International Journal of Geo-Information

*Article*

# The Problem of Reference Rot in Spatial Metadata Catalogues

**Sergio Martin-Segura *, Francisco Javier Lopez-Pellicer, Javier Nogueras-Iso, Javier Lacasta and Francisco Javier Zarazaga-Soria**

Department of Computer Science and System Engineering, University of Zaragoza, 50018 Zaragoza, Spain; fjlopez@unizar.es (F.-J.L.-P.); jnog@unizar.es (J.N.-I.); jlacasta@unizar.es (J.L.); javy@unizar.es (F.-J.Z.-S.)
* Correspondence: segura@unizar.es

**Abstract:** The content at the end of any hyperlink is subject to two phenomena: the link may break (*Link Rot*) or the content at the end of the link may no longer be the same as it was when it was created (*Content Drift*). *Reference Rot* denotes the combination of both effects. Spatial metadata records rely on hyperlinks for indicating the location of the resources they describe. Therefore, they are also subject to *Reference Rot*. This paper evaluates the presence of *Reference Rot* and its impact on the 22,738 distribution URIs of 18,054 metadata records from 26 European INSPIRE spatial data catalogues. Our *Link Rot* checking method detects broken links while considering the specific requirements of spatial data services. Our *Content Drift* checking method uses the data format as an indicator. It compares the data formats declared in the metadata with the actual data types returned by the hyperlinks. Findings show that 10.41% of the distribution URIs suffer from *Link Rot* and at least 6.21% of records suffer from *Content Drift* (do not declare its distribution types correctly). Additionally, 14.94% of metadata records only contain intermediate HTML web pages as distribution URIs and 31.37% contain at least one HTML web page; thus, they cannot be accessed or checked directly.

**Keywords:** metadata; spatial data infrastructures; *Reference Rot*; *Link Rot*; *Content Drift*



**Citation:** Martin-Segura, S.; Lopez-Pellicer, F.J.; Nogueras-Iso, J.; Lacasta, J.; Zarazaga-Soria, F.J. The Problem of Reference Rot in Spatial Metadata Catalogues. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 27. https://doi.org/10.3390/ijgi11010027

Academic Editor: Wolfgang Kainz

Received: 6 November 2021
Accepted: 26 December 2021
Published: 31 December 2021



## 1. Introduction

Spatial Data Infrastructures (SDI) can be defined as a coordinated approach to technologies, policies and institutional arrangements that facilitate the availability and access to spatial data. SDIs are organized as a hierarchical network of nodes where the main technological components in each node are spatial data, metadata, middle-ware services (to locate, visualize and download data among other purposes) and final user applications. Catalogue services are essential as they provide the mechanism for searching and discovering its geographic data and services [1,2].

A good example of SDI is the INSPIRE directive, an initiative to share a common SDI across the European Union. It has the purpose of facilitating the accessibility and interoperability of the resources whilst focusing on sustainable development. This directive proposes a Technical Guidance (the INSPIRE *Implementation Guidance*) for the implementation of spatial metadata [3]. It proposes the ISO 19115 [4] as the metadata standard and provides some restrictions on how to implement or declare any relevant information about it. The standard describes the resource itself along with their resource distributions. Each distribution describes a different way of accessing the same information and may differ in format (data type), location (URI), and so forth.

An SDI relies entirely on spatial metadata [5]. 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 [6]. Therefore, spatial metadata quality has been studied from different perspectives and frameworks by maintainers, stakeholders, and researchers [7,8].

The way the metadata records refer to their distribution locations, via distribution URIs, meets the conditions outlined for being susceptible to suffering *Reference Rot*. *Reference Rot* [9] denotes the combination of two problems: *Link Rot* and *Content Drift*. *Link Rot* (or *Broken Link*) occurs when an URI no longer gives access to resource representations since the resource has been moved or deleted. For example, a metadata record that has an URI pointing to a web map service that was shut down years ago will suffer form *Link Rot*. *Content Drift* occurs when an URI returns resource representations that do not represent the resource that was intended to be referenced by that URI. For example, a metadata record that describes a Web Map Service whose URI 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 STM articles [9], 58% of web citations in Agricultural Library [10] or 27% of *American Political Science Review* links [11]).

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 paper.

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 paper, we made the following contributions:

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 URIs from 26 officially registered INSPIRE Discovery Services of EU and EFTA 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*.

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. 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. We only use the data types as *Content Drift* indicator. Therefore, any other mismatch between the metadata and the resource, such as dates, spatial data extent and so forth, cannot be not detected. We only fetched the partial content of the HTTP responses (the reasons are detailed in Section 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 4.4).

The rest of this paper is organized as follows. First, we collect some related works in Section 2. In Section 3, we analyze the multiple dimensions of the problem and the challenges we will face. In Section 4, we describe the methodology we designed to detect *Reference Rot* in spatial catalogues along with the details of the experiment. In Section 5, we

present the results of the execution over real catalogues with a brief analysis. In Section 6 we discuss these results. Finally, we expose the conclusions and future works in Section 7.

## 2. Related Work

Since the early days of the Web, researchers realized that *Link Rot* was one of its notorious problems [12,13]. Various studies have been conducted over time on different corpora: web, digital libraries metadata, academic electronic journals, legal documents, and so forth. [6,10,14–19]. First experiments reported different degrees of incidence of *Link Rot*, varying from 18.3% for URIs in dermatology journals from 1999 to 2004 [18], to 81% for three-year-old references in undergraduate term papers in 2000 [16]. 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* [12,20] 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 URI maintenance policies; and (3) a lack of synchronization between authors and metadata curators.

Rajabifard et al. [21] 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. [22] 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 [23–28]. One of the main findings was that some types of data are more likely to change than others. For example, HTML pages change more frequently than PDF files.

The Hiberlink Project [29] 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 [9,30].

The closest approach for measuring *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 [31], the *Metadata Quality Assessment (MQA)* by the European Data Portal [32], and the Dataset-Service Linkage Service by INSPIRE [33].

Both ODPW and MQA work with general purpose metadata that follow the Data Catalogue Vocabulary (DCAT) schema [34], 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 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. [35] 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 URIs were broken and only 52.61% of the resources matched their declared type.

## 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 URIs. Distributions contain information about its distributors, its online locations, and its formats. The online locations are the place where

the access URIs are declared. The formats define the expected data types or protocols in which the resource will be served.

*Distribution Link Rot* happens when the URI included in a specific distribution cannot retrieve any content. This manifests as a *connection error* (i.e., `invalid URI` 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 URI. Overlooking *Link Rot* leads not only to a detriment in the usefulness of the distribution URIs 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 URIs. The INSPIRE *Implementation Guidance* suggests including full URIs 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 URI 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 URI 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 measuring this fenomenon. This way, we detect *Distribution Content Drift* as a mismatch between the expected resource format and the real one found in the URI. *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 SDIs 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 Types [36] 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 *"URI-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 URI may not match the declared format when the URI 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.

## 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 standard [37]. The whole process involved the following steps:

1. *Link extraction*. Identify URIs 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 URIs 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 (see details in Section 4.5).
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.

### *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. The retrieved metadata records were stored for further processing.

### *4.2. Link and Format Extraction*

Stored records were parsed to extract each potential distribution URI, 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 URI 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 4.7.3.

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.

### 4.3. Request Phase

In this phase, we performed an HTTP request to each extracted distribution URI. Once the URI 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 URI 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, URIs 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.

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 [38]. 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 Sections 4.4 and 4.5.2.

### 4.4. Type Guessing Phase

In this phase, we analyzed the content of the HTTP responses for each distribution URI using the tool *Libmagic* to guess its file format. Then we mapped the inferred file format to our controlled vocabulary (see Section 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 patters that denote other spatial formats such as GeoJSON. Finally, we tried to detect some common *OGC Error* messages that most OGC Service implementations return. This is necessary and useful as explained in the next Section 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 4.5.2).

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 4.2. We can directly compare these keywords because we used the same controlled vocabulary.

### 4.5. Spatial Specific Cases

In Section 3 we explained how a *naive* approach to *Reference Rot* measuring may report wrong results for Spatial Metadata Catalogues. In this phase, we explain the methods used to manage these situations.

#### 4.5.1. Incomplete Service URIs

A common case of *Link Rot false positives* may happen when the distribution URI 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 URIs we previously detected as *OGC Errors* in the type guessing phase. Then, the new URIs were requested and guessed again.

#### 4.5.2. 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 URI 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 URI returned a `png`.
- The distribution URI returned a compressed file whose content type we could not guess. That is, distribution URI 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 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.

**Table 1.** Decision Table.

| 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* |

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.

#### 4.5.3. 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` URI too. We also considered as *wrong OK status* the `empty` responses that returned an `HTTP OK` status code.

### *4.6. Metadata Reference Rot Analysis*

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

We classified each metadata record based on the combination of metadata-wide *Link Rot* and *Content Drift* analysis (see Figure 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 URI and its type is correctly declared;
- No direct access: The record has at least one accessible URI but none of them provides direct access to the resource;
- More data needed: Covers any of the scenarios explained in Section 4.5.2 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 URIs;
- No expectations: The record does not declare any data type;
- *Link Rot*: All URIs are broken;
- Without Links: The resource has no URIs.

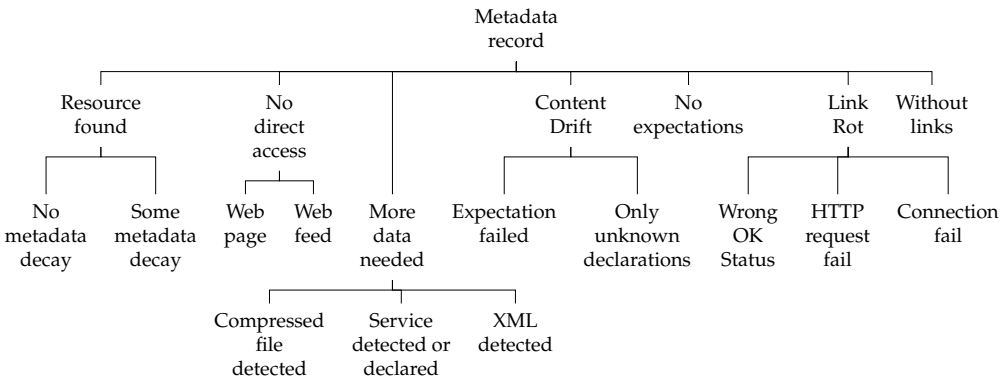


**Figure 1.** *Metadata Reference Rot* Categories.

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 URI to the resource;
- Web feed: An Atom or RSS that may contain an URI 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;
- 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 URIs are broken but some obtained an HTTP error response from the server;
- Connection fail: All URIs are broken and none of the succeed to connect to any server.

### *4.7. Experiment*

#### 4.7.1. Metadata Collection

For this experiment, we used 26 out of 35 officially registered INSPIRE Discovery Services of EU and EFTA countries at INSPIRE Geoportal (https://inspire-geoportal.ec.europa.eu/harvesting_status.html, accessed on 1 September 2021).

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 URI and not declaring it correctly. The selected catalogues are listed in Table A1.

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.

#### 4.7.2. Distribution Count and Temporal Perspective

As each metadata record can have zero, one or many distribution URIs, we analyzed the number of distributions provided by each metadata record. The results can be seen in Table 2. The most common case is a metadata record (28.64%) with one distribution URI. Next, the second most common case is a metadata record (20.79%) with 7 distribution

URIs. 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 URIs. Finally, 4.56% of metadata records have zero distributions.

**Table 2.** Distribution count on metadata records.

| 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% |

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.

#### 4.7.3. Declared Data Types

As explained in Section 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 (see Table A2) 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 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.

## 5. Results

This section describes the results of the execution of the analysis process.

### *5.1. Link Rot*

Table 3 shows the detailed response status code count for each unique distribution URI. By unique, we mean that a URI 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 URIs returned an `HTTP OK` status code while the content of the resource suggests an OGC Error (see Section 4.5.3).

**Table 3.** Distribution URI status.

| 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 | 20,372 | 89.59% |
| | *Wrong OK status* | *Wrong OK status* | 315 | 1.39% |

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 URIs 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.

### 5.2. Resource Types

Table 4 shows the distribution types obtained in the type guessing process. A more detailed table can be found in Table A2. 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 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 guessed records combining the `OGC Errors` and the few `empty responses`. The rest of the data types are, as expected, spatial datasets and services.

**Table 4.** Resource types (overview).

| 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% |

### 5.3. Reference Rot Presence over Time

Figure 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*);
- 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 URIs 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 URIs. 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.

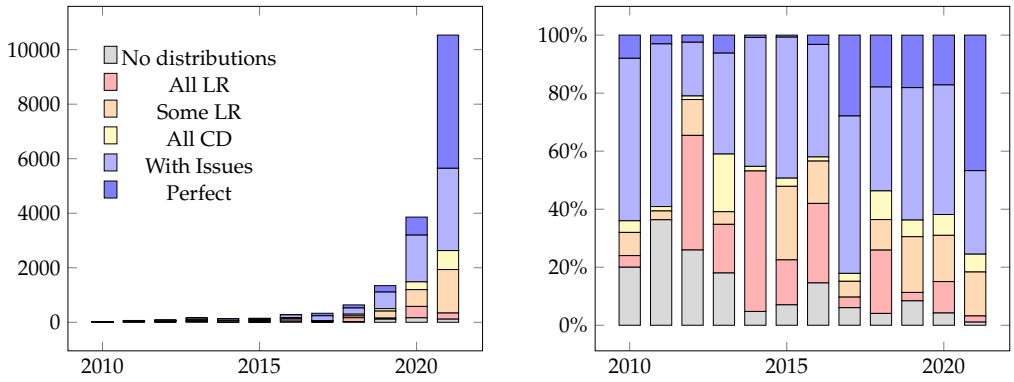


**Figure 2.** *Reference Rot* presence over time.

### 5.4. Metadata Wide Reference Rot

Figure 3 shows the metadata-wide *Reference Rot* categories we defined previously in Section 4.6.

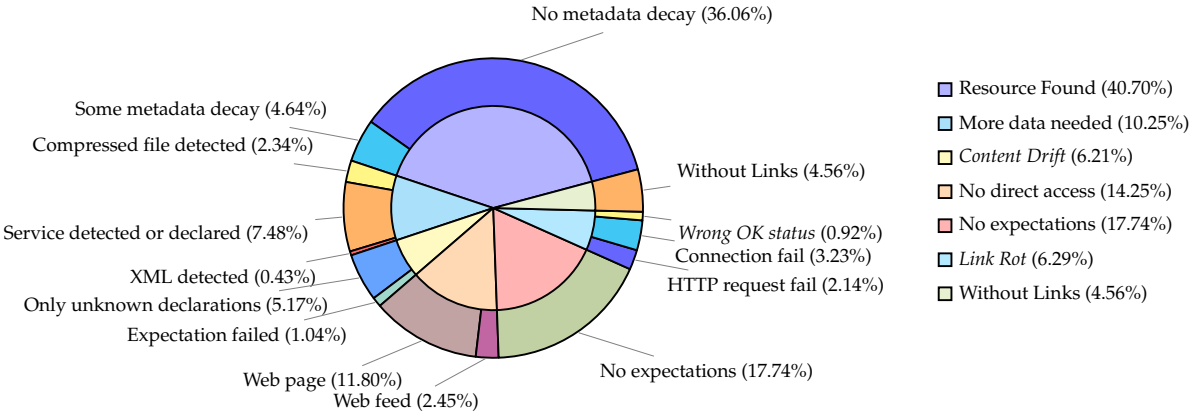


**Figure 3.** Metadata-wide*Reference Rot*.

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. The other 35% will be analyzed below.

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. 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 article. Of the 6.29% of metadata records with all their URIs 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.

Table A3 breaks down the results for each individual catalogue. This reveals that each one has its own metadata practices and policies, and each strategy affects its data accessibility differently.

For example, 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 URIs, 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`.

### *5.5. Indirect Access*

Section 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 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).

## 6. Discussion

The SDI literature always highlights that a *Spatial Catalogue* is an essential component for discovering and sharing datasets and services. In this study 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 article, more than 10% of the distribution URIs 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 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` URI 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 URI (sometimes this is difficult when there is a lack of context). It also hides the final distribution URI

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 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 4.7.3, 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 [39] 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 URI 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 URI 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 [35]. 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 [6]. 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 [12] or Hyper-G [40], which tried to maintain referential integrity in ultra-large-scale web-based systems.

Systems such as Handle [41] and its subsystem DOI [42] 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. [43] discussed the relevance of giving DOI identifiers 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 [44] (nowadays, the largest web pages' snapshot archive), The Memento Protocol [45] (a protocol for accessing web page snapshots

compatible with the Wayback Machine, among others) and WebCite [46] (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.

## 7. Conclusions

In this study, 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 URIs 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 service URIs. This reveals gaps in the usefulness of current metadata for tasks beyond description and management, such as discovery and access to resources.

Further studies may perform this analysis over larger and less curated catalogues to compare the results, expecting a lower quality. They 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. 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. Repeating the experiment fetching the whole response contents would increase the storage and time requirements but also the quality of the type guessing. A more mature and robust *spatial data type guessing* tool could also be implemented.

We want the spatial data community, and especially the stakeholders involved in the administration of spatial data catalogues, to become aware of these issues. We look for better spatial catalogues that allow humans and automatic user agents to discover, access and re-use spatial resources. We believe that those are the building blocks for the spatial information systems of the future.

**Author Contributions:** Methodology, software and writing—original draft preparation, Sergio Martin-Segura; Conceptualization, validation and writing—review and editing, Francisco Javier Lopez-Pellicer; resources, formal analysis and writing—review and editing Javier Nogueras-Iso and Javier Lacasta; supervision, project administration and funding acquisition, Francisco Javier Zarazaga-Soria. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are openly available in FigShare at https://figshare.com/s/d4af7b49a89d2e2562fe, accessed on 5 November 2021.

**Acknowledgments:** This paper is part of the R&D project PID2020-113353RB-I00, supported by the Spanish MCIN/ AEI/10.13039/501100011033/ and the project T59_20R supported by the Aragon regional Government.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyzes, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results

## Appendix A

**Table A1.** Catalogues under study.

| Country | Count | Internal Codename | Discovery Service Title |
|---|---|---|---|
| AT | 1160 | *Austrian Catalogue* | INSPIRE Suchdienst Österreich |
| BE | 211 | *Belgian Catalogue (Wallonia)* | Service de découverte pour la Wallonie |
| BE | 6648 | *Belgian Catalogue (Flanders)* | Metadata Vlaanderen CSW |
| BE | 140 | *Belgian Catalogue (Federal)* | Federal Discovery Service |
| BE | 106 | *Belgian Catalogue (Brussels)* | Geobru_inspire |
| BG | 144 | *Bulgarian Catalogue* | Bulgarian INSPIRE Discovery Service |
| CH | 120 | *Swiss Catalogue* | geocat.ch |
| CZ | 342 | *Czech Catalogue* | Národní geoportál INSPIRE |
| DK | 1086 | *Danish Catalogue* | Geodata-info søgetjeneste |
| EE | 201 | *Estonian Catalogue* | Estonian INSPIRE Discovery Service |
| EL | 80 | *Greek Catalogue* | Greek geospatial data catalogue |
| ES | 527 | *Spanish Catalogue* | Spanish Official Catalogue of INSPIRE Dataset and Services |
| FI | 1457 | *Finnish Catalogue* | Paikkatietohakemiston CSW-rajapinta |
| FR | 250 | *French Catalogue* | Geocatalogue Catalogue Server Priority data |
| HR | 310 | *Croatian Catalogue* | NIPP kataloška usluga |
| IE | 93 | *Irish Catalogue* | Irish CSW |
| LI | 20 | *Liechtensteiner Catalogue* | geocat.ch direct partners (LI) |
| LT | 262 | *Lithuanian Catalogue* | GIS-Centras metadata catalogue |
| LU | 364 | *Luxembourg Catalogue* | Luxemburg's national official geoportal |
| LV | 477 | *Lettish Catalogue* | Latvijas metadatu katalogs. ĢDS |
| NL | 565 | *Dutch Catalogue* | CSW Nationaal Georegister |
| PL | 423 | *Polish Catalogue* | Geoportal—Polska Usługa Wyszukiwania INSPIRE |
| PT | 1213 | *Portuguese Catalogue* | Direção-Geral do Território |
| RO | 179 | *Romanian Catalogue* | Serviciul de căutare al geo-portalului INSPIRE al României |
| SI | 183 | *Swedish Catalogue* | Inspire (SI) (Geodetska uprava RS) |
| UK | 1495 | *British Catalogue* | GEMINI—CSW Server |
| Total | 18,054 | | |

## Appendix B

**Table A2.** Data types.

| Type | Name | Count | Ratio |
|---|---|---|---|
| Spatial Data Type (42.09%) | kml | 5221 | 25.24% |
| | shp | 1093 | 5.28% |
| | gml | 665 | 3.21% |
| | jpeg | 452 | 2.18% |
| | png | 319 | 1.54% |
| | geojson | 233 | 1.13% |
| | pdf | 173 | 0.84% |
| | text | 128 | 0.62% |
| | csv | 107 | 0.52% |
| | tiff | 103 | 0.50% |
| | geopackage | 82 | 0.40% |
| | netcdf | 76 | 0.37% |
| | mapinfo | 31 | 0.15% |
| | ole | 10 | 0.05% |

**Table A2.** *Cont.*

| Type | Name | Count | Ratio |
|---|---|---|---|
| | sgml | 8 | 0.04% |
| | geodatabase | 4 | 0.02% |
| | edigeo | 2 | 0.01% |
| Spatial Service Type (28.51%) | wms | 3752 | 18.14% |
| | wfs | 1956 | 9.46% |
| | ows | 138 | 0.67% |
| | wcs | 50 | 0.24% |
| | wps | 1 | 0.00% |
| Indirect Medium (25.94%) | html | 4245 | 20.52% |
| | atom | 931 | 4.50% |
| | metadata | 190 | 0.92% |
| Undecidable Content (1.94%) | compressed | 356 | 1.72% |
| | xml | 44 | 0.21% |
| | OTHER | 2 | 0.01% |
| *Wrong OK status* (1.52%) | OGC Error | 307 | 1.48% |
| | Empty response | 8 | 0.00% |

## Appendix C

Table A3 shows the *Metadata Reference Rot* per catalogue.

**Table A3.** Metadata-wide *Reference Rot* per Catalogue.

| 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% |

## References


1. Nebert, D. Interoperable Spatial Data Catalogs. *Photogramm. Eng. Remote Sens.* **1999**, *65*, 3.
2. Nogueras-Iso, J.; Zarazaga-Soria, F.J.; Béjar, R.; Álvarez, P.J.; Muro-Medrano, P.R. OGC Catalog Services: A Key Element for the Development of Spatial Data Infrastructures. *Comput. Geosci.* **2005**, *31*, 199–209. [CrossRef]
3. INSPIRE MIG. *Technical Guidelines for Implementing Dataset and Service Metadata Based on ISO/TS 19139:2007. INSPIRE Maintenance and Implementation Group (MIG)*, Version 2.0.1; Technical Report; INSPIRE MIG: Brussels, Belgium, 2017.
4. *ISO 19115-1:2014*; Geographic Information—Metadata—Part 1: Fundamentals. International Organization for Standardization: Geneva, Switzerland, 2014.
5. Nebert, D.D. The SDI Cookbook. 2001. Available online: http://www.gsdi.org/pubs.html (accessed on 2 February 2021).
6. Tyler, D.C.; McNeil, D.C.B. Librarians and Link Rot: A Comparative Analysis with Some Methodological Considerations. *Portal Libr. Acad.* **2003**, *3*, 309–314. [CrossRef]
7. Ureña-Cámara, M.A.; Nogueras-Iso, J.; Lacasta, J.; Ariza-López, F.J. A Method for Checking the Quality of Geographic Metadata Based on ISO 19157. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 1–27. [CrossRef]
8. Quarati, A.; De Martino, M.; Rosim, S. Geospatial Open Data Usage and Metadata Quality. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 30. [CrossRef]
9. Klein, M.; Van De Sompel, H.; Sanderson, R.; Shankar, H.; Balakireva, L.; Zhou, K.; Tobin, R. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE* **2014**, *9*, e115253. [CrossRef] [PubMed]
10. Sife, A.S.; Bernard, R. Persistence and Decay of Web Citations Used in Theses and Dissertations Available at the Sokoine National Agricultural Library, Tanzania. *Int. J. Educ. Dev. Using Inf. Commun. Technol. (IJEDICT)* **2013**, *9*, 85–94.
11. Gertler, A.L.; Bullock, J.G. Reference Rot: An Emerging Threat to Transparency in Political Science. *PS-Polit. Sci. Polit.* **2017**, *50*, 166–171. [CrossRef]
12. Ingham, D.; Caughey, S.; Little, M. Fixing the "Broken-Link" Problem: The W3Objects Approach. *Comput. Netw. ISDN Syst.* **1996**, *28*, 1255–1268. [CrossRef]
13. Nielsen, J. Fighting Linkrot. 1998. Available online: https://www.nngroup.com/articles/fighting-linkrot/ (accessed on 13 January 2021).
14. Harter, S.P.K. ARCHIVE: Electronic Journals and Scholarly Communication: A Citation and Reference Study. *J. Electron. Publ.* **1997**, *3*. [CrossRef]
15. Koehler, W. An Analysis of Web Page and Web Site Constancy and Permanence. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 162–180. [CrossRef]
16. Davis, P.; Cohen, S. The Effect of the Web on Undergraduate Citation Behavior 1996-1999. *J. Am. Soc. Inf. Sci. Technol.* **2001**, *52*, 309–314. [CrossRef]
17. Casserly, M.F.; Bird, J.E. Web Citation Availability: Analysis and Implications for Scholarship | Casserly | College & Research Libraries. *Am. Commun. J.* **2003**, *9*, 300–317. [CrossRef]
18. Wren, J.D.; Johnson, K.R.; Crockett, D.M.; Heilig, L.F.; Schilling, L.M.; Dellavalle, R.P. Uniform Resource Locator Decay in Dermatology Journals: Author Attitudes and Preservation Practices. *Arch. Dermatol.* **2006**, *142*, 1147–1152. [CrossRef] [PubMed]
19. Dimitrova, D.V.; Bugeja, M. Raising the Dead: Recovery of Decayed Online Citations. *Am. Commun. J.* **2007**, *9*, 2.
20. Rhodes, J.S. Web Sites That Heal. 2002. Available online: http://web.archive.org/web/20160315090512/http://www.webword.com/moving/healing.html (accessed on 15 March 2016)
21. Rajabifard, A.; Kalantari Soltanieh, S.; Binns, A. SDI and Metadata Entry and Updating Tools. In Proceedings of the GSDI 11 World Conference, Rotterdam, The Netherlands, 15–19 June 2009; GSDI Association: Manouba, Tunisia.
22. Olfat, H.; Kalantari, M.; Rajabifard, A.; Williamson, I.P.; Pettit, C.; Williams, S. Exploring the Key Areas of Spatial Metadata Automation Research in Australia. In Proceedings of the GSDI 12 World Conference: Realising Spatially Enabled Societies, Singapore, 19–22 October 2010; Leuven University Press: Leuven, Belgium.
23. Brewington, B.E.; Cybenko, G. Keeping up with the Changing Web. *Computer* **2000**, *33*, 52–58. [CrossRef]
24. Cho, J.; Garcia-Molina, H. The Evolution of the Web and Implications for an Incremental Crawler. In Proceedings of the Conference on Very Large Databases, Cairo, Egypt, 10–14 September 2000; p. 18.
25. Koehler, W. Web Page Change and Persistence—A Four-Year Longitudinal Study. *J. Am. Soc. Inf. Sci. Technol.* **2002**, *53*, 162–171. [CrossRef]
26. Fetterly, D.; Manasse, M.; Najork, M.; Wiener, J.L. A Large-Scale Study of the Evolution of Web Pages. *Softw. Pract. Exp.* **2004**, *34*, 213–237. [CrossRef]
27. Ntoulas, A.; Cho, J.; Olston, C. 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), New York, NY, USA, 17–20 May 2004; pp. 1–12. [CrossRef]
28. Adar, E.; Teevan, J.; Dumais, S.T.; Elsas, J.L. 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), Barcelona Spain, 9–12 February 2009; pp. 282–291. [CrossRef]
29. Sanderson, R.; Van de Sompel, H.; Burnhill, P.; Grover, C. 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), Indianapolis, IN, USA, 25 July 2013; p. 21. [CrossRef]

30. Burnhill, P.; Mewissen, M.; Wincewicz, R. Reference Rot in Scholarly Statement: Threat and Remedy. *Insights UKSG J.* **2015**, *28*, 55–61. [CrossRef]
31. Neumaier, S.; Umbrich, J.; Polleres, A. Automated Quality Assessment of Metadata across Open Data Portals. *J. Data Inf. Qual.* **2016**, *8*, 2:1–2:29. [CrossRef]
32. European Data Portal. Metadata Quality Dashboard—Methodology. 2020. Available online: https://www.europeandataportal.eu/mqa/methodology?locale=en# (accessed on 29 March 2021).
33. INSPIRE Joint Research Centre. *Geoportal Workflow for Establishing Links between Data Sets and Network Services*; Technical Report; INSPIRE Joint Research Centre: Brussels, Belgium, 2020.
34. W3C. *Data Catalog Vocabulary (DCAT)—Version 2*; Technical Report; World Wide Web Consortium: Cambridg, MA, USA, 2020.
35. Nogueras-Iso, J.; Lacasta, J.; Ureña-Cámara, M.A.; Ariza-López, F.J. Quality of Metadata in Open Data Portals. *IEEE Access* **2021**, *9*, 60364–60382. [CrossRef]
36. Freed, N.; Borenstein, N. *Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies*; RFC 2045; RFC Editor: Marina del Rey, CA, USA, 1996.
37. Nebert, D.; Whiteside, A.; Vretanos, P. *Open GIS Catalogue Services Specification (Version: 2.0.2)*; Technical Report; Open Geospatial Consortium, 2007.
38. Kessler, G. File Signatures, 2002. Available online: https://www.garykessler.net/library/file_sigs.html (accessed on 4 February 2021).
39. GO FAIR. FAIR Principles. 2016. Available online: https://www.go-fair.org/fair-principles/ (accessed on 22 September 2021).
40. Andrews, K.; Kappe, F.; Maurer, H. The Hyper-G Network Information System. *J. Univers. Comput. Sci.* **1995**, *1*, 206–220. [CrossRef]
41. Sun, S.; Reilly, S.; Lannom, L.; Petrone, J. *Handle System Protocol (Ver 2.1) Specification*; RFC 3652; RFC Editor: Marina del Rey, CA, USA, 2003.
42. *ISO 26324:2012*; Information and Documentation—Digital Object Identifier System. International Organization for Standardization: Geneva, Switzerland, 2012.
43. Klump, J.; Huber, R.; Diepenbroek, M. DOI for Geoscience Data—How Early Practices Shape Present Perceptions. *Earth Sci. Inform.* **2016**, *9*, 123–136. [CrossRef]
44. The Internet Archive. Wayback Machine. 2001. Available online: https://web.archive.org/(accessed on 13 January 2021).
45. Van de Sompel, H.; Nelson, M.; Sanderson, R. *HTTP Framework for Time-Based Access to Resource States—Memento*; RFC 7089; RFC Editor: Marina del Rey, CA, USA, 2013.
46. WebCite Consortium. WebCite. 1998. Available online: https://www.webcitation.org/(accessed on 28 December 2020).