical Report 264.

Kessler, G. (2002). File Signatures. https://www.garykessler.net/library/file_sigs.html.

Klein, M., Van De Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., and Tobin, R. (2014). Scholarly context not found: One in five articles suffers from reference rot. *PLoS ONE*, 9(12).

Klump, J., Huber, R., and Diepenbroek, M. (2016). DOI for geoscience data - how early practices shape present perceptions. *Earth Science Informatics*, 9(1):123–136.

Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 50(2):162–180.

Koehler, W. (2002). Web page change and persistence—A four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53(2):162–171.

Kotsev, A., Minghini, M., Cetl, V., Penninga, F., Robbrecht, J., and Lutz, M. (2021). INSPIRE - a public sector contribution to the european green deal data space. Scientific Analysis or Review, Policy Assessment, Anticipation and Foresight, Technical Guidance KJ-NA-30832-EN-N (online),KJ-NA-30832-EN-C (print), Publications Office of the European Union, Luxembourg (Luxembourg).

Kottman, C., editor (1999). *The OpenGIS Abstract Specification. Topic13: Catalog Services (version 4)*. Open Geospatial Consortium Inc (Open GIS Consortium Inc), OpenGIS Project Document 99-113.

Kügeler, A. and Jirka, S. (2021). Geospatial Trends 2021. Technical report, data.europa.eu.

Lacasta, J., Lopez-Pellicer, F. J., Espejo-García, B., Nogueras-Iso, J., and Zarazaga-Soria, F. J. (2017). Aggregation-based information retrieval system for geospatial data catalogs. *International Journal of Geographical Information Science*, 31(8):1583–1605.

Larson, R. R. (2009). Geographic Information Retrieval and Digital Libraries. In Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., and Tsakonas, G., editors, *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 461–464, Berlin, Heidelberg. Springer.

Larson, R. R. and Frontiera, P. (2004). Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries. In Heery, R. and Lyon, L., editors, *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 45–56, Berlin, Heidelberg. Springer.

Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (2005). *Geographical Information Systems and Science*. John Wiley & Sons.

López-Franco, R., Martin-Segura, S., and Zarazaga-Soria, F. J. (19–21 June 2024, 2024). Estudio de las capacidades geográficas de los LLMs: Integración de GPT 3.5 en la interfaz de voz de un visor de mapas. In *XX Conferencia de La Asociación*

Española Para La Inteligencia Artificial, CAEPIA 2024, pages 95–100, A Coruña, Spain.

Maganto, A. S., Iso, J. N., and Ballari, D. (2008). Normas sobre metadatos (ISO19115, ISO19115-2, ISO19139, ISO 15836). *Mapping*, 123:48–57. ISSN 1-131-9-100.

Martin-Segura, S., Béjar, R., and Zarazaga-Soria, F. J. (2024a). DGGSTools: An Open Source Python Package for the Manipulation of Vector and Raster Datasets in the rHEALPix Discrete Global Grid System. *AGILE: GIScience Series*, 5:1–6.

Martin-Segura, S., Lopez-Pellicer, F. J., Béjar, R., Nogueras-Iso, J., and Zarazaga-Soria, F. J. (2024b). An empirical study of the limitations of minimum bounding boxes for defining the extent of geospatial resources: The use of DGGS and other alternatives for improving the performance of spatial searches. *International Journal of Geographical Information Science*, 0(0):1–20.

Martín Segura, S., López-Pellicer, F. J., García, J. V., and Zarazaga-Soria, F. J. (2018). Geolake Search (el futuro de las IDE está en mejorar su catálogo). In *REVISTA INTERNACIONAL MAPPING*, volume 28, pages 24–30. Revista Mapping.

Martin-Segura, S., Lopez-Pellicer, F. J., Nogueras-Iso, J., Lacasta, J., and Zarazaga-Soria, F. J. (2022). The Problem of Reference Rot in Spatial Metadata Catalogues. *ISPRS International Journal of Geo-Information*, 11(1):27.

Martín Segura, S., Lopez-Pellicer, F., Valiño-García, J., and Zarazaga-Soria, F. (2019). A next-generation geospatial catalogue: a proof of concept. In *Proceedings of the 22nd AGILE Conference on Geo-information Science*. Cyprus University of Technology 17-20 June 2019, Limassol, Cyprus.

Masser, I. (2019). *Geographic Information Systems to Spatial Data Infrastructures: A Global Perspective*. Boca Raton, FL.

Nebert, D. (2004). Developing Spatial Data Infrastructures: The SDI Cookbook. Technical report.

Nebert, D., Whiteside, A., and Vretanos, P. (2007). Open GIS Catalogue services specification (version: 2.0. 2). Standard, Open Geospatial Consortium.

Nebert, D. D. (2001). The SDI cookbook. *http://www. gsdi. org/pubs. html*.

Neumaier, S., Umbrich, J., and Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. *Journal of Data and Information Quality*, 8(1):2:1–2:29.

Nielsen, J. (1998). Fighting Linkrot. https://www.nngroup.com/articles/fighting-linkrot/.

Nogueras-Iso, J., Lacasta, J., Ureña-Cámara, M. A., and Ariza-López, F. J. (2021). Quality of Metadata in Open Data Portals. *IEEE Access*, 9:60364–60382.

Nogueras-Iso, J., Latre, M. Á., Muro-Medrano, P. R., and Zarazaga-Soria, F. J. (2004). *Electronic Government, Lecture Notes in Computer Science (LNCS)*, volume 3183 of *Lecture Notes in Computer Science*, chapter Building e-Government services over Spatial Data Infrastructures, pages 387–391. Zaragoza, Spain.

Ntoulas, A., Cho, J., and Olston, C. (2004). What's new on the web? the evolution of the web from a search engine perspective. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Olfat, H., Kalantari, M., Rajabifard, A., Williamson, I. P., Pettit, C., and Williams, S. (2010). Exploring the key areas of spatial metadata automation research in Australia. Leuven University Press.

Parliament, E. and Council, T. E. (2007). Directive of the european parliament and of the council establishing an infrastructure for spatial information in the european community (inspire). joint text approved by the conciliation committee provided for in article 251(4) of the ec treaty. 2004/0175(cod), pe-cons 3685/06.

Perry, J. W., Kent, A., and Berry, M. M. (1955). Machine literature searching X. Machine language; factors underlying its design and development. *American Documentation*, 6(4):242–254.

Philipp Bunge, Aner Ben-Artzi, Jarda Bengl, Prasenjit Phukan, and Sacha van Ginhoven (2018). Open Location Code. https://web.archive.org/web/20180301114837/http://openlocationcode.com/.

PROJ contributors (2023). *PROJ Coordinate Transformation Software Library*.

Purss, M. B. J., Gibb, R., Samavati, F., Peterson, P. R., and Ben, J. (2016). The OGC® Discrete Global Grid System core standard: A framework for rapid geospatial integration. pages 3610–3613.

Purves, R., Clough, P., Jones, C., Hall, M., and Murdock, V. (2018). Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends® in Information Retrieval*, 12:164–318.

Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A. K., Vaid, S., and Yang, B. (2007). The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7):717–745.

Quarati, A., De Martino, M., and Rosim, S. (2021). Geospatial Open Data Usage and Metadata Quality. *ISPRS International Journal of Geo-Information*, 10(1):30.

Rajabifard, A., Kalantari Soltanieh, S., and Binns, A. (2009). SDI and Metadata Entry and Updating Tools. In *GSDI 11 World Conference*. GSDI Association.

Rao, M. V., Babu, V. S., Chandra, S., and Chary, G. R. (2015). Need for Cadastral Level Land Resource Information for Sustainable Development—A Case Study in Chikkarasinakere Hobli, Mandya District, Karnataka. In *Integrated Land Use Planning for Sustainable Agriculture and Rural Development*, pages 67–92. Apple Academic Press, 0 edition.

Renteria-Agualimpia, W., Lopez-Pellicer, F. J., Lacasta, J., Muro-Medrano, P. R., and Zarazaga-Soria, F. J. (2015). Identifying geospatial inconsistency of web services metadata using spatial ranking. *Earth Science Informatics*, 8(2):427–437.

Rhodes, J. S. (2002). Web Sites That Heal. http://web.archive.org/web/20160315090512/http://www.webword.com/moving/healing.html.

Roth, J. (2011). The Approximation of Two-Dimensional Spatial Objects by Two Bounding Rectangles. *Spatial Cognition & Computation*, 11(2):129–152.

Sahr, K., White, D., and Kimerling, A. J. (2003). Geodesic Discrete Global Grid Systems. *Cartography and Geographic Information Science*, 30(2):121–134.

Sanderson, R., Van de Sompel, H., Burnhill, P., and Grover, C. (2013). Hiberlink: Towards time travel for the scholarly web. In *Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts*, DPRMA '13, page 21, New York, NY, USA. Association for Computing Machinery.

Sharma, P., Samal, A., Soh, L.-K., and Joshi, D. (2022). A spatially-aware algorithm for location extraction from structured documents. *GeoInformatica*.

Sife, A. S. and Bernard, R. (2013). Persistence and decay of web citations used in theses and dissertations available at the Sokoine National Agricultural Library, Tanzania. Technical Report 2.

Sun, S., Reilly, S., Lannom, L., and Petrone, J. (2003). Handle System Protocol (ver 2.1) Specification. RFC 3652, RFC Editor / RFC Editor.

The Internet Archive (2001). Wayback Machine. https://web.archive.org/.

Tyler, D. C. and McNeil, D. C. B. (2003). Librarians and Link Rot: A Comparative Analysis with Some Methodological Considerations.

United Nations GGIM (2021). Geospatial Industry Advancing the SDGs. Technical report, United Nations initiative on Global Geospatial Information Management.

Ureña-Cámara, M. A., Nogueras-Iso, J., Lacasta, J., and Ariza-López, F. J. (2019). A method for checking the quality of geographic metadata based on ISO 19157. *International Journal of Geographical Information Science*, 33(1):1–27.

U.S. Federal Register (1994). Executive Order 12906. Coordinating Geographic Data Acquisition and Access: the National Spatial Data Infrastructure (U.S.). *The April 13,1994, Edition of the Federal Register*, 59(71):17671–17674.

Van de Sompel, H., Nelson, M., and Sanderson, R. (2013). HTTP Framework for Time-Based Access to Resource States – Memento. RFC 7089, RFC Editor / RFC Editor.

Vretanos, P. A., editor (2002). *Web Feature Server Implementation Specification. Version 1.0.0.* Open Geospatial Consortium Inc (Open GIS Consortium Inc), OpenGIS project document OGC 02-058.

W3C (2020). Data Catalog Vocabulary (DCAT) - Version 2. Standard 2, W3C.

Walker, D. R. F., Newman, I. A., Medyckyj-Scott, D. J., and Ruggles, C. L. N. (1992). A system for identifying datasets for GIS users. *International Journal of Geographical Information Systems*, 6(6):511–527.

WebCite Consortium (1998). WebCite. https://www.webcitation.org/.

What3words Limited (2023). What3words | About us. https://what3words.com/en/about.

Wren, J. D., Johnson, K. R., Crockett, D. M., Heilig, L. F., Schilling, L. M., and Dellavalle, R. P. (2006). Uniform Resource Locator Decay in Dermatology Journals: Author Attitudes and Preservation Practices. *Archives of Dermatology*, 142(9).

Xu, Y., Xie, Z., Chen, Z., and Xie, M. (2021). Measuring the similarity between multipolygons using convex hulls and position graphs. *International Journal of Geographical Information Science*, 35(5):847–868.

Zarazaga-Soria, F. J., Nogueras-Iso, J., Béjar, R., and Muro-Medrano, P. R. (2004). *Electronic Government, Lecture Notes in Computer Science (LNCS)*, volume 3183 of *Lecture Notes in Computer Science*, chapter Political aspects of Spatial Data Infrastructures, pages 392–395. Zaragoza, Spain.

Figures

Tables

Annex

Annex A

Results of the Reference Rot analysis

Table A.1 shows the results of the analysis of the accessibility of the metadata disaggregated by country.

Catalogue	Metadata	Distrib.	Found	*Link Rot*	*Content Drift*	More Needed	No Direct Ac.	No Expect.	No URIs
Luxembourg Catalogue	364	1424	**61.81%**	1.10%	1.37%	25.55%	9.62%	0.55%	0.00%
Swedish Catalogue	183	304	54.10%	2.19%	2.19%	0.55%	32.79%	0.00%	8.20%
Belgian Catalogue (Brussels)	106	223	31.13%	**28.30%**	3.77%	19.81%	7.55%	9.43%	0.00%
Danish Catalogue	1086	645	25.41%	17.96%	21.55%	9.48%	13.63%	9.48%	2.49%
Austrian Catalogue	1160	1357	29.48%	2.93%	8.88%	18.79%	19.05%	15.86%	5.00%
Lettish Catalogue	477	418	0.00%	11.11%	0.00%	0.00%	0.00%	77.15%	11.74%
Czech Catalogue	342	458	9.06%	14.33%	11.99%	19.88%	12.87%	31.29%	0.58%
Romanian Catalogue	179	131	3.35%	17.88%	7.82%	5.03%	15.64%	27.37%	**22.91%**
Croatian Catalogue	310	374	0.65%	9.35%	5.81%	5.48%	33.87%	44.84%	0.00%
Greek Catalogue	80	44	**65.00%**	1.25%	0.00%	0.00%	0.00%	23.75%	10.00%
Bulgarian Catalogue	144	89	0.00%	6.94%	0.00%	1.39%	2.78%	**81.25%**	7.64%
Irish Catalogue	93	57	8.60%	15.05%	19.35%	38.71%	6.45%	11.83%	0.00%
Belgian Catalogue (Federal)	140	169	18.57%	2.86%	5.00%	15.00%	3.57%	19.29%	**35.71%**
Polish Catalogue	423	673	5.91%	**36.41%**	14.42%	4.49%	1.18%	37.59%	0.00%
British Catalogue	1495	1850	0.40%	3.21%	28.36%	2.34%	43.28%	22.41%	0.00%
Estonian Catalogue	201	297	31.34%	2.49%	1.99%	3.98%	14.43%	45.77%	0.00%
Belgian Catalogue (Flanders)	6648	8134	**69.87%**	0.21%	0.47%	10.74%	11.54%	1.43%	5.75%
Belgian Catalogue (Wallonia)	211	476	0.00%	0.47%	0.00%	0.00%	0.00%	**99.05%**	0.47%
Portuguese Catalogue	1213	1305	51.44%	16.49%	4.70%	20.03%	3.79%	0.00%	3.54%
French Catalogue	250	808	32.80%	6.00%	2.40%	31.60%	15.20%	10.80%	1.20%
Liechtensteiner Catalogue	20	19	5.00%	0.00%	5.00%	5.00%	**80.00%**	0.00%	5.00%
Swiss Catalogue	120	227	3.33%	5.83%	9.17%	0.00%	45.00%	6.67%	**30.00%**
Spanish Catalogue	527	942	32.45%	4.36%	1.90%	3.98%	15.56%	41.18%	0.57%
Lithuanian Catalogue	262	352	**71.37%**	0.00%	0.76%	22.90%	4.58%	0.00%	0.38%
Dutch Catalogue	565	720	22.30%	4.25%	4.78%	6.19%	7.26%	54.16%	1.06%
Finnish Catalogue	1457	1246	21.55%	12.77%	2.68%	3.23%	11.87%	42.48%	5.42%
Total	18,054	22,738	40.70%	6.29%	6.21%	10.25%	14.25%	17.74%	4.56%

Table A.1: Results of the Reference Rot analysis